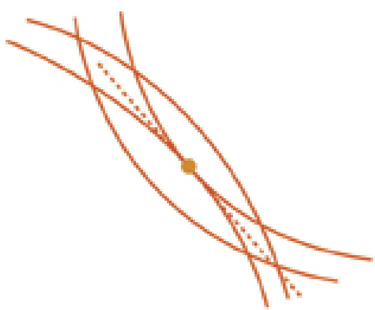
E C O N O M I C S
C A L C U L U S
E C O N O M I C S
C A L C U L U S
E C O N O M I C S
C A L C U L U S

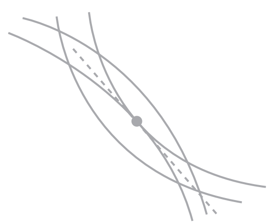
Economics *with* Calculus



Michael C Lovell

Economics
with
Calculus

Economics *with* Calculus



Michael C Lovell
Wesleyan University, USA

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

This page intentionally left blank

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401–402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Economics with calculus / Michael C. Lovell.
p. cm.
Includes bibliographical references and index.
ISBN 9812388257 (alk. paper) -- ISBN 9812388575 (pbk. : alk. paper)
1. Economics, Mathematical. 2. Calculus. I. Title.

HB135 .L68 2004
330.01'515--dc22
2004052464

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Printed in Singapore.

I dedicate this book to my family

Adrienne Goolkasian Lovell

Leslie and Kenneth Chausse: Nathaniel, Colleen, Laurel

Stacie and Steve Markoski: Michael, Andrew, Alexander

George Lovell and Carrie Cihak

Martin Lovell and Sharon Hunter: Mica

This page intentionally left blank

Preface

How should one approach the study of economics? There is no one answer to this question because the appropriate learning style depends in large measure upon the reader's prior training and interests. Those blessed with a working knowledge of calculus are prepared to absorb the basic analytical framework of economics by a much easier route than those who lack this background.

Each year approximately 90,000 high school students achieve a 4 or a 5 score on the Advanced Placement Calculus Examination (AB or BC). Countless more complete their first year of calculus in college before enrolling in the introductory economics course. My experience from teaching a calculus-based introductory course is that students blessed with this background can learn economics more easily and with more enjoyment by making use of their specialized training. The author of a prominent introductory economics text has commented that "to mathematically sophisticated students the introductory [textbook] models look naïve and simplistic; these students are discouraged from going into economics because it is too simple."[1] Students with a working knowledge of calculus will find this text provides an exciting challenge. Far from being dumbed down, this book is designed to help the reader smarten up.

Some economics textbooks claim they enable the student to learn economics without tears. Some even claim they can provide a useful introduction to economics without either graphs or equations! This book presents an honest introduction to economics that can be covered in one semester, not by thinning the soup, but by assuming that the reader has a working knowledge of the calculus. I strive to make the argument as simple as

[1] David Colander, "Telling better stories in introductory macro," *American Economic Review*, May 2000.

possible, but no simpler. We do not use the calculus to complicate the analysis but to simplify the presentation.

The standard introductory economics textbook presents economic theory in translation — it is a translation of concepts developed with mathematical tools into non-threatening English. In order to avoid talking calculus to the uninitiated, the standard introductory textbook introduces such basic concepts as "marginal revenue," "marginal cost," and "marginal utility" without telling the reader that they are first derivatives. The end of chapter problems in the standard textbook are designed for the numerically challenged. The exercises in this text challenge students to put their quantitative skills to work. Students with a strong mathematics background find that it is easier to learn the basic principles of economics using the calculus rather than reading economics in translation.

Scope

There is a danger in the introductory course that one will try to do too much. There is the task of allocating scarce time to diverse pedagogical objectives. The task is complicated, because for many students the introductory course may be the only economics course they will ever take. Therefore, this book is not just hardcore economics — readers must be exposed to the breadth of concepts they have to know in order to function effectively as economic citizens. A student who invests a semester or two in the study of economics should acquire the background necessary for understanding articles in the newspaper's financial section and for segregating enlightened insight from economic nonsense on the editorial page. Because many students find that the introductory course turns out to be only the beginning, one of the responsibilities of the introductory text is to provide students planning to major in economics with a good overview of the discipline and an understanding of the types of intellectual demands imposed upon the student of economics. This text is designed to help continuing students master the basic analytic tools they will be expected to bring with them when they enter more advanced courses.

I also hope that this volume will be an interesting independent read for anyone who majored in science or engineering in college but now wishes to pursue the study of economics. It should be of interest to anyone with a quantitative background who wishes to study economics in preparation for a career shift into the worlds of business or finance. It may also interest anyone with a quantitative bent who enjoys reading about economics and business developments in the popular press but wants to take a deeper and more structured look at how economists analyze how the system works.

Mathematical prerequisites

In developing these materials I have given a high priority to not requiring more mathematics than is covered in the standard high school Advanced Placement Calculus course (AB or BC). I do not assume that the reader has studied probability or statistics. Because Lagrangian multipliers are not covered in AP Calculus, I solve problems of constrained maximization by substituting the constraint into the objective function, but Lagrangians are covered in an appendix for any reader who has a stronger mathematical background than is assumed in this text. Partial differentiation, the one indispensable calculus topic that is not covered in the standard Calculus Advanced Placement Course, is carefully and patiently introduced to the student during the discussion of demand functions in Chapter 3. When developing models of economic growth I have found that many of my students find e^{rt} intimidating; therefore, we work in discrete time and write $(1+r)^t$.

I have found that students who have earned a 4 or a 5 on the Calculus AP course (AB or BC) do not find the mathematics employed in this text too much of a stretch.[2] Students with a lower score should consider studying more math before attempting a course based on this text or else take a conventional introductory economics course. The level of mathematics in this text may seem pedestrian to upper division students in the sciences who have a stronger mathematical background than this text presumes. It is obviously *not* the type of mathematics that economists use in serious research these days. But after completing the overview of economics provided by this text, the mathematically sophisticated reader will have the economics background required to read more advanced economics texts focusing on whatever areas of economics may be of particular interest.

Organization

In deciding how to organize the material in this text I have kept several principles in mind.

[2] If their calculus is rusty, they may wish to consult Stephen Silver's review at http://www.citadel.edu/faculty/silver/calculus.pdf.

- First, I place micro before macro in order that the students will have a foundation in micro that can be drawn upon in developing macroeconomic concepts.
- Second, I carefully pace the use of mathematical concepts so as to enable rusty students to gradually recall the details of their calculus course. Thus, Chapter 2 refreshes the student's memory of how to use the calculus to solve maximization problems and contains the simplest possible example of constrained maximization. Once our students have flexed their calculus muscles on economic problems the pace of the analysis can accelerate.
- Third, I try to move from the simple to the complex. For example, I discuss monopoly before competitive markets because it is easier. Again, I do not begin the discussion of macroeconomics with growth theory, however logical that might be, because students find it easier to start worrying about fiscal and monetary policy issues before moving on to the complexities of growth theory in Chapter 12.
- Fourth, I decided not to segregate international trade topics to a chapter near the end of the book. This decision was motivated in part by my fear that this chapter might be skipped over at the end of semester rush. But the primary reason is that international trade permeates practically every aspect of economics and provides wonderful examples for illustrating the applicability of economic theory to important policy issues. So the theory of comparative advantage is discussed when the production transformation curve is introduced in Chapter 2. In Chapter 3 on supply and demand the student is introduced to the topics of foreign exchange rates, tariffs and quotas. In almost every chapter, the student encounters trade issues.

There is room for flexibility for those who would prefer to sequence the topics in a different order. For example, one can start the study of macroeconomics with Chapter 12 on growth and then drop back, perhaps selectively, to read about economic indicators and monetary and fiscal policy.

Classroom alternatives — pick and choose

This book obviously presents more material than most instructors will want to use, particularly in a one semester course covering both micro and macroeconomics. Some parts are essential. Other sections may be judiciously sampled in accordance with the interests of the students and the objectives of the instructor. Here are some alternative strategies for using this book in the classroom:

1. I have found it possible to cover in one semester most but not all the material in this book with a group of highly motivated students with strong mathematical skills but no prior coursework in economics. I have supplemented the text with a paperback book of real-world readings from the financial press, newspapers and popular magazines.[3] An instructor teaching such a course will probably want to make a point of covering the following barebones core sections of the text plus other sections that would be of particular interest in the light of current economic developments or special concerns of the students.

   The Micro Core: Ch 1, Ch 2, Ch 3.1–3.6, Ch 4.1–4.3.1, Ch 5.1, 5.2, 5.4, 5.5.1, 5.5.2, 5.6, Ch 6.1–6.3, 6.6, Ch 7.1–7.2.1.

   The Macro Core: Ch 1.5.1–1.5.4, Ch 8, Ch 9.1–9.3, 9.5, Ch 10.1–10.4, and Ch 12.1, 12.4, 12.5.1–12.5.4.

2. In a yearlong micro-macro principles sequence there is ample time to cover the Micro and Macro Cores plus Ch 10.5 (aggregate demand and supply), Ch 10.6 (Monetarists versus the Keynesians), Ch 11 (expectations, uncertainty and inflation), Ch 12.2 (Malthusian population dynamics), Ch 12.3 (classical growth), Ch 12.4 (growth accounting), Ch 12.6 (population trends), Ch l2.7 (exhaustible resources) and Ch l2.8 (over fishing).

   For greater depth in a two-semester first course, this text may be read alongside a conventional introductory economics textbook. Read the conventional text for institutional details, but turn to this book whenever the text starts to develop the theory. Alternatively, one can progress chapter by chapter through this text, but look for real world applications on the internet or in the *Economic Report of the President*, the *New York Times*, the *Wall Street Journal*, *The Economist*, or an appropriate compendium of supplemental readings.

---

[3] *Economic: Annual Editions*, ed: Don Cole, McGraw Hill-Dushkin. The web now provides a wonderful source for current event materials. This approach is articulated by Professors Shyamala Raman at Saint Joseph College, Jean Shackelford at Bucknell University and Kim Sosin at the University of Nebraska Omaha in their "just-in-time syllabus" home page at http://ecedweb.unomaha.edu/jits.htm.

3. I have used the book as a supplement to a traditional non-calculus (delta) intermediate microeconomics text, assigning Chapters 1 through 7 plus Chapter 11.2 on market dynamics and speculation, 11.4 on rational expectations, 12.7 on exhaustible resources and 12.8 on over-fishing.
4. Students enrolled in a standard intermediate macro course may supplement their readings in a conventional nonmathematical text with Ch 1.5.1–1.5.4, Ch 8–10, Ch 11.3–11.5 and Ch 12.1–12.6.

Controversy

Economics is controversy, among professional economists no less than in Washington and in the financial press. There is no shortage of controversy in microeconomics. Thus MIT economist Franklin M. Fisher served as the key witness for the prosecution at the Microsoft antitrust trial while his MIT colleague Richard Schmalensee was the chief economist for the defense. The disagreement among macroeconomists is extreme, having at times reached the point where the discipline was all too accurately described as being in a state of chaos. At the 1997 American Economic Association convention, several prominent macroeconomists responded to the question of whether there was a common core to their discipline. Most thought there was a core to macroeconomics, but they could not agree on what it was. One participant at the session commented that there is a core of practical macroeconomics, but went on to say "This believable core model falls well short of perfection, leaves many questions unanswered and is subject to substantial stochastic errors."[4] It is likely that *any* macroeconomics text will be judged to be in serious error by at least 2/3rds of the profession.

There is a temptation in the textbook, no less than in the classroom, to present economics as revealed truth. Or we may confuse our students by laboriously partitioning the discipline into "classical," "neo-classical," "new-classical," "monetarist," "post-Keynesian," and "neo-Keynesian" schools of thought. This text discusses how macroeconomic thought evolved in reaction to unanticipated historical developments. I present the multiplier, the IS-LM apparatus, and short-run and long-run aggregate demand and supply functions. I develop the concept of rational expectations, the Lucas supply function, a modified Solow growth model subject to diminishing returns to scale, and real business cycle theory.

[4] Alan S. Blinder, "Is there a core of practical macroeconomics that we should all believe?" *American Economic Review*, May 1997.

Exercises

The exercises at the end of the chapters are designed to help readers test and strengthen their understanding of analytical materials. Symbols distinguish two types of questions:

* Indicates questions that elaborate on the analysis of the text, often by considering different applications of the techniques or asking the students to work with slightly more complicated problems. Sometimes they involve extensive independent projects, such as the question at the end of Chapter 1.

\# Distinguishes questions that are more demanding mathematically than the material in the text.

Counter Culture

This is *not* a traditional introductory economics text. It is a slim volume and does not claim to cover all the material in traditional introductory textbooks, which often run to more than 1,000 pages and weigh in at over 4 1/2 pounds. It does not come with a CD Rom presenting videos of the author. It's produced in black and white rather than gaudy color. It's just a book.

Supplements, Updates and Feedback

For supplements and updates, readers should check on the text's web site: **http://mlovell.web.wesleyan.edu/EconCalc**. Readers with comments, suggestions and complaints are invited to email the author at **mlovell@wesleyan.edu**, Subject: EconCalc.

Acknowledgements

My first debt is to Professors Burton C. Hallowell, Gerald M. Meier and Robert A. Rosenbaum for inviting me, ages ago, to moonlight at Wesleyan University in order to teach a special introductory economics course for students with a working knowledge of calculus. My department chair at Yale approved this assignment reluctantly, pointing out that this would not be a career-building activity. He was obviously right, but teaching introductory economics to calculus-aware students was so much fun! So it was with great excitement, several years later, that I resigned my professorship at Carnegie-Mellon University to accept an invitation to become a full-time member of the Wesleyan faculty. My colleague Tom Whitin and I shared responsibility for teaching this special introductory course for many years. This book grew out of the course that I initially taught as a moonlighter.

The book has benefited from the helpful suggestions of many friendly critics. I owe much to the critical comments of Dr. András Simonovits of the Hungarian Academy of Science, Professor Matthew J. Baker of the United States Naval Academy, Professor Lawrence Klein of the University of Pennsylvania, and Professor Stephen J. Silver of the Citadel. I am indebted to all my students, but particularly to Yu-hsin Chang, Adriel Gerard, Benjamin Landis, Ehimika Ohiorhenuan, Gergory Ramkhelawan, Sherida Powell, Eli Staub, Mark Umbarger, and Shenyi Wu for the thoughtful feedback and careful suggestions that helped to make this book more student friendly. Professors Timothy A. Park of the University of Georgia and Roman Weil of the University of Chicago Business School kindly advised me on technical topics.

This book would not have seen the light of day if it had not been for the skilled professional team at World Scientific Publishing. I am particularly indebted to Yubing Zhai and Juliet Lee Ley Chin for their wonderful editorial support and their patience in nursing the book through to completion.

Michael C. Lovell
January 2004

Contents

This page intentionally left blank

1

Introduction

1.1 Economics defined

In a general way, everyone knows what economics is about. All of us participate in the market, buying our bread or the daily paper, searching for a job or relaxing with an inheritance, borrowing for a car or investing in the stock market. We are all part of a system, a system in which some fare much better than others. We are all threatened, but some more than others, by the prospect of unemployment. All of us must worry that in the years to come inflation will erode the value of our savings. All of us can hope to benefit from vigorous economic growth and the maintenance of economic stability. But how should economics be defined?

A modern dictionary[1] defines our subject as follows:

[1] *The American Heritage Talking Dictionary*, 3rd edition, 1994.

ec·o·nom·ics *n. Abbr. econ. (used with a sing. verb). The social science that deals with the production, distribution, and consumption of goods and services and with the theory and management of economies or economic systems.*

Writing more than a century and a half ago, philosopher-economist John Stuart Mill [1806–1873] presented a quite similar statement in his popular *Principles of Political Economy*:

> [The] subject is wealth. Writers on Political Economy profess to teach, or to investigate, the nature of Wealth, and the laws of its production and distribution: including, directly or remotely, the operation of all the causes by which the condition of mankind, or of any society of human beings ... is made prosperous or the reverse.

Renowned British economist Alfred Marshall [1842–1924] defined economics as follows in his *Principles of Economics*:[2]

> Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of well-being.

In contrast, a modern intermediate microeconomics text explains:[3]

> Economics is traditionally defined as the study of *the allocation of scarce resources among competing end uses.* This definition stresses two important features of economics. First, productive resources are scarce — they do not exist in sufficient amounts to satisfy all human wants. This scarcity imposes a variety of constraints on both the choices available to a society and the opportunities open to its members. [Second,] choices must be made about how ... resources will be used. The necessity to make choices leads to the second feature of economics: the concern with how those choices are actually made.

By the time you have finished this book, this last definition may well make the most sense.

---

[2] Alfred Marshall's highly successful text, first published in 1890, went through eight editions.

[3] Walter Nicholson, *Microeconomic Theory: Basic Principles and Extensions*, Dryden, 1995, p 3.

1.2 The scope of economics

These definitions do not indicate the full range of issues studied by economists. For starters, a good way to understand the scope of economics is to consider the following short list.

- How markets work to determine prices and allocate resources.
- How governments influence, for better or for worse, market outcomes through tax policy, tariffs, subsidies, patent protection, environmental policy, etc.
- How a nation's central bank (e.g., the Federal Reserve System of the United States) may influence the money supply, interest rates, unemployment, inflation, and the rate of growth of output.
- How we measure income inequality, inflation, unemployment, and productivity growth.

This is only a short list. Economic researchers today are interested in a much longer list of topics.

One might define economics as what economists study and economists as those who study economics. While such a definition is obviously circular, something of the flavor of what economics is all about can be obtained by perusing the titles of working papers reported on Table 1.1.[4] The topics range from the stock market to monetary policy and from software development to smoking cessation. The papers were produced by the distinguished group of economists associated with the National Bureau of Economic Research (NBER). Founded in 1920, the NBER is a private, nonprofit, nonpartisan research organization dedicated to promoting a greater understanding of how the economy works. The more than 500 professors of economics and business now teaching at universities around the country who are NBER researchers are leading scholars in their fields. As the list makes clear, these scholars are putting the research techniques of economists to work on a wide range of exciting topics.[5]

---

[4]A working paper, such as those listed on the table, is a preliminary draft research report that the author circulate for comment and suggestions before the final version of the paper is published in an economics journal, often more than a year after the working paper has been made available to interested scholars.

[5]The NBER Website, http://www.nber.org, contains a complete list of the working papers produced over the years. Included are abstracts summarizing in a couple of paragraphs the main points of each paper. More than this, faculty and students at universities that subscribe to the service may download over the Internet the complete text of any NBER working paper in PDF format.

Table 1.1. Working papers produced by NBER scholars (week of August 1–7, 1999).

Tradable Deficit Permits: Efficient Implementation of the Stability Pact in the European Monetary Union, Alessandra Casella #7278
Race and Home Ownership, 1900 to 1990, William J. Collins and Robert A. Margo #7277
Price Stability as a Target for Monetary Policy: Defining and Maintaining Price Stability, Lars E.O. Svensson #7276
Size and Growth of Japanese Plants in the United States, Bruce A. Blonigen and KaSaundra Tomlin #7275
The Effects of Direct Foreign Investment on Local Communities, David N. Figlio and Bruce A. Blonigen #7274
The Competition between Competition Rules, Hans-Werner Sinn #7273
Liquidity Crises in Emerging Markets: Theory and Policy, Roberto Chang and Andrés Velasco #7272
Consumption Over the Life Cycle, Pierre-Olivier Gourinchas and Jonathan A. Parker #7271 (IFM, EFG)
ABC at Insteel Industries, V.G. Narayanan and Ratna G. Sarkar #7270
An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output; Olivier Blanchard and Roberto Perotti #7269
Taxing Retirement Income: Nonqualified Annuities and Distributions from Qualified Accounts, Jeffrey R. Brown, Olivia S. Mitchell, James M. Poterba, and Mark J. Warshawsky #7268
No Contagion, Only Interdependence: Measuring Stock Market Co-movements, Kristin Forbes and Roberto Rigobon #7267
Is Hospital Competition Socially Wasteful? Daniel P. Kessler and Mark B. McClellan #7266 (HC)
International Institutions for Reducing Global Financial Instability, Kenneth Rogoff #7265
Trade and Growth: Import-Led or Export-Led? Evidence from Japan and Korea, Robert Z. Lawrence and David E. Weinstein #7264
Can Capital Mobility be Destabilizing? Qinglai Meng and Andrés Velasco #7263
Determinants of Smoking Cessation: An Analysis of Young Adult Men and Women, John A. Tauras and Frank J. Chaloupka #7262
Optimal Monetary Policy Inertia, Michael Woodford #7261
Quality Certification and the Economics of Contract Software Development: A Study of the Indian Software Industry; Ashish Arora and Jai Asundi #7260
A Tax on Output of the Polluting Industry is not a Tax on Pollution: The Importance of Hitting the Target; Don Fullerton, Inkee Hong, and Gilbert E. Metcalf #7259
Is There Monopsony in the Labor Market? Evidence from a Natural Experiment Douglas Staiger, Joanne Spetz, and Ciaran Phibbs #7258
The Band Pass Filter, Lawrence J. Christiano and Terry J. Fitzgerald #7257
Assessing the Impact of Organizational Practices on the Productivity of University Technology Transfer Offices: An Exploratory Study, Donald Siegel, David Waldman, and Albert Link #7256
The Japanese Recession of the 1990s: An Exploration of Its Causes, Albert Ando #7255

Table 1.1. (Continued)

Distortionary Taxation, Excessive Price Sensitivity, and Japanese Land Prices, Kiyohiko G. Nishimura, Fukujyu Yamazaki, Takako Idee, and Toshiaki Watanabe #7254
Are All Banking Crises Alike? The Japanese Experience in International Comparison, Michael Hutchison and Kathleen McDill #7253
Tax Policy and Consumer Spending: Evidence from Japanese Fiscal Experiments, Katsunori Watanabe, Takayuki Watababe, and Tsutomu Watanabe #7252
Determinants of the Japan Premium: Actions Speak Louder than Words, Joe Peek and Eric S. Rosengren #7251
The Japanese Banking Crisis: Where Did It Come From and How Will It End?, Takeo Hoshi and Anil Kashyap #7250
Environmental Policy and Firm Behavior: Abatement Investment and Location Decisions under Uncertainty and Irreversibility; Anastasios Xepapadeas #T0243

1.3 Allocating resources

Although economists investigate a wide range of problems, there is one unifying concern that provides coherence to the investigations of economists. Economists are concerned with resource allocation. Here is a set of questions about resource allocation that every society must resolve, one way or another:

- Who will work at what job?
- What will be produced?
- Who will receive what?

For a colony of ants and for a hive of bees the answers to these questions are determined genetically — some are born to be workers, some may be drones, but only one is destined to be the queen. Quite complex societal relationships can be genetically coded, but programmed species cannot adapt rapidly to change. In medieval Europe, who tilled the fields, who shoed the horses, and who was lord of the manor was determined by what one's father had done — arrangements based on custom rather than genetically programmed can breakdown within a generation or two in response to changing technologies.

Our modern economy is a decentralized system. No central planning agency makes basic decisions about what to produce, who shall work at what job, or who will get to consume how much of what goods. In our decentralized system what happens is the result of millions of individual decisions. The amazing thing is that this decentralized system somehow works:

> Consider, if you will, the yellow lead pencil, one of the simplest devices manufactured in our economy today. Yet no one person knows how to make this simple product. The lumberjack who cuts down the cedar tree knows his trade well but does not know how to mine the graphite. The miner does not know the process by which the graphite is mixed with lead to make sure that the final product is a number 2 1/2 pencil and not a 4 or a 2. Neither the lumberjack nor the miner knows how to mix the yellow lacquer used in coloring the pencil or how to make the eraser that smudges the page when you try to correct a mistake.

Somehow, the market system manages to coordinate the diverse activities of individual decision-makers in a way that results in a useful product rather than chaos. Adam Smith provided a first step toward understanding how this process works more than two centuries ago.

1.4 Adam Smith and the invisible hand of free enterprise

Economics is more than a set of research techniques. It is more than a set of loosely connected topics. Economists share a common interest in a controversial proposition eloquently stated in 1776 by Adam Smith, a Scottish Professor of Moral Philosophy, in *An Inquiry into the Nature and Causes of the Wealth of Nations*:

> Every individual endeavors to employ his capital so that its product may be of greatest value. He generally neither intends to promote the public interest nor knows how much he is promoting it. He intends only his own security, only his own gain.

While Smith may sound cynical in asserting that the selfish pursuit of one's own interest rather than altruism is the prime motivator of economic behavior, he went on to argue that greed is good:

> And he is in this led by an invisible hand to promote an end that is no part of his intention. By pursuing his own interest he frequently promotes that of society more effectively than when he really intends to promote it. It is not from the benevolence of the butcher, the baker and the candlestick maker that we expect our dinner, but from their regard to their own advantage.

Adam Smith was clearly articulating two basic principles underlying economic thinking to this day.

- First of all, we have the behavioral assumption that people are motivated by the desire to maximize their own wellbeing or satisfaction.
- Second, and even more controversial, is the proposition that self-interested behavior can contribute to the betterment of society.

In talking about the "invisible hand" Smith was not referring to the intervention of a big brother or a divine force. He was talking about a force of nature, like gravity or magnetism. The crux of the argument is the proposition that the selfish motivation of economic agents will be held in check by market forces, at least under competitive conditions. As Smith himself emphasized, the pursuit of self-interest must not be totally unconstrained. To take but one example, if property rights are not protected, the butcher, the baker and the candlestick maker will find it to their advantage to close up shop rather than have the fruits of their labor taken by greedy thieves. Adam Smith was optimistic, given his assumption that self-interest is the prime motivator of human behavior, in concluding that the pursuit of individual self-interest frequently contributes more effectively than altruism to the betterment of society.

Smith was arguing in his *Wealth of Nations* against excessive government regulations of economic activity and economic planning. He favored free enterprise and free international trade unconstrained by excessive government regulation, tariffs or quotas. While Smith was eloquent, his proposition that self-interested behavior is in the public interest is counterintuitive rather than self-evident. Conventional wisdom does not condone selfish behavior. Generosity is customarily considered a virtue, particularly in others.

Ever since the publication of the *Wealth of Nations*, economists have debated the validity of Smith's argument. Greed may be good but subject to constraints — obviously, the market system cannot function if property rights are not protected and contracts are not enforceable. Over the years economists have devoted much effort to determining the precise conditions under which self-interested behavior, guided by the market mechanism, will most effectively contribute to the public good. Much of this textbook is devoted to the study of the controversial issues raised by Adam Smith more than two centuries ago.

1.5 Economic performance: An overview

To study economics is to study the economy and ways of looking at it. Indeed, economics could be defined as the study of how the system works and why it sometimes fails. In this introductory chapter we shall take a brief advanced peak at historic economic achievements and recurring problems. Since this is just a quick overview, we will leave to Chapter 8 a detailed explanation of the precise way in which economic performance is measured.

1.5.1 *Economic growth*

How the output per capita produced by the United States economy has fluctuated historically is reported on Figure 1.1.[6] During the 20th century the U.S. sustained an average annual growth rate in per capita output of 2.1% — that may seem like a small percentage, but compounded over 100 years it constitutes a remarkable century of progress. The *Economic Report of the President* for year 2000[7] summed up the implications of this achievement as follows:

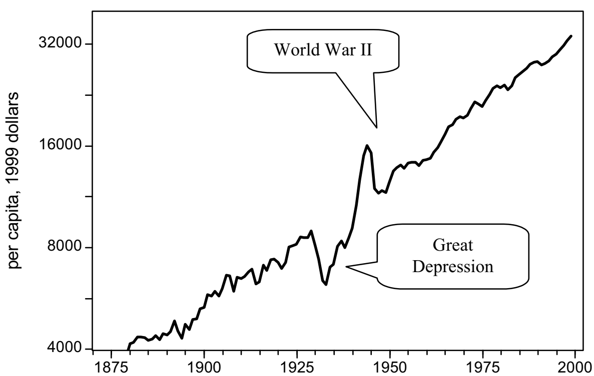


Fig. 1.1. Growth of the American economy
Output per capita, adjusted for inflation (GDP measured in dollars of 1999 purchasing power).

[6]Output per capita is calculated by dividing the Gross Domestic Product (GDP) by population. Such concepts as GDP and the consumer price index will be explained in Chapter 8.

[7]*Economic Report of the President together with the Annual Report of the Council of Economic Advisers*, United States Government Printing Office, 2000, p. 35.

> Over the past century the U.S. economy ... has found the 2% answer to the American dream. [When] living standards rise at 2% annually, they double every 35 years.[8] This means that by the time they reach their mid-30s, parents can provide their children with a standard of living that is twice the level that they themselves enjoyed as children. When incomes grow at this pace, each generation experiences a far more affluent lifestyle than the previous one, and over the course of a lifetime, Americans can expect, on average, a fourfold increase in living standards.

If economic growth were to continue at this pace into the 21st century, today's generation of college students might reasonably expect to enjoy a doubling or tripling of living standards long before they reach retirement age!

The President's *Economic Report* went on to exclaim (p. 278):

> To appreciate how far we have come, it is instructive to look back on what American life was like in 1900. At the turn of the century, fewer than 10% of homes had electricity, and fewer than 2% of people had telephones. An automobile was a luxury that only the very wealthy could afford. Many women still sewed their own clothes and gave birth at home. Because chlorination had not yet been introduced and water filtration was rare, typhoid fever, spread by contaminated water, was a common affliction. One in 10 children died in infancy. Average life expectancy was a mere 47 years. Fewer than 14% of Americans graduated from high school.

Table 1.2. A century of progress.

	1900	1950	2000
Life Expectancy at Birth			
Men	46.3	65.6	73.9
Women	48.3	71.1	79.4
Infant Mortality Rate	99.9	29.2	7.2

Source: *Economic Report of the President*, 2000, p. 166.
Note: The infant mortality rate is the number of deaths of children under one year per 1,000 live births in a calendar year.

[8] $(1 + 0.02)^{35} = 1.9999$.

The international comparisons of output per capita presented on Table 1.3 reveal that the United States has no monopoly on growth. Citizens in all these countries experienced a remarkable increase in material comforts, but there were considerable variations in living standards. In 1820 the United Kingdom led the world, producing almost three times as much per capita as Japan and 40% more per worker than the United States. By 1989 the UK had fallen to 5th place and Japan had soared to 2nd place. The next table compares the growth of China with that of the major western countries over a six hundred year time span. In 1400, the best available evidence indicates, China was ahead of the West in terms of output per capita. No wonder Marco Polo [1254?–1324?], the legendary Italian merchant-explorer, had been amazed by the great wealth he observed on his travels to China. In Chapter 12 we shall be looking at the sources of economic growth.

Table 1.3. Comparative economic performance.

	(US$, 1985)				1820–1989	
Country	1820	1913	1950	1989	Growth factor	Growth rate % per annum
Australia	1,242	4,523	5,931	13,584	11	1.4%
Germany	937	2,606	3,339	13,989	15	1.6%
Italy	960	2,087	2,819	12,955	13	1.6%
Japan	588	1,114	1,563	15,101	26	1.9%
United Kingdom	1,405	4,024	5,651	13,468	10	1.3%
United States	1,048	4,854	8,611	18,317	17	1.7%

Source: Reprinted by permission of Oxford University Press from Angus Maddison, *Dynamic forces in capitalist development: A long-run comparative view*, 1991, pp. 6–7.

Table 1.4. Comparative performance: China and the West.

	(Population in millions; GDP per capita in 1985, US$)			
Year	1400	1820	1950	1989
China				
Population	74	342	547	1,120
GDP per capita	500	500	454	2,361
The West (Western Europe plus Australia, Canada and the U.S.)				
Population	43	122	412	587
GDP per capita	430	1,034	4,902	14,413

Source: Reprinted by permission of Oxford University Press from Angus Maddison, *Dynamic forces in capitalist development: A long-run comparative view*, 1991, p. 10.

We must also ask why some nations have been left behind in misery while much of the world has surged ahead. More than half the children in Bangladesh suffer from malnutrition. Male life expectancy is only 50 years in Kenya. In Haiti less than half the population can read.[9] Harvard professor Benjamin M. Friedman explains:[10]

> ... the most pressing economic problem of our times is that so many of what we usually call 'developing economies' are, in fact, not developing ... [M]any if not most of the world's poorest countries, where very low incomes and incompetent governments combine to create ... appalling human tragedy, are making no progress — at least on the economic front.

Some progress has been made, but it is painfully slow. Xavier Sala-i-Martin estimates that the proportion of the World's population subsisting on less than $2.00 per day has fallen from 44% to 18% over the last quarter century.[11] The improvement has been uneven, and many remain in abject poverty. Far from starting to catch up, some of the world's poorest countries have been slipping further and further behind.

For developing countries that are heavily dependent on agricultural exports, what happens from one year to the next depends not only on the size of their harvest but also on the price that world markets offer for their products. For example, when Vietnam became a major coffee supplier in the 1990s, the world price of coffee dropped precipitously. Honduras, El Salvador and Uganda were particularly hard hit by the price collapse because coffee was their major export.

1.5.2 *The transformation of agriculture*

U.S. agriculture provides a spectacular example of how advances in productivity transform society:

---

[9] For information about global poverty, see *World Development Report 2000/2001: Attacking Poverty*, Oxford University Press, 2001: http://www.worldbank.org/poverty/wdrpoverty/.

[10] Benjamin M. Friedman, "Globalization: Stiglitz's Case," *New York Review of Books*, 15 August 2002.

[11] Xavier Sala-i-Martin, "The disturbing 'rise' of global income inequality," NBER working paper no. 28904, April 2002.

- The value of output produced by each farmer increased on average by about 3% per year throughout the 20th century, more than doubling every 25 years.
- During the past 75 years, production of corn has increased five times over even though the number of acres planted in corn was cut by 16%.
- Since 1975 meat production has increased by 11.5% even though the number of cattle and calves has dropped by about a quarter.
- In 1900 about 40% of civilian workers in the United States were employed on the farm; today only about 2% of the workforce is in agriculture.
- Because the ability of America to produce food has far outstripped the needs of our growing population, roughly 25% of today's farm output is exported.

Mechanization, education, hybrid seed corn, commercial fertilizers and chemical pesticides all contributed to a remarkable expansion in output that is produced with fewer resources. Increased farm productivity has permitted a substantial expansion in farm output coupled with a spectacular decline in farm employment.

The task of adjusting to the increasing bounty of nature generated by technological progress proved far from easy. From the farmers' viewpoint, increased farm productivity had its downside. As will be explained in Chapter 3, increasing productivity contributed to a *fall* in the prices farmers received for their product. A decline of farm incomes relative to what could be earned elsewhere pushed the farmers from their land. The exodus of workers from the farms meant that a large segment of the population had to abandon a cherished way of life. In later chapters we shall be looking at the variety of programs that the government adopted in an effort to ease the plight of the farmer.

1.5.3 *Unemployment*

The historical record makes clear that under capitalism the path of economic expansion is not always smooth sailing — this can be seen by looking back at the output data on Figure 1.1 and the unemployment record on Figure 1.2. The Great Depression of the 1930s stands out on Figure 1.2 as an economic disaster of the first order, for the unemployment rate climbed to 25%. One worker in four could not find a job! The graph also shows that the pace of economic advance is frequently interrupted by recession periods in which output falls below trend and unemployment

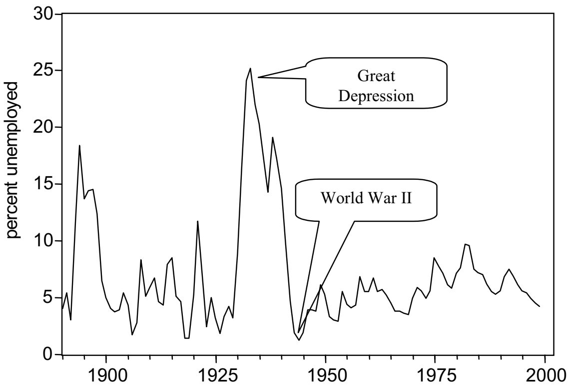


Fig. 1.2. A century of unemployment — USA
The unemployment rate is the proportion of those willing and able to work who cannot find jobs. The unemployment rate peaked at 25% in the depths of the Great Depression of the 1930s. It reached an all time low of 1.2% during World War II.

sharply increases. President William Clinton was one of only a few presidents so fortunate as to escape having an economic recession named in his honor. Presidents Eisenhower, Nixon, Ford, Carter, Reagan and both Bushes all suffered from recessions during their terms of office. The electorate blamed President Herbert Hoover for the Great Depression of the 1930s, voting him out of office after only one term.

That the United States does not have a monopoly on recessions is clear form the international evidence on unemployment presented on Table 1.5. All modern industrialized countries suffer from periods of recession, but some are more susceptible to this economic disease than others. It is interesting to observe that for the first couple of decades after World War II the United States was second only to Canada in terms of the seriousness of the unemployment problem. But in more recent decades the United States has done much better relative to all the other countries listed on the table. The prolonged boom that the United States enjoyed during the last decade of the 20th century was not fully shared with the rest of the industrialized world. And to the surprise of almost everyone, shortly into the 21st century the United States economy slipped into a serious recession.

After explaining how unemployment and recessions are measured in Chapter 8, several chapters will be devoted to an analysis of what is known about the causes of unemployment and the way in which government policy makers attempt to cope with them.

Table 1.5. International comparisons of unemployment rates (%).

YEAR	U.S.	Canada	Australia	Japan	France	Germany	Italy	Netherlands	Sweden	UK	Average
1960	5.5	6.5	1.6	1.7	1.5	1.1	3.7	NA	1.7	2.2	2.8
1969	3.5	4.4	1.8	1.1	2.3	0.6	3.5	NA	1.9	3.1	2.5
1975	8.5	6.9	4.9	1.9	4.2	3.4	3.4	5.1	1.6	4.6	4.5
1979	5.8	7.5	6.3	2.1	6.1	2.9	4.4	5.1	2.1	5.4	4.8
1983	9.6	11.9	10.0	2.7	8.6	6.9	5.9	11.4	3.5	11.8	8.2
1989	5.3	7.5	6.2	2.3	9.6	5.7	7.8	7.0	1.6	7.2	6.0
1990	5.6	8.1	6.9	2.1	9.1	5.0	7.0	6.2	1.8	6.9	5.9
1991	6.8	10.3	9.6	2.1	9.6	5.6	6.9	5.9	3.1	8.8	6.9
1992	7.5	11.2	10.8	2.2	10.4	6.7	7.3	5.6	5.6	10.1	7.7
1993	6.9	11.4	10.9	2.5	11.8	7.9	10.2	6.6	9.3	10.5	8.8
1994	6.1	10.4	9.7	2.9	12.3	8.5	11.2	7.2	9.6	9.7	8.8
1995	5.6	9.4	8.5	3.2	11.8	8.2	11.8	7.0	9.1	8.7	8.3
1996	5.4	9.6	8.6	3.4	12.5	8.9	11.7	6.4	9.9	8.2	8.5
1997	4.9	9.1	8.6	3.4	12.4	9.9	11.9	5.3	10.1	7.0	8.3
1998	4.5	8.3	8.0	4.1	11.8	9.4	12.0	4.0	8.4	6.3	7.7
1999	4.2	7.6	7.2	4.7	11.1	9.0	11.5	NA	7.1	6.1	7.6
Averages	**U.S.**	**Canada**	**Australia**	**Japan**	**France**	**Germany**	**Italy**	**Netherlands**	**Sweden**	**UK**	**Average**
1959–1974	5.0	5.2	2.1	1.4	2.0	0.8	3.4	3.4	1.9	2.9	2.74
1975–1999	6.6	9.1	7.7	2.6	9.2	6.2	7.4	7.2	4.3	8.5	6.88
1990–1999	5.8	9.5	8.9	3.1	11.3	7.9	10.2	6.0	7.4	8.2	7.84
1959–1999	6.0	7.5	5.5	2.2	6.4	4.1	5.8	6.9	3.4	6.3	5.27
Maximum	9.7	11.9	10.9	4.7	12.5	9.9	12.0	11.5	10.1	11.8	8.80
Minimum	3.5	3.4	1.3	1.1	1.2	0.3	2.4	3.1	1.2	2.0	2.14

Notes: NA ~ not available
The data for Germany after 1990 relate to unified Germany
Source: Department of Labor web page.

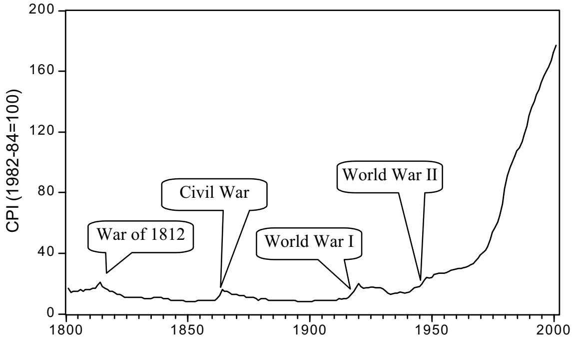


Fig. 1.3. Two centuries of inflation — USA
This graph shows that while inflation had its ups and downs in the 19th century, prices ended up at about the same level at the end of that century as they had been at the beginning. The last half of the 20th century was obviously a much more inflationary story.

1.5.4 *Inflation*

No country has escaped inflation — a general tendency for prices to rise — but some countries have suffered much more than others over the years. A careful study of Figure 1.3 will reveal that inflation is particularly likely to occur when military conflict leads to a substantial increase in government spending and the quantity of money in circulation.[12] Inflation was a world wide problem in the 1970s, thanks in part to the Organization of Petroleum Exporting Countries (OPEC) success in pushing up the price of oil.

Over the years inflation takes its toll on the purchasing power of a nation's currency. The statisticians at the United States Bureau of Labor Statistics estimate that in year 2000 a representative market basket of goods cost the consumer 7.1 times as much as that same basket would have cost in 1950 — prices increased at an average annual rate of 4%. Figure 1.4 reveals that in recent decades Germany has had somewhat less inflation than the U.S. while Japan has had more. But all these inflations are moderate when compared with the runaway inflations experienced by Indonesia and Israel, as can be seen by comparing Figure 1.5 with Figure 1.4, once the difference in scale is noted. In Indonesia's worse year prices on average increased by more than 1000%! In 1985 Israeli prices jumped by 375%!

[12] In Chapter 8 we will learn how to interpret other types of graphs which provide a more accurate indication of the extent of inflation.

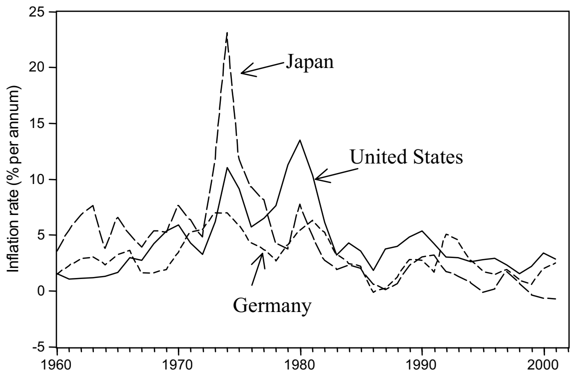


Fig. 1.4. Inflation in Germany, Japan and the United States

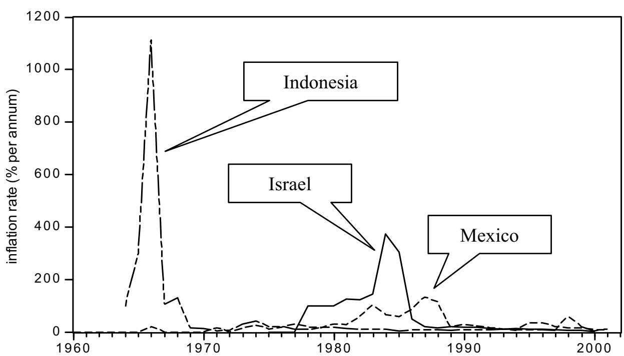


Fig. 1.5. Inflation in Indonesia, Israel, and Mexico

Chapter 8 explains how inflation is measured and discusses strategies by which the public tries to adapt to rapidly rising prices. Later chapters look at the causes of inflation and the policies that countries have adopted in attempting to control the problem.

1.5.5 *Foreign exchange rates*

Students traveling abroad find it necessary to convert their own currency into that of the country they are visiting. American students planning to spend a semester in France or Italy will want to know how many Euros they will be able to get for a dollar — that is the foreign exchange rate.

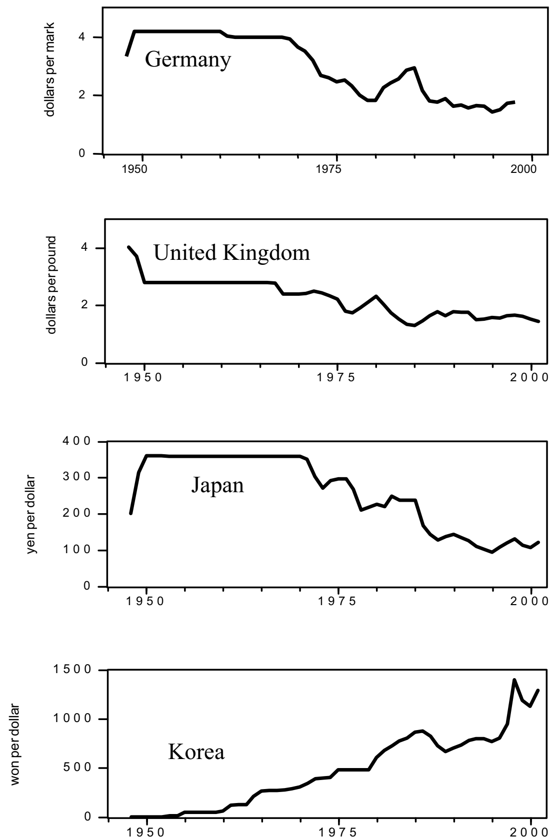


Fig. 1.6. Foreign exchange rates
The graphs reveal that while Americans now get fewer Japanese yen for a dollar, over the years the German mark, the Korean won and the British pound have all lost value relative to the dollar.

Figure 1.6 reveals that exchange rates can fluctuate markedly over the years and indeed from one week to the next. It is not unheard of for students from abroad whose study in America appeared to be more than adequately financed to have suddenly found themselves short of dollars when the value of their currency dropped substantially in the foreign exchange marketplace. In Chapter 3 we will be studying how prices are determined in the market place, including the price of foreign currency.

There is an intriguing feature about these graphs of exchange rate fluctuations that distinguishes them from plots of inflation and unemployment.

At times exchange rates have been remarkably stable. As a matter of government policy, exchange rates are sometimes fixed rather than allowed to fluctuate in response to changing economic conditions. In Chapters 3 and 11 we will be learning how governments can, for better or for worse, try to stabilize certain prices, including the prices of agricultural commodities as well as foreign exchange rates.

While exchange rates are at times remarkably stable, thanks to government intervention, sometimes they are subject to extreme fluctuations. Look at the unexpected collapse of the Korean won on Figure 1.6![13] Exchange rate crises were experienced in Mexico in 1994–1995, in many Asian countries in 1997–1998, in Russia in 1998 and in Argentina in 2001. Currency crises are tremendously disruptive. When a country's currency collapses the nation's importers find that they must pay much more in terms of the domestic currency (peso or won) on purchases made in foreign markets. Banks become unstable. Unemployment soars. The people may riot in the street. Governments collapse when they cannot find a politically acceptable way out of crisis.

1.5.6 *Inequality*

In thinking about how well the economy functions, we must worry not only about how much it manages to produce but about who gets what share of the output. During the last quarter century, not all groups in the United States have shared in the expanding economy. Figure 1.7 shows that those in the 95th percentile (the top 5% of families) have had expanding opportunities while those below the median (the bottom half of the population) have had little or no growth in family income since the late 1970s. In Chapter 7 we shall learn how inequality is measured. We shall find that while per capita income is much higher in the United States than in India, income inequality is about the same in the two countries. Income is much less equally distributed in Brazil and Mexico than in the United States. Income is more equally distributed than in the United States in the majority of industrialized countries, notably Japan. We will also find that inequality among the different countries of the world is much greater than inequality within nations.

[13]The exchange rate scale is quoted as won per dollar on this graph, which means that a rise in the curve signifies a reduction in the value of the won *vis-a-vis* the dollar.

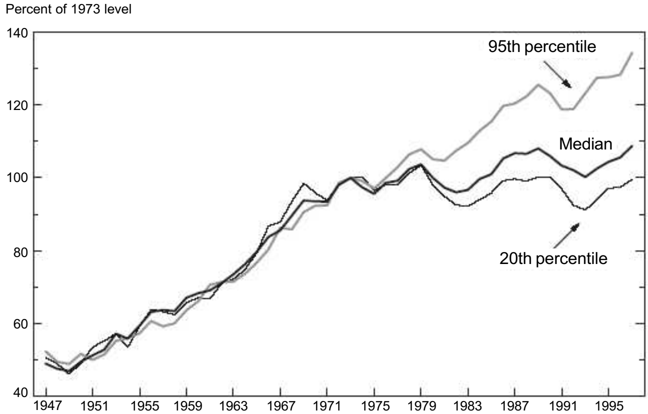


Fig. 1.7. Growth in real family income, 1947–1997
Growth in real family income has slowed and inequality has increased since 1973.
Source: *Economic Report of the President*, 1999.

1.6 Prospectus

This introductory chapter presented several definitions of economics. But no definition can fully capture the true spirit of what economics is all about. Economists bring with them when they study an issue a unique mindset or unifying viewpoint that differs from that of political scientists, sociologists, and others who may be investigating the same phenomenon. Part of that viewpoint is provided by the common concern of economists with the issues of free trade and resource allocation that were articulated so clearly by Adam Smith in his *Wealth of Nations*. The underlying theme of economics since the days of Adam Smith is the theme of this book. It is the proposition that markets do work to allocate resources. This optimistic proposition is tempered by the realization that markets do not work perfectly. Markets sometimes fail. Government does have a role to play in the market place. But how broad a role should this be? Through the study of economics we may hope to learn about what policy remedies work and what economic medicines may do more harm than good.

Economics as a discipline advances not only from the pace of its own momentum but also from the task of trying to explain unanticipated economic developments, such as the 25% unemployment rate suffered in the 1930s, the surprising success of the American economy in mobilizing for World War II, the unpredicted great inflation of the 1970s, and the happy blend of full employment coupled with low inflation in the 1990s. To study economics is to advance one's understanding of how the system works and why it sometimes fails.

Summary

1. There is no shortage of definitions of economics. One dictionary defines economics as the social science that deals with the production, distribution and consumption of goods and services. A more technical definition, which will make more sense by the time you have finished this book, states that "economics is the study of the allocation of scarce resources among competing end users ...".
2. A list of working papers produced by scholars associated with the National Bureau of Economic Research indicated something of the range of topics that economists study.
3. Economists are concerned with resource allocation: Who will work at what job? What will be produced? Who will receive what?
4. In his *Wealth of Nations*, published in 1776, Adam Smith argued that individuals are motivated by their own self interest, but in pursuing it they are guided as if by an invisible hand so as to promote the good of society. The question of when and how markets work to channel self interested behavior for the good of society is a central theme of economic thought.
5. For much of the world the 20th century was a remarkable period of progress. The infant mortality rate fell from 99.9 per 1,000 to 7.2 per 1,000 and life expectancy increased from about 47 to 76 years. Living standards in the United States doubled every 35 years.

Exercises

***Note*:** The exercises at the end of each chapter are designed to help readers test and strengthen their understanding of analytical materials. Symbols distinguish two types of questions:

* Indicates questions that elaborate on the analysis of the text, often by considering different applications of the techniques or asking the students to solve slightly more complicated problems. Sometimes they involve extensive independent projects.

\# Distinguishes questions that are more demanding mathematically than the material in the text.

*Project

This book obviously cannot cover all the topics of interest in as broad a field as economics. Find an article in an economics journal on a topic of particular interest to you that is written by a professional economist for economists. You might look in the *American Economic Review*, the *Quarterly Journal of Economics*, or *Econometrica.* Avoid the *Wall Street Journal*, the *Harvard Business Review*, or other sources that are not written for professional economists.

Do not get bogged down in the details but try to understand the gist of the article by focusing on the introduction and concluding sections of the paper. Put your article aside and read it again when you have finished working through this book. You will be surprised at how much more sense the article makes by the end of the semester. Of course, you should not be surprised to find that there remain some technical details that you cannot understand after only one semester of economics.

You must provide the topic, but how will you find the article? Your best strategy depends on what library resources are available.

Search Strategy #1

You can look on the NBER website mentioned in footnote 5: http://www.nber.org. It has a search facility that will help you uncover any articles written on your chosen topic.

Search Strategy #2

If your library has access to *EconLit* you can search in his comprehensive source for articles published on your topic in any economic journal since 1969.

Search Strategy #3

If your library has *JSTOR* you can not only search for articles on your topic published in any of the 25 leading economics or finance journals in this wonderful data base. You will be able to download the article you found from this archive and print it on your computer.

Write a review

Write a three or four page review of your article. You should explain the method of analysis used by the author (theoretical, statistical?), the type of evidence presented (time series or survey data, experimental?), and the major conclusions of the paper. Did you find the argument convincing? See examples on the EconCalc Web Page: http://mlovell.web.wesleyan.edu/StudentPapers/.

2

Production Possibilities

2.1 Overview

Economics may be defined as "the study of how scarce resources are best allocated to achieve given ends." This chapter will clarify what this means by focusing on some particularly simple but unrealistic choice situations having to do with the allocation of scarce resources. The production transformation curve is the basic analytical construct developed in this chapter. It is an extremely simple device, but it will suffice for introducing the fundamental concepts of "economic efficiency" and "opportunity cost." We shall use this device in developing a simple example of maximizing behavior and in presenting the concept of "relative prices." The production transformation curve will also help in explaining David Ricardo's famous Theory of Comparative Advantage. The concluding section considers Ricardo's argument for the elimination of tariffs and quotas in order to establish efficient trade among nations.

2.2 The production transformation curve

While some nations are wealthier than others, even the richest of nations has only a limited range of resources. At any point in time a nation has only so much in the way of natural resources, factory buildings, machinery and computers. Only a finite number of workers are available for employment. Consequently, the output of even the best endowed of nations is limited. The production transformation curve is a useful analytical technique for examining such limits.

2.2.1 *Hypothetical numerical example*

In order to simplify the exposition, let us start with the simplest possible example. Consider a farmer who can produce only two commodities: Let A be the quantity of the first good (e.g., baskets of apples) and B the quantity of the second (e.g., bushels of barley).[1] The shaded region on Figure 2.1 indicates the set of feasible output combinations that the farmer is capable of producing. According to the graph, the farmer could produce 5 units of A if he focused all his effort on the production of that commodity. If instead he focused on Good B, he could obtain a maximum of 5 units. But by dividing his effort he might produce 3 units of A and 4 units of B. The outer boundary of the set of feasible outputs is called the *production transformation curve*. Sometimes it is referred to as the *production possibility frontier*. The hypothetical data for this production transformation curve are recorded on Table 2.1.

The production transformation curve has a negative slope, which reflects the fact that you cannot have more of Good B without giving up some of Good A. Producing the two types of goods are competing activities — land

[1] Quite frankly, only two goods are considered because this makes it easier to graph the key issues underlying the production transformation concept. The technique of linear programming, explained in Section 4, facilitates the extension of the analysis to a large number of commodities.

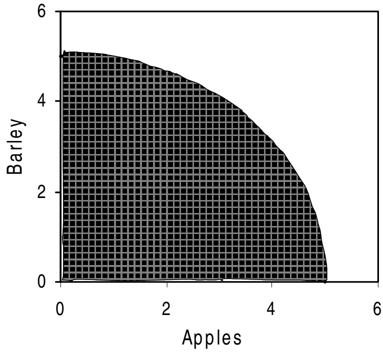


Fig. 2.1. Production possibilities curve
The production transformation curve indicates the set of efficient points. All the output combinations in the crosshatched area on the graph are feasible.

Table 2.1. Production possibilities.

A(pples)	B(arley)
0	5.0
1	4.9
2	4.6
3	4.0
4	3.0
5	0.0

and labor allocated for the production of Good A is not available for the production of Good B.

Let us consider the numerical example on which the graph is based. Suppose that the set of all feasible outputs is

$$A^2 + B^2 \leq 25\,, \quad A \geq 0\,,\ B \geq 0\,. \tag{1}$$

Along the production transformation curve we must have a strict equality:

$$A^2 + B^2 = 25\,, \quad 0 \leq A \leq 5 \ \text{ and } \ 0 \leq B \leq 5\,. \tag{2}$$

To solve for B as a function of A note first that

$$B^2 = 25 - A^2\,, \quad 0 \leq A \leq 5\,. \tag{3}$$

Taking the square root of both sides yields:

$$B = (25 - A^2)^{\frac{1}{2}} = +\sqrt{25 - A^2}\,, \quad 0 \leq A \leq 5\,. \tag{4}$$

This is the relationship that was plotted on Figure 2.1.[2]

NOTE: In thinking about the production transformation curve (and many other economic relationships) it is convenient to pretend that the variables are measured in continuous units. Realistically, of course, this makes more sense for a liquid like cider than it does for apples. But if the numbers represented tons of coal instead of baskets of apples, it would be easier to think in terms of fractions of a unit. Moreover, it will facilitate the application of the calculus if we measure A and B as continuous units.

2.2.2 *Production transformation — Mobilizing for World War II*

For a second production transformation curve example, consider the mobilization of the United States economy for World War II. The curve plotted on Figure 2.2 is a rough but not unreasonable estimate of the tradeoffs between producing civilian goods versus production for the war effort. On the ordinate (y-axis) we plot total expenditures on war. On the abscissa (x-axis) we have civilian utilization, defined as Gross Domestic Product minus wartime expenditure. The two points on the graph refer to the output mix in prewar 1939 and the wartime peak year of 1944.[3]

> In 1939 we had civilian utilization of \$891 billion and war output of \$12 billion.
> In 1944 we had civilian output of \$1121 billion and war output of \$593 billion.

The point labeled 1939 is much lower than that for 1944 because military expenditure was so much less than at the wartime peak. The remarkable expansion of output made it possible to have more guns and more butter — more goods were available for non-military (i.e., civilian) use

[2]Since $A^2 + B^2 = 25$ is the equation for a circle with radius 5, our production function is a quarter circle confined to the positive quadrant, given that output cannot be negative.
[3]Data on unemployment and GDP plotted on Figure 1.1 and Figure 1.2 place the remarkable economic expansion of World War II in perspective.

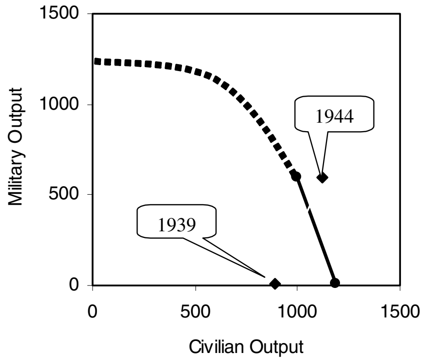


Fig. 2.2. World War II production transformation curve
In 1939 output in the United States was substantially below the production transformation curve because of massive unemployment. In 1944 the economy was operating beyond the production transformation curve. Output was maintained at non-sustainable levels by appealing to workers to support the war effort by accepting overtime, by encouraging retirees to go back to work, by recruiting housewives to work in the factory, and by pushing the unemployment rate down to an unprecedented 1.2%.
Note: Both the civilian and the military output are measured in billons of dollars corrected for the distortions of inflation.

during the war.[4] The reason for this is that the point for 1939 is substantially below the production transformation curve because the unemployment rate in that depression year was 17.2%. The production transformation curve is drawn slightly below the point for 1945 to indicate that output was forced above sustainable levels at the peak of the war. At the peak level of wartime mobilization, unemployment was at an all time low of only 1.2%. "Over-full employment" was achieved by encouraging everyone to show their patriotism and support the war effort by working overtime. Women were asked to leave the home to work in the shipyards and factories while retired workers were encouraged to rejoin the work force for the duration of the war.

2.3 Opportunity cost and the MRT

The old adage that "one cannot obtain something for nothing" has more than a germ of truth. In terms of the example that was provided by

[4]The output figures are not distorted by inflation because output is measured in terms of billions of dollars of 1996 purchasing power. The technique for measuring real output in this way will be explained in Chapter 8.

Figure 2.1, a farmer who was initially producing 3 baskets of apples and 4 bushels of barley decides to produce an additional basket of apples, it will be necessary to cut back barley production by one bushel. The negative slope of the production possibility curve on the graph captures the painful truth that in order to get more of one thing one has to give up something else. One defines the "opportunity cost" of a basket of apples in terms of the sacrifice of barley required to obtain it.

The concept of opportunity cost is widely used by economists. While the public is accustomed to thinking of costs in financial terms, in dollars and cents, economists are not always so crass. It is useful to think of cost in terms of the *sacrificed alternative.*

Example #1

The cost paid by a student in going to a movie is inadequately measured by the price of admission; going to the movie may involve a reduction in time spent studying and a lower grade on tomorrow's examination. Thus the facts not learned and the lower grade must be included in thinking about the opportunity cost of going to that movie. (Of course, the opportunity cost of going to a movie is much less once exams are out of the way.)
It is useful to think of the *opportunity cost* of an item as the value of the sacrificed alternative.

Example #2

Consider a student who can only take four courses this semester. Enrolling in economics means not enrolling in other possible courses. The opportunity cost of enrolling in economics may be defined in terms of the foregone opportunity to enroll in the next best alternative course.

Example #3

A student pays $25,000 tuition per year to attend the university. But if she were not going to college she would be earning $16,000 per year. Thus the opportunity cost of a year at college might be calculated as the sum of the tuition expense *plus* the foregone earnings, or $41,000 per year!

Example #4

A manufacturer paid $500,000 for a milling machine, but it is used only 20 hours per week. The firm has been asked to bid on a new line of business that would keep the machine busy for an additional 10 hours per week. The opportunity cost of using the machine for the new activity is only the wear and tear of the additional usage it will receive, for the alternative to using it is idle time.

Example #5

A freshly minted lawyer, discouraged by the grind in a Madison Avenue Law firm, invests his inheritance in a retail ice cream franchise. The accountant won't record the cost of the owner supplied funds as a cost of doing business, but the opportunity cost of using these funds is the interest (or stock market) earnings foregone that would have been earned if the inheritance had not been invested in the franchise.

Example #6

The government drafted a young man to serve in the Army during the Vietnam War. It paid him $350 a month plus room and board, but he would have earned $1,500 a month in his civilian job. Assuming that his civilian wages are an accurate measure of his productivity in civilian life, the foregone civilian earnings should be included in calculating the opportunity cost of having the draftee serve in the Army. Thus the Pentagon's budget understated the true cost of the war by whole orders of magnitude.

2.3.1 *Marginal rate of transformation (MRT)*

It is clear from Figure 2.1 that the size of the sacrifice of barley required if one is to obtain additional apples depends upon how much in the way of barley is already being produced. For example, if one were producing 3 baskets of apples and 4 bushels of barley, it would be possible to obtain an additional basket of apples by giving up a single bushel of barley. Once we are producing 4 apples and 3 barley, the cost of one more apple in terms of foregone barley is higher, as can be seen from the graph. This prosaic example illustrates a key concept:

> The marginal rate of transformation (MRT), is defined as minus the slope of the production transformation curve; more precisely, MRT $= -dB/dA$.

For our apples-barley example with $B = (25 - A^2)^{1/2}$, we have $dB/dA = 1/2(25 - A^2)^{-1/2}(-2A) = -A/(25 - A^2)^{1/2}$ and so the marginal rate of transformation is $A/(25 - A^2)^{1/2} = A/B$.

The MRT (sometimes called the marginal rate of substitution) is a measure of opportunity cost. The marginal rate of transformation tells us the rate at which we must give up one good (e.g., barley) in order to obtain more of the other (apples).

All this makes sense, but why shouldn't the production transformation curve be drawn as a straight line, so that the marginal rate of transformation is constant? Such linearity might make sense if there were no variation in the productivity of land and labor. But here is one counter argument:

> Some land is better suited for the production of Good A whereas other land may be better suited for Good B. Similarly, some workers may be better than average at producing Good A but may not excel as much at harvesting Good B.
>
> Now suppose that the farmer decides that half the labor and half the land are to be utilized for the production of Good A and the remainder for Good B. Our farmer decides to allocate the land and workers who are better at producing Good A to that task and those whose strength lies in producing Good B to that good.
>
> If the farmer now changes his mind and decides to abandon the production of Good B and reallocate all of production effort to Good A, output of A will not double, even though twice as much labor and land are now allocated to its production. Output of A will not double because the new workers and land will be less productive in that use than those who had already been allocated to the production of A.

This is but one of a number of arguments explaining why the production transformation curve will bow out, as indicated on the graph, which means that the more you consume of the commodity the higher its opportunity cost. This key concept is sometimes referred to as the "Law of Increasing Opportunity Cost."[5]

[5] The Law of Increasing Opportunity Cost is not a universal truth. For example, consider the production of honey and apples. An orchard will yield more apples if it is pollinated by honey bees, and the bees will yield more honey when they find blossoming apple trees. Thus the two types of output are complementary. Indeed, modern farmers hire beekeepers to bring truckloads of hives into their orchards at blossom time.

2.3.2 *A simple supply function*

Suppose that our farmer is trying to decide the appropriate mix of apple and barley output. Our farmer believes that the price of apples will be $p_a = \$2.00$ a basket and the price of barley $p_b = \$1.00$ per bushel. For simplicity let us suppose that the costs of inputs are negligible so that all revenue is profit. The objective is to maximize total revenue. Now revenue will be

$$R(A, B) = p_a A + p_b B\,, \tag{5}$$

Or with the specified prices

$$R(A, B) = 2A + B\,. \tag{6}$$

Our farmer wishes to maximize this function of two variables, but output is constrained by the productive capacity of his farm as summarized by the production transformation curve. Thus the farmer wants to solve the following problem:

Maximize

$$R(A, B) = 2A + B \tag{7}$$

subject to the production function constraint, equation (4):

$$B = \sqrt{25 - A^2}\,, \quad 0 \le A \le 5\,. \tag{8}$$

Thus the problem is to maximize a function of two variables subject to a single constraint.

Before solving this problem with the calculus, it is helpful to consider it graphically. Figure 2.3 plots the production transformation curve and several iso-revenue lines. The *iso-revenue lines*, derived from equation (7), show all combinations of A and B yielding the specified revenue. An iso-revenue line partitions the graph into two regions. All points below the iso-revenue line yield less revenue than that specified for the line. All points above the iso-revenue line yield more revenue.

> Since $p_a = \$2.00$ and $p_b = \$1.00$, the equation for the iso-revenue line constituting the set of points yielding revenue R is $B = R - 2A$; that is why the iso-revenue lines are all graphed with the same slope.

It is clear from the $R = \$10$ iso-revenue line on the graph that there are a number of feasible output combinations yielding \$10 of revenue. However, it is not possible to obtain \$17.00 of revenue because no point on the $R = \$17$

iso-revenue line is feasible. The revenue maximizing level of output is at point e where the $R = \$11.18$ iso-revenue line is tangent to the production transformation curve; maximum profit is \$11.18.

It may be useful to reiterate that the slope of the iso-revenue line depends upon the prices of the two goods. To see precisely how, note that we can obtain from equation (5)

$$B = \frac{R}{p_b} - \left(\frac{p_a}{p_b}\right) A\,. \tag{9}$$

Thus the intercept is R/p_b and the slope is $-p_a/p_b$. An increase in R will shift the iso-revenue curve outwards but will not change its slope. An increase in p_b will make the iso-revenue lines flatter, and this will lead the revenue maximizing producer to slide to the left along the production transformation curve so as to produce more barley and fewer apples. This process will be clarified by solving the problem with the calculus.

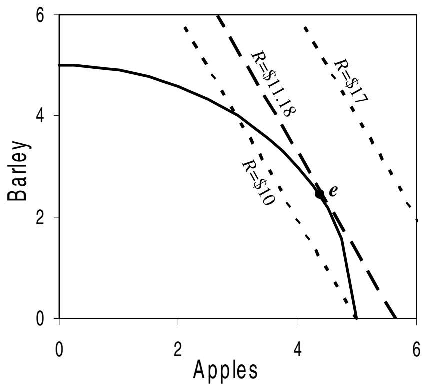


Fig. 2.3. Maximizing revenue when p = \$2 and p = \$1.
Each iso-revenue line is the set of points yielding the specified revenue. All feasible points to the north-east (up and to the right) of the \$10 iso-revenue curve yield more than \$10 of revenue. None of the points on the \$17 iso-revenue curve is feasible. Maximum revenue is obtained at point e where the \$11.18 iso-revenue curve is tangent to the production transformation curve.

One way to use the calculus to solve our problem of maximizing a function of two variables subject to one constraint is to convert our problem by substitution into the more familiar task of maximizing a function of a single variable without a constraint. To do this we substitute equation (8) into (7) to obtain

$$R(A) = 2A + \sqrt{25 - A^2}\,. \tag{10}$$

This new function reveals revenue as a function of A alone because it implicitly takes into account the loss in revenue derived from the production of less barley when more apples are produced. To maximize revenue we set the first derivative equal to zero:

$$\frac{dR}{dA} = 2 + \frac{1}{2}(25 - A^2)^{-\frac{1}{2}}(-2A) = 0\,. \tag{11}$$

Hence,

$$(25 - A^2)^{-\frac{1}{2}} = 2A^{-1}\,, \tag{12}$$

which has solution $A = \sqrt{20} = 4.47$ and $B = 2.24$.[6] The farmer will obtain $R = 2 \times 4.47 + 2.24 = \$11.18$ when the goods are brought to market.

More generally, we can derive the supply function for A showing how the quantity of apples supplied by our maximizing farmer responds to changes in price. Substituting production constraint (8) into equation (5) yields revenue function

$$R(A) = p_a A + p_b\sqrt{25 - A^2}\,. \tag{13}$$

Differentiating with respect to A, just as before, yields

$$\frac{dR}{dA} = p_a + \frac{1}{2}p_b(25 - A^2)^{-\frac{1}{2}}(-2A) = 0\,; \tag{14}$$

To solve for the supply function of A, first note that (14) yields $p_a/p_b = (25 - A^2)^{-1/2}A$; therefore, $(p_b/p_a)^2 = (25 - A^2)/A^2 = 25/A^2 - 1$, or $1 + (p_b/p_a)^2 = 25/A^2$, which leads to

$$A = 5\left[1 + \left(\frac{p_b}{p_a}\right)^2\right]^{-\frac{1}{2}}\,. \tag{15}$$

For example, if $p_b = 0$ then $A = 5$ for any $p_a > 0$. For large p_b/p_a, A approaches 0. Or if $p_a = p_b = 4$, $A = B = 5/2^{1/2} = 3.53$. Again, if prices double so that $p_a = p_b = 8$, we will again have $A = B = 5/2^{1/2}$. It is the ratio p_a/p_b that matters.

[6] The conscientious student may want to check that the second order conditions for a maximum are satisfied; i.e., $d^2R/dA^2 < 0$.

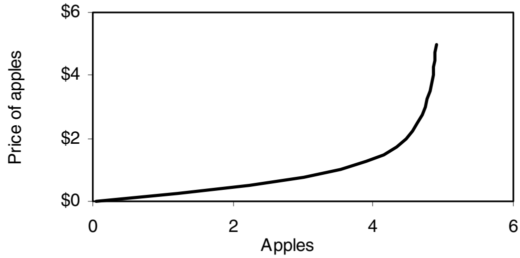


Fig. 2.4. Supply function for Good A, given $p \ $ = \$1.00
The supply function may be derived from the production possibility curve, assuming for simplicity that the firm maximizes sales revenue because there are no labor, material, or other production costs. To see why, note that the slope of the production transformation curve is equal to $-p \ /p$ at the tangency point of maximum revenue. When p increases, given p , the iso-revenue line becomes steeper; therefore, the point of maximum revenue slides to the right signifying that more of Good A is supplied.

The model has a rather surprising implication:

> If the prices of the two goods were to double in a period of inflation, the quantity of the goods supplied would be unaffected because p_b/p_a would remain unchanged.

The fundamental point is that the supply of Good A cannot be determined by just knowing its price, p_a. The supply of Good A depends on the price p_a *relative* to the price of B: p_a/p_b. Quite generally, for much more elaborate models than the simplest case being considered here, economists find that it is *relative prices* that matter.[7] It is possible to graph a *supply function* showing how much of a good will be supplied as a function of its price, but only if the price of the other good is specified, as illustrated on Figure 2.4.

2.4 Linear programming

2.4.1 *Introduction*

In the 1940s a mathematical technique known as *linear programming* was developed in the United States by George Dantzig and independently in the Soviet Union by L. Kantorovich. The first practical application of linear programming (LP) in the United States dealt with the efficient deployment of military forces and the solution of military planning problems. In the

[7]When economists speak of "relative prices" we mean prices compared to each other as contrasted with their absolute dollar level.

early 1950s, A. Charnes and W. W. Cooper pioneered cost saving industrial applications of LP, focusing on the petroleum industry. Planners in socialist countries (Soviet type) used linear programming techniques to derive schedules for the optimal allocation of resources.

Despite the similarities of name, linear programming has no direct connection with computer programming, except that computer programs are used to solve linear programming problems, usually with the Simplex Algorithm developed by George Dantzig. The standard linear programming problem involves the maximization of a linear function in k variables

$$Z = c_1X_1 + c_2X_2 + \cdots + c_kX_k \,, \tag{16}$$

subject to n inequality constraints of the form

$$a_{i1}X_1 + a_{i2}X2 + \cdots + a_{ik}X_k \leq b_i \,, \quad i = 1 \ldots n \,, \tag{17}$$

where the a_{ij} and b_i are regarded in many applications as being technologically determined. Industrial applications may involve solving problems with thousands of inequality constraints (n) and variables (k). A production transformation example will indicate one application of this mathematical technique.[8]

2.4.2 *Simplified example*

Suppose a manufacturer has 20 machines and 10 workers. To produce a unit of Good X requires the use of 2 machines and the labor of 2 workers for one week. To produce a unit of Good Y requires 1 worker and 4 machines. The production processes are characterized by *constant returns to scale*, which means that if all inputs to the production process are doubled then output will precisely double as well.[9] How can we determine the production possibility curve showing the range of feasible weekly outputs that this manufacturer can produce?

This problem cannot be solved using the calculus because of the two inequality constraints! Fortunately, the problem is simple enough to permit a graphical solution. Observe that the limited number of machines means that the following inequality constraint must be satisfied:

---

[8]The simplex technique for solving such problems is presented in a number of texts, including: Alpha C. Chiang, *Fundamental Methods of Mathematical Economics*, Irwin/McGraw-Hill; 1984. Michael Hoy, John Livernois, Chris McKenna, RayRees and Thanasis Stengos, *Mathematics for Economists*, MIT Press, 2001.

[9]The constant returns to scale concept will be clarified in Chapter 5.4.

$$2X + 4Y \leq 20\,. \tag{18}$$

The triangular region plotted on the upper left-hand panel of Figure 2.5 constitutes the set of all points satisfying this inequality and the non-negativity constraints ($X \geq 0$ and $Y \geq 0$). Next observe that the limited labor supply means that the following constraint must also be satisfied:

$$2X + Y \leq 10\,. \tag{19}$$

Plotted on the upper right-hand panel is the triangular region of output combinations that satisfy this labor constraint. Finally, the lower panel displays the set of output combinations satisfying *both* the machine and the labor constraint. This feasible region is the intersection of the two triangular regions plotted on the two preceding graphs. The production transformation curve is the border of this feasible region. Note that the production transformation curve is piecewise linear. Also, the existence of the kink at point $\langle 3\ 1/3, 3\ 1/3\rangle$ means that the function constituting this frontier is not differentiable at that point.[10]

Now suppose that our manufacturer wishes to maximize revenue.[11] What quantities of X and Y should be produced if X sells for \$1.00 per unit and Y sells for \$1.00 per unit? To find the optimal level of output first consider revenue at the three points $\langle 0, 5\rangle$, $\langle 5, 0\rangle$, and $\langle 3\ 1/3, 3\ 1/3\rangle$. The first point, which yields the maximum output of X, provides revenue of \$5.00. The second point, which maximizes the output of Y, also yields \$5.00 of revenue. But the vertex (kink point) at $\langle 3\ 1/3, 3\ 1/3\rangle$ is better for it yields revenue of \$6 2/3. Figure 2.6 helps to understand what is going on by adding iso-revenue curves to the right-hand panel of Figure 2.5. Each iso-revenue curve constitutes the locus of all points yielding the same revenue. No points on the \$8.00 iso-revenue curve can be produced; many points on the \$5.00 iso-revenue curve are feasible. Point $\langle 3\ 1/3, 3\ 1/3\rangle$ is feasible and lies on the \$6 2/3 iso-revenue curve. Since the \$6 2/3 iso-revenue curve just touches the production transformation curve only at point $\langle 3\ 1/3, 3\ 1/3\rangle$, graphical intuition suggests that this point yields the revenue maximizing output.

[10] To determine the kink-point note that it must lie on both $2X + 4Y = 20$ and $2X + Y = 10$. Solving these two linear equations for the two unknown coordinates yields $X = Y = 3\ 1/3$.

[11] It is only for simplicity that we assume revenue rather than profit maximization. Revenue maximization might yield the same result as profit maximization if there were no raw materials and the number of workers and machines could not be changed.

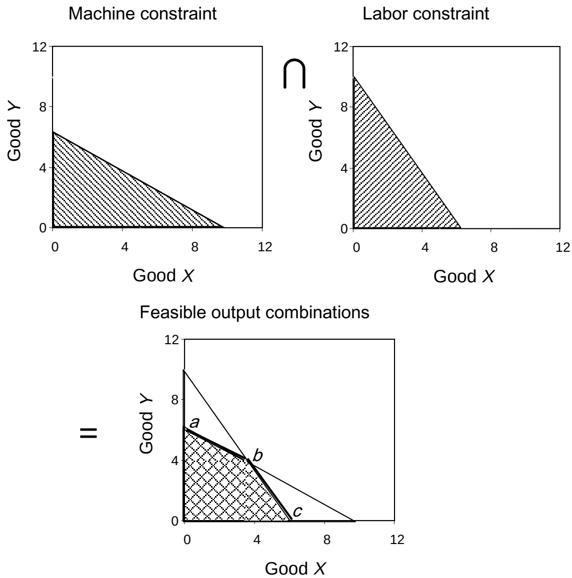


Fig. 2.5. Production transformation curve — linear programming
Only those points in the intersection of the set satisfying the machine inequality constraint with the set satisfying the labor constraint are feasible. Therefore, points a, b and c indicate the production transformation curve.

The revenue maximizing level of output will remain at point $\langle 3\ 1/3, 3\ 1/3 \rangle$ for moderate variation in prices, but suppose the price of Y were to increase to \$3.00 while the price of X remained at \$1. Then that point would now yield \$13 1/3 in revenue, but feasible point $\langle 0, 5 \rangle$ is even better, yielding \$15. Alternatively, if $p_x = \$1$ and $p_y = \$2$, a whole range of points, $0 < X < 3\ 1/3$, would yield maximum revenue because an entire segment of the production transformation curve would be tangent to the iso-revenue curve. Further scrutiny of the graph reveals that, as with Figure 2.3, the revenue maximizing product mix depends only on *relative prices*. Specifically, revenue is maximized with output at $\langle 0, 5 \rangle$ if $p_x/p_y < 1/2$; it will shift to $\langle 5, 0 \rangle$ if $p_x/p_y > 2$. In between these bounds, revenue can be maximized at point $\langle 3\ 1/3, 3\ 1/3 \rangle$. It is easy to see that it is the price ratio that matters because the price ratio determines the slope of the relevant iso-revenue curves. The supply curve for Good Y, plotted on Figure 2.7, provides a convenient graphical summary showing how the quantity that the maximizing business will bring to market depends on p_y, given that $p_x = \$1.00$.

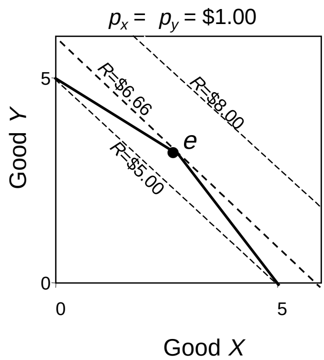


Fig. 2.6. Revenue maximization — linear programming
The iso-revenue curves for $p_x = p_y = \$1.00$ are constructed in exactly the same way as on Figure 2.3. Maximum revenue of $6.66 is achieved at point e. Because the iso-product curve is piecewise linear with a kink at e, the output mix will not shift for moderate price changes.

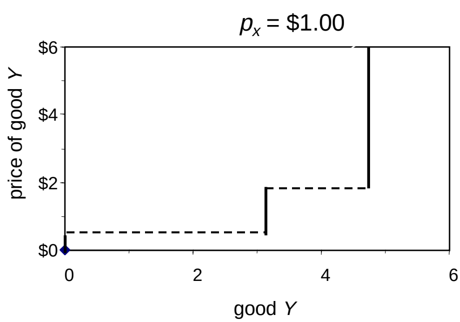


Fig. 2.7. Supply curve for Good Y
The supply curve, derived from the piecewise linear production possibilities curve of Figure 2.6, has steps because the product mix is unaffected by moderate price changes.

This particular example is absurdly simple, but it does illustrate the basic principles of the class of problems know as "linear programs." The linear programming approach has been used for decades by petroleum refineries to determine the optimal product mix. Agribusinesses use the simplex algorithm to solve cost minimization problems, such as determining the cost minimizing mix of available types of feed for chickens, given the restriction that appropriate nutritional constraints must be satisfied.

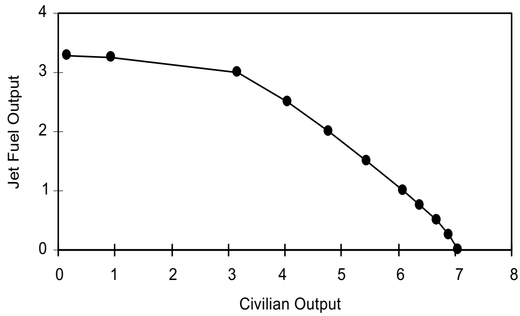


Fig. 2.8. Military versus civilian petroleum output
The alternative output possibilities recorded on Table 2.5 are plotted on the graph.

Table 2.2. U.S. petroleum refining.

(Millions of barrels per day) JP-4 jet fuel	non jet fuel
0.00	7.07
0.26	6.91
0.50	6.68
0.75	6.39
1.00	6.09
1.50	5.45
2.00	4.78
2.50	4.05
3.00	3.16
3.25	0.93
3.29	0.15

2.4.3 *Linear programming — Korean War*

Figure 2.8 shows the possible tradeoffs between the production of jet fuel versus non-jet fuel petroleum products for the U.S. refining industry at the time of the Korean War. The estimates were calculated by Stanford Professor Alan S. Manne utilizing linear programming.[12] The graph of the production transformation curve is composed of linear segments, just as in the case of Figure 2.5, but there are many more kinks because there are

[12] Alan S. Manne, "A linear programming model of the U.S. petroleum refining industry," *Econometrica*, 1957. While dated, it is presented here because it is much simpler than more recent studies of resource allocation in times of conflict.

many more inequality constraints arising from the large number of types of capacity involved in the refining of petroleum.

How can we calculate the opportunity costs from these data? Consider, for example, the trade off between the production of JP-4 jet fuel and non-jet fuel petroleum if the economy were producing 0.1 million barrels of JP-4 Jet fuel. Since the production transformation curve is piece-wise linear, the first step is to calculate the slope of the relevant line-segment. Observe that the change from producing 0 to 0.26 of jet fuel is accompanied by a reduction in non-petroleum output from 7.07 to 6.91. Thus the slope of this segment of the production transformation curve is

$$\frac{0.26 - 0.0}{6.91 - 7.07} = \frac{0.26}{-0.16} = -1.625 \tag{20}$$

Therefore, 1.625 is the marginal rate of transformation involved in producing an additional barrel of non-jet fuel petroleum. It is the opportunity cost of producing an additional barrel of non-jet fuel, which involves the sacrifice of 1.625 barrels of jet fuel.

We must still find the opportunity cost of jet-fuel! If we had plotted jet-fuel on the abscissa and non-jet fuel on the ordinate, the slope of the relationship would have been $1/1.625 = 0.61538$. Thus the marginal rate of transformation in the production of additional *jet fuel* is the reciprocal of the marginal rate of transforming jet fuel into non-jet fuel; i.e., it is 0.61538. This is the opportunity cost of jet-fuel. Note that it is low when relatively few resources are being focused on its production.

It is important to note that the *slope* of the production possibility curve becomes flatter when more jet fuel is produced. For example, suppose the economy were producing 0.3 million barrels of JP-4 Jet fuel. To find the slope of the relevant segment of the production transformation curve observe that the change from producing 0.26 to 0.50 of jet fuel is accompanied by a reduction in non jet fuel output from 6.91 to 6.68. Thus the slope of this segment of the production transformation curve is now

$$\frac{0.50 - 0.26}{6.68 - 6.91} = \frac{0.24}{-0.23} = -1.04348\,. \tag{21}$$

This means that the opportunity cost of producing an additional unit of non-jet fuel petroleum is now 1.04. Its reciprocal, the opportunity cost of producing additional jet fuel, is 0.958.

The key point is that opportunity cost equals the marginal rate of transformation; it equals minus the slope of the production possibility curve, but with a positive sign. The opportunity cost changes when the slope changes. Thus the opportunity cost of jet-fuel increases as we focus more and more of the nation's productive energy on its production. If still more resources are concentrated on the production of jet-fuel its opportunity cost will step upwards, as predicted by the law of increasing opportunity cost (cf. Section 2.3.1).[13]

2.5 The gains from trade

2.5.1 *Introduction*

While it is sometimes said that economists never agree, there is one proposition that has received almost unanimous support from professional economists ever since David Ricardo presented his *Theory of Comparative Advantage* in *The Principles of Political Economy and Taxation*, 1817. Ricardo argued against Mercantilism, the conventional wisdom of his day. Mercantilist doctrine held that in order to realize its full economic potential a country must restrict imports from abroad.

Ricardo argued that protective tariffs and restrictions on the free flow of international trade should be abolished because they do more harm than good:

> Under a system of perfectly free commerce, each country naturally devotes its capital and labor to such employments as are most beneficial to each. This pursuit of individual advantage is admirably connected with the universal good of the whole. By stimulating industry, by rewarding ingenuity, and by using most efficaciously the peculiar powers bestowed by nature, it distributes labor most effectively and most economically: while, by increasing the general mass of productions, it diffuses general benefit, and binds together, by one common tie of interest and intercourse, the universal society of nations throughout the civilized world. It is this principle that determines that wine shall be made in France and Portugal, that corn shall be grown in America and Poland, and that hardware and other goods shall be manufactured in England.

Ricardo's argument is relevant to this day, for the ghosts of Mercantilism haunt the halls of Congress. In the early 1990s, many strenuously objected that the North American Free Trade Agreement (NAFTA) would suck jobs

[13] Since the production possibility curve is piecewise linear, it is more precise to say that the production of jet fuel is subject to the Law of Non-Decreasing Opportunity Cost. Or we might say it obeys the *weak form* of the law of increasing opportunity cost.

out of the American economy to low wage Mexico. After much debate NAFTA was adopted, taking effect at the beginning of 1994. NAFTA provides that tariffs on most goods traded between the United States, Mexico and Canada will be eliminated by 2009.

2.5.2 *Numerical example*

Ricardo skillfully supported his argument in favor of free trade by using a clever numerical example similar in many respects to the following argument. Suppose that there are two countries, number 1 and number 2, each producing two commodities, A and B. Suppose that the production transformation function for Country 1 is

$$A_1^2 + B_1^2 = 25 \tag{22}$$

and that for Country 2 is

$$A_2^2 + B_2^2 = 36\,. \tag{23}$$

The two production transformation curves are displayed on Figure 2.9. Note that Country 2 is clearly more productive than Country 1. Such a difference might arise if there are more workers in Country 2. Even if their populations are comparable in number, Country 2 may be more productive because there is more land or machinery per worker. Or it may be that

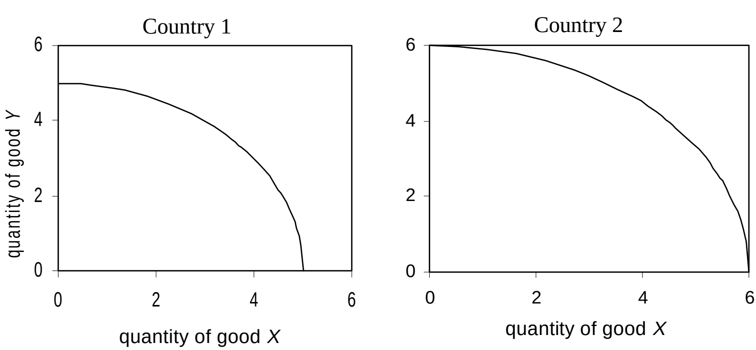


Fig. 2.9. Production possibilities — two nations
Comparing the two production possibility curves reveals that Country 2 is much more productive than Country 1. The discrepancy may arise because there are more workers in Country 2. Or perhaps they work harder or for longer hours, or perhaps the workers have better equipment or supplies or the land is more fertile.

workers are better educated or trained in Country 2. Or perhaps they work longer hours or take fewer holidays.

Would Country 1 or Country 2 gain if trade opened up between these two economies? Because they are more productive, workers in Country 2 may be paid more; they are likely to worry that the opening up of free trade will expose them to competition from the cheap labor in Country 1. At the same time workers in Country 1 will be worrying that they will lose out from the opening up of free trade because they cannot compete with the more productive workers in Country 2. But we will find, in accordance with the argument of Ricardo, that *both* countries may simultaneously gain from trade.

As a first step toward developing the concept of efficiency, it seems reasonable to state:

> An allocation is *efficient* if we cannot reallocate production assignments so as to increase the output of one commodity without reducing the output of at least one other commodity.

An example will illustrate that having each country operating efficiently does not imply that the two countries, considered together, are operating efficiently. Suppose that Country 1 is producing $A_1 = 2.25$ and $B_1 = 4.47$ while Country 2 produces $A_2 = 5.75$ and $B_2 = 1.71$. Thus the total output of commodity A is $2.25+5.75 = 8$ and that of commodity B is $4.47+1.71 = 6.18$, as summarized on Table 2.3.

Country 1 is operating efficiently because it is on its production transformation curve, equation (22): $A_1^2 + B_1^2 = 2.25^2 + 4.47^2 = 25$. And Country 2 is also operating efficiently because $5.75^2 + 1.71^2 = 36$, as required by equation (23). Nevertheless, we shall find that even though each country is operating efficiently it is still possible through an appropriate reallocation of production tasks to increase the combined output of at least one commodity without reducing that of the other!

To show that the initial allocation is inefficient we must find a reallocation of production assignments between the two countries that would lead to

Table 2.3. Initial allocation of production (Allocation #1).

	Output of Good A	Output of Good B	MRT $= -dB/dA$
Country 1	2.25	4.47	0.50
Country 2	5.75	1.71	3.35
TOTAL	8.00	6.18	

Table 2.4. A gain from trade (Allocation #2).

	Output of Good A	Output of Good B	MRT $= -dB/dA$
Country 1	3.00	4.00	0.75
Country 2	5.00	3.32	1.51
TOTAL	8.00	7.32	
After Country 1 trades 3/4ths of a unit of Good A for 3/4ths of a unit of Good B:			
	Consumption of Good A	Consumption of Good B	
Country 1	2.75	4.75	
Country 2	5.75	2.67	
TOTAL	8.00	7.32	

Note that both countries now have the same amount of Good A but more of Good B than they did with Allocation #1.

an increase in total output of at least one commodity and no reduction in the other. Suppose that we set $A_1 = 3$ rather than 2.25; then Country 1 must cut back its production of the other commodity to $B_1 = (25 - 3^2)^{0.5} = 4$, as recorded in the top row of Table 2.4. Now if Country 2 produces 5 units of A_1 it can produce $B_2 = 3.32 = (36 - 5^2)^{0.5}$. With this reallocation of productive effort we still have total output of $3 + 5 = 8$ units of commodity A, just as before. But total output of B is now 7.32, which is obviously better than the 6.18 of before!

The fact that the combined output of the two countries would be larger with the reallocation of Table 2.4 means that there is a potential *gain from trade*! Trade would enable both countries simultaneously to realize an improvement from the reallocation of production. For example, suppose that Country 1 trades 3/4ths of a unit of Good A for 3/4ths of a unit of Good B. Then both countries have the same amount of Good A as with the initial allocation, which was recorded on Table 2.3. But while both are as well off in terms of Good A, with Allocation #2 both have more of Good B than in the initial allocation. Thanks to the reallocation of production assignments coupled with the trade, *both* countries are better off! Clearly, the initial allocation of resources was inefficient, when both countries are considered together, because the reallocation of production assignment made the total output of B larger without any reduction in the total quantity of A.

The fact that production allocation #2 dominates the first does not establish that it is an efficient allocation. Indeed it is not. Table 2.5 presents

Table 2.5. An efficient allocation of resources (Allocation #3).

	Output of Good A	Output of Good B	MRT $= -dB/dA$
Country 1	3.64	3.43	1.06
Country 2	4.36	4.12	1.06
TOTAL	8.00	7.55	

a third feasible allocation that dominates the second in the sense that it has more of at least one good and no reduction in the supply of the other.

While it is clear that Allocation #3 dominates both of the previous allocations, how can we test to determine if it is an efficient allocation of production? That is to say, how can we establish that there is no reallocation of productive assignments between the two countries that would produce even more of at least one commodity without any reduction in the total output of the other? The trick to establishing that the allocation is efficient is to note that only for this third allocation are the marginal rates of transformation (reported in the last column) of the two countries identical. We shall establish the following proposition:

Proposition: Equality of the marginal rates of transformation for the two countries is necessary for efficiency.

Here is one proof of the proposition. An alternative proof is presented as Appendix 2.1.

Let $dA_1 > 0$ be a small increase in the output of commodity A by Country 1. Then the resulting cutback in that country's production of the other commodity must be $dB_1 = dB_1/dA_1 \cdot dA_1$, where dB_1/dA_1 is the slope of Country 1's production transformation curve (i.e., the opportunity cost of producing Good A in Country 1 in terms of the foregone output of B. Similarly, $dB_2 = dB_2/dA_2 \cdot dA_2$. Now suppose we set $dA_2 = -dA_1$, so as to leave the total output of that commodity unchanged. The resulting change in the total output of Good B will be:

$$\begin{aligned} dB &= \frac{dB_1}{dA_1} \cdot dA_1. + \frac{dB_2}{dA_2} \cdot (-dA_1) \\ &= \left(\frac{dB_1}{dA_1.} - \frac{dB_2}{dA_2} \right) \cdot dA_1 \,. \end{aligned} \tag{24}$$

If the expression in parentheses after the second equality were positive, then $dA_1 > 0$ would yield an increase in the supply of B with no reduction in A, implying that the initial allocation was inefficient. Or if the expression in parentheses were negative, an adjustment in the opposite direction, $dA_1 < 0$, would yield an increase in B, which again implies inefficiency. Hence the expression in parentheses must be zero when resources are allocated efficiently. This argument suggests that efficiency requires:

$$\frac{dB_1}{dA_1.} = \frac{dB_2}{dA_2.} \tag{25}$$

That is to say, efficiency requires that the marginal rate of substitution between the two commodities in Country 1 must equal the marginal rate of substitution in Country 2. Or to put it another way, when output is allocated efficiently among countries the opportunity cost of producing commodity A in terms of the foregone output of commodity B is the same everywhere.

Since in competitive markets profit maximizing entrepreneurs adjust output so as to equate the marginal rate of transformation with the price ratios, if free competition among nations, undistorted by tariffs and quotas, makes the price ratio the same in both Country 1 and Country 2, this last condition will be satisfied automatically! Competition is the best of all possible worlds. This argument of Ricardo neglects shipping costs, but a more complicated analysis taking such costs appropriately into account shows that the argument is still valid. The argument also depends on the assumption that the costs of adjustment and transitional unemployment are not too severe and neglects the possible effects of the removal of trade restrictions on the distribution of income within each country.

The Two-Country (World) Production Transformation Curve constituting the set of all efficient points is plotted on Figure 2.10. At each point on this curve the MRT for Country 1 is equal to that for Country 2. For example, point e is obtained by having Country 1 produce at point x. To show this, Country 2's production transformation curve is displaced from the origin to point x in order to reveal the feasible two-country outputs attainable by varying the output of Country 2 while keeping Country 1 producing at point x. At point e Country 2's marginal rate of transformation is the same as Country 1's, as required for efficiency.

2.5.3 *Market forces achieve efficiency*

One might think that the achievement of an efficient allocation of resources would require skilled central economic planners capable of determining the precise mix of production activities needed in each country. More than

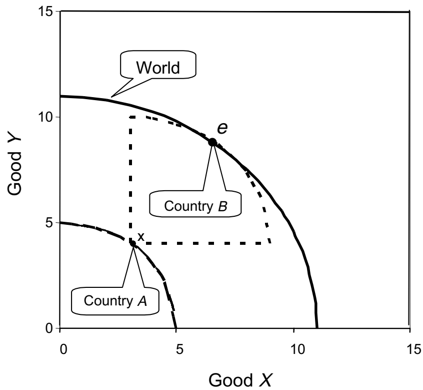


Fig. 2.10. Two-country world production transformation curve
Here is the production transformation curve showing possible output for the two country world that is generated by the production transformation curves for Country A and Country B. Country B's production transformation curve is displaced from the origin in order to show feasible world outputs if Country A produces at point x. It turns out that only tangency point e is efficient, given that A is producing at x. By moving point x and therefore Country B's production transformation curve to the left and to the right along Country A's production transformation curve one can map out additional efficient points. The outer envelop of points developed in this way is the two-country production transformation curve.

this, one might think that implementation of the efficient allocation would require the negotiation of an international agreement. But generations of economists have argued that this is not the case. The free play of natural market forces can lead to the efficient allocation of resources without the need for either an omniscient planner or an international agreement.

Establishing the validity of this proposition is straightforward, at least if we are allowed to make the simplifying assumption that shipping costs are negligible.

Step #1: The first step is to show that market forces will lead to output adjustments that will cause the price ratios in the two countries to adjust to the same value. Suppose, for example, that apples and bananas both sell for \$1 each in Country 1 while in Country 2 apples cost one peso but bananas two pesos. Thus the price ratio p_a/p_b is 1 in Country 1 but 1/2 in Country 2. The fact that the ratios are not equal creates a profit opportunity: Consider a trader in Country 1 who has 100 apples. Our trader can purchase 100 bananas in Country 1 and ship them to Country 2 where he can trade them for 200 apples. Shipping the apples back to

Country 1 yields 200 bananas, or a profit of 100 bananas for the round trip. Our happy trader has taken advantage of an arbitrage opportunity:

> *Arbitrage* is trading for profit without risk. It is the act of buying a commodity or currency in one market and simultaneously selling it for a profit in another market.

Not only has the trader made a profit; the act of arbitrage has moved apples from Country 2, where they are relatively cheap, to Country 1, where they are relatively more expensive. This will tend to drive down the price of apples in Country 1 (p_a^1) and push up their price in Country 2 (p_a^2). Simultaneously, the decreased quantity of bananas in Country 1 will tend to drive up their price (p_b^1), thus leading to a decrease in the price ratio p_a^1/p_b^1. Arbitrageurs will continue to ship as long as the activity is profitable, and it will be profitable as long as $p_a^1/p_b^1 \neq p_a^2/p_b^2$. Because it is profitable, arbitraging activity will continue until equilibrium is achieved with

$$\frac{p_a^1}{p_b^1} = \frac{p_a^2}{p_b^2}\,. \tag{26}$$

Step #2: Now in competitive markets profit maximizing entrepreneurs adjust output so as to equate the marginal rate of transformation with the price ratios. Thus $\mathrm{MRT}^1 = p_a^1/p_b^1$; similarly, in Country 2, $\mathrm{MRT}^2 = p_a^2/p_b^2$. Therefore, free trade among nations (undistorted by tariffs and quotas) will make the relative price ratio in Country 1 equal to that of Country 2, thus assuring that the efficiency condition specified by equation (25) will be satisfied automatically!

This argument neglects transportation costs in the interest of simplicity, but a more complicated analysis taking such costs appropriately into account shows that Ricardo's argument is still valid. And the argument is readily generalized from two commodities and two countries to n commodities and m countries. The argument does depend on the assumption that the costs of adjustment and transitional unemployment are not too severe and neglects the possible effects of the removal of trade restrictions on the distribution of income within each country.

Way back in 1817 Ricardo illustrated the gains from trade with a numerical example similar to our apples and bananas example, but he used wool and wine for his two commodities and England and Portugal as his two trading partners. Ricardo's numerical example demonstrated that international trade does not benefit one country at the expense of the other. Both countries gain from trade. Ricardo drew a fundamental economic moral from his numerical example:

> *Countries that restrict imports and discourage exports lose the benefits of free trade. Nations that impose tariffs on imports are hurting themselves as well as their trading partners.*

Fewer propositions in economics have received as strong an endorsement as has been provided by generations of economists for the doctrine of free trade. But Ricardo's lesson has seldom been put into practice. Those who produce goods that compete with imports will worry that they will lose their jobs. Legislators, ever alert to the special pleadings of their constituents, have largely ignored Ricardo's argument. Time and time again legislators have voted for tariffs and restrictive quotas in an attempt to protect the special interests of domestic producers from foreign competition at the expense of the consumer. Legislators listen to the pleading of their most vocal constituents rather than the arguments of economists.

2.6 Methodological note

This chapter has relied upon a very simple mode of analysis, the production transformation curve, in order to present as clearly and simply as possible certain fundamental principles of economics. Just as Adam Smith had argued in 1776 (recall the quotations in Chapter 1), we found that the behavior of economic agents may be explained by the assumption that individuals pursue their own self-interest. We also found support for Adam Smith's second point: the individual pursuit of self-interest can lead to an efficient allocation of resources. Our mode of analysis was quite primitive, but it paralleled the argument for free international trade put forward by David Ricardo in 1817.

In addition to presenting basic issues, this chapter has also illustrated certain principles of economic analysis. Economists are inclined to develop simplifying models of the phenomenon under study. It can be objected that the analysis of this chapter — like the argument of Ricardo — was much too simple. Albert Einstein is reported to have said that "theory should be as simple as possible, but no simpler." The assumption that only two commodities are produced is obviously unrealistic. The assumption that farmers maximized revenue, neglecting all other costs, was blatantly false.

While the mode of analysis presented in this book will gradually increase in complexity as we progress from chapter to chapter, it cannot be claimed that any of the models are realistic, although some will be more realistic than others. But in defense it can be argued that if it is to help us to understand basic principles, a theory must simplify if it is to clearly reveal the heart of the matter; it must focus on the key issues and neglect the inessential.

To illustrate, let us digress for a moment to consider our knowledge of the ocean's tides: Sir Isaac Newton [1642–1727] laid the foundation for the modern theory of the tides in his *Principia*, published in 1687. Newton simplified the problem by focusing his attention on the behavior of water in a canal circling the earth. He further simplified by neglecting the effect of the earth's rotation upon his canal tide and the friction generated by the tidal movement of water along the ocean's floor. Newton's model certainly was not realistic. It could not explain why the swing from high to low tide is only about three feet in the Mediterranean but more then twenty feet in the Bay of Fundy. Nevertheless, his analysis constituted a major step forward in our understanding of tidal forces. As science advances it becomes capable of explaining much more, analytical techniques are refined, and the theory comes to conform closer and closer to reality. So too in this textbook. This early chapter presented a grossly simplified explanation of basic concepts that will be elaborated upon in subsequent chapters.

Summary

1. The production transformation curve is based on the proposition that increasing the production of one good necessarily involves a reduction in the output of at least one other good, given that only a limited set of resources are available for production in a world of scarcity.
2. The concept of *opportunity cost* recognizes that the cost of acquiring a good is the cost of the sacrificed alternative.
3. When a depressed economy suffers from excessive unemployment, output falls short of its potential and the economy operates below the production possibility curve frontier. Even if all resources are fully employed, the economy will produce below the production transformation curve if resources are being allocated inefficiently.
4. The production transformation curve illustrated what is meant by inefficient resource allocation. Resources are said to be allocated inefficiently if it is possible to increase the output of at least one good without reducing the output of any other good.

5. The production transformation curve was also used to show how the quantity that a maximizing producer will supply of a commodity depends on prices, or more precisely, on the price of that commodity *relative* to the price of other goods. The argument predicted that if all prices doubled, as with uniform inflation, the quantity of goods supplied would be unaffected. It is relative not absolute prices that matter.
6. Ricardo's Principle of Comparative Advantage is the proposition that free trade among nations unimpeded by protective tariffs and quotas leads to an efficient allocation of resources.
7. Arbitrage activity (buying a commodity in one market in order to sell it at a profit in another market) contributes to the attainment of an efficient allocation of resources.

Appendix 2.1. Alternative proof of a fundamental proposition

Consider two producers of two commodities. We shall see that efficiency requires that the marginal rates of transformation between the two commodities be the same for both producers.

Let $Y_1 = T_1(X_1)$ and $Y_2 = T_2(X_2)$ be the production transformation curves for countries one and two. Then the total output of Good Y is

$$Y = Y_1 + Y_2 = T_1(X_1) + T_2(X_2)\,. \tag{27}$$

Now we wish to find the maximum amount of Y that can be produced, given that the output of X is at some fixed level k (e.g., 8); i.e.,

$$X_1 + X_2 = k\,. \tag{28}$$

Since $X_2 = k - X_1$ we have on substituting into equation (27):

$$Y = T_1(X_1) + T_2(k - X_1)\,. \tag{29}$$

Differentiation with respect to X_1 will yield a necessary condition for maximizing Y, given the constraint that $X_1 + X_2 = k$. First, note by the chain rule that the derivative of the second term of (29) is $dT_2(k - X_1)/dX_1 = dT_2/dX_2 d(k - X_1)/dX_1 = -dT_2/dX_2$. Therefore,

$$\frac{dY}{dX_1} = \frac{dT_1}{dX_1} + \frac{dT_2}{dX_2}(-1)\,. \tag{30}$$

Setting this derivative equal to zero yields

$$\frac{dT_1}{dX_1} = \frac{dT_2}{dX_2}\,. \tag{31}$$

Since $dT_1/dX_1 = -\mathrm{MRT}_1$ and $dT_2/dX_2 = -\mathrm{MRT}_2$, a maximum can be attained only if the marginal rate of transformation is the same in the two economies; i.e., efficiency requires[14]

$$\mathrm{MRT}_1 = \mathrm{MRT}_2\,. \tag{32}$$

Key Concepts

Exercises

1. Consider the following production transformation curve: $x + y^2 = 25$; $x \geq 0$, $y \geq 0$.
   a. Determine the output of y if $x = 0$, 9, 16, 21, and 25.
   b. Plot the production transformation curve.
   c. Determine the output of y as a function of x.
   d. Estimate from your graph the marginal rate of transformation when $x = 9$?
   e. Determine the marginal rate of transformation as a function of x.

   Note: You should treat x as a continuous variable, like milk, so you can solve this problem with the calculus; do not restrict your answers to integer values.

[14]The second order condition for a maximum will be satisfied, thanks to the law of increasing opportunity cost.

2. Suppose that the production possibilities for Goods x and y are specified by $x^2 + y^2 = 16$, $x > 0$, $y \geq 0$.
   a. Plot the production transformation curve.
   b. Find the equation for y as a function of x. Then find dy/dx.
   c. Determine the marginal rate of transformation and the opportunity cost of x (in terms of foregone y output) when $x = 2$.
   d. Determine the quantity of x and y that will maximize total revenue if $p_x = p_y = \$1.00$.
   e. Determine the quantity of x and y that will maximize total revenue if $p_x = p_y = \$3.00$.
   f. Derive the supply curve for x. That is to say, find the quantity of x produced as a function of p_x/p_y (the relative price of x).
3. In his study of the supply of jet fuel for the Korean War, Allan Mann suggested that enemy action might shift the production transformation curve as illustrated on Table 2.6.

Table 2.6. Reduced refining capacity.

	JP-4 jet fuel (Millions of barrels per day)	Petroleum-non jet fuel (Millions of barrels per day)
a	0.00	3.28
b	0.50	3.17
c	1.00	2.58
d	1.50	1.65
e	1.64	1.00
f	1.71	0.38

   a. Plot the reduced capacity data points on a sheet of graph paper or with a computer spreadsheet, forming a production possibilities curve with non-jet (civilian) fuel on the abscissa and jet fuel on the ordinate. [Suggestions: Label the axes of your graph carefully; mark each data point with the appropriate letter — a, b, c, etc.; connect adjacent points with straight lines.]
   b. If 3 million barrels of petroleum non-jet fuel are being produced, how much JP-4 Jet fuel can be refined? What is the MRT involved in producing an additional barrel of jet fuel when the output of non-jet fuel is 3.00 million barrels per day? What is the opportunity cost of an additional barrel of jet fuel?
   c. What is the opportunity cost of an additional barrel of civilian output (non-jet fuel) in terms of foregone barrels of jet fuel when

civilian output is 2 million barrels? Explain how the opportunity cost of civilian fuel is related to the opportunity cost of jet fuel and the slope of the production transformation curve.

4. Congratulations, you have inherited a boat and $1,000 from your late Uncle Jake. In the United States corn sells for $1.00 per bushel and wheat for $2.00 a bushel. In England, corn and wheat both sell for one British pound a bushel.

   a. Explain how you can double your $1,000 through an artful arbitrage operation, assuming that shipping costs are negligible.
   b. If a large number of traders attempt to profit in this way, will the opportunity for profit be diminished? Why?

   Note: You do *not* have to know the foreign exchange rate in order to answer this question.

5. You and a shipmate are stranded on a deserted island in the Pacific. You can each produce two goods, coconuts and fish. Your production transformation function is $f^2/4 + c^2 = 10,000$. Your shipmate's is $f^2 + c/4 = 10,000$.

   a. Graph the two production possibility frontiers on separate graphs.
   b. If you produce 100 fish, what is the maximum number of coconuts you can produce? If your shipmate also produces 100 fish what is the maximum attainable output of coconuts.
   c. Is there a way of increasing the total output of coconuts without reducing your combined output of fish below 200? Explain.
   d. If the price of fish is $2.00 and the price of coconuts is $1.00, how many fish and coconuts would you produce? How many would your shipmate produce of each commodity? Would the resulting output be efficient?
   e. Determine how many fish you would supply as a function of the price of fish, given that coconuts sell for $1.00.

6.* In his 1817 treatise, *The Principles of Political Economy and Taxation*, David Ricardo presented the following numerical example, which is similar to what is now known as a *linear programming* problem.

> "Suppose that 80 English workers can produce 80 barrels of wine or 100 bolts of cloth in one year but that in Portugal 60 workers can produce 120 barrels of wine or 60 bolts of cloth. In England they are producing 40 barrels of wine and 50 bolts of cloth while in Portugal they produce 60 barrels of wine and 30 bolts of cloth."

Assuming, as did Ricardo, that the relationship is linear, the production transformation curve with 80 English workers is the function $W_E = T_E(C_E) = 80 - 0.8C_E$. This implies that the opportunity cost of a bolt of cloth in England is 0.8 barrels of wine. (Therefore, the opportunity cost of a barrel of wine is $1/0.8 = 1.25$ bolts of cloth.)

a. Plot the production transformation curve for England.
b. Determine the production transformation curve for Portugal's 60 workers; plot it on a separate graph.
c. If both countries concentrate on producing only wool, what is the maximum quantity they can produce? If they concentrate in both countries on producing only wine, how much can they produce (assuming workers stay sober)?
d. Determine a price ratio for wine and wool in England (p_w^E/p_c^E) that would induce maximizing British farmers to produce both wine and cloth. Is England producing efficiently (i.e., is it operating on its production possibility frontier)? What price ratio (p_w^P/p_c^P) must prevail in Portugal if maximizing producers are to decide to produce both wine and cloth? Is Portugal operating efficiently?
e. Suppose you have 10 bolts of cloth in your London warehouse. The price ratios you determined in *d* prevail in both England and Portugal. How much wine can your cloth be exchanged for in England? Would you do better to ship your cloth to Portugal, buy wine there, and bring it home to sell it in England. How much profit would you make from the round trip?
   Hint: You will find it helpful to consider the four feasible allocations on Table 2.7.
f. Determine the production possibility function for the combined output of England and Portugal; plot it on a graph. (It will have two liner segments, like the bottom panel of Figure 2.5.)

Table 2.7. Four alternative allocations.

	England		Portugal		Total	
	Cloth	Wine	Cloth	Wine	Cloth	Wine
#1	100	0	60	0	160	0
#2	0	80	0	120	–	200
#3	100	0	0	120	100	120
#4	0	–	60	–	60	–

Implications

Ricardo would say that England has a "comparative advantage" in the production of cloth because the opportunity cost of producing cloth in England (0.8 barrels of wine) is less than the opportunity cost of producing cloth in Portugal (2 barrels of wine). Because the opportunity cost of producing wine is the reciprocal of the opportunity cost of producing cloth, Portugal must have a comparative advantage in producing wine.

It would be inefficient to have England producing any wine if Portugal is producing any cloth. Why? Because if England is producing some wine, she could get an extra bolt of cloth by cutting wine output by 0.8 of a barrel. Then Portugal could cut cloth output by 1 bolt and generate 2 barrels of wine. Thus the world would have just as much cloth and $-0.8 + 2 = 1.2$ more barrels of wine!

Ricardo's point was that trade can benefit two countries that are producing the same two commodities unless the opportunity costs are the same in both countries. He was implicitly assuming that both countries were operating on their production possibility frontier (i.e., efficiency and full employment).

7.# Honors Option: A Two-Country World

Country 1 has production transformation curve $A_1^2 + B_1^2 = 25$.

Country 2 has production transformation curve $A_2^2 + B_2^2 = 36$.

Will their combined (two-country world) output be $A = [25-(5/11)B^2]^{1/2} + [36-(6/11)B^2]^{1/2}$? Explain.

Hint: Invoke the efficiency proposition, equation (25).

8.# Honors Option: Another Two-Country World

Country 1 has the following production transformation curve: $x_1^2 + y_1^2 = 4$.

Country 2 has the following production transformation curve: $x_2^2 + y_2^2 = 16$.

Find the production transformation curve for the combined output of the two countries.

3

Supply and Demand: Where do Prices come from?

3.1 Overview

Where do prices come from? Why do they fluctuate? While we found in the preceding chapter that prices can guide production decisions, we did not learn what determines prices. The task of this chapter is to find out how markets work to determine price and output.

Here are some of the issues explored in this chapter: How do crop failures, taxes, subsidies, and government price controls influence prices? Why are shortages the natural but unintended consequence when governments impose price controls in an attempt to ease the pain of rising prices? Why do price floors generate surpluses? Does the minimum wage cause unemployment?

We shall learn how to measure the responsiveness of quantity consumed to changes in income and prices. We shall also learn how to measure the excess burden that taxes impose on producers and consumers.

Caveat: The discussion in this chapter will be limited to the case of *competition*. That is to say, we shall be assuming that there are a large number of buyers and sellers. Later, Chapter 6 will analyze a variety of alternative market forms, including monopoly.

3.2 The Middletown housing market, a parable

To start with the simplest possible example, consider the market for rental housing in Middletown. To keep things simple, suppose that there are 1,000 identical housing units. The landlords would obviously like to collect as much rent as possible. Tenants prefer to pay low rent. How will these opposing interests be reconciled?

What the rent will turn out to be depends in part on how much consumers are willing to spend. If the rent is low enough, many will want their own place. At a higher rent, people will double up (students will go home to live with their parents after they graduate from college). And if the rent is extremely high, everyone will sleep in the park. Suppose that a market survey reveals the following facts about the demand for housing in Middletown:

- No one is willing to spend more than $900 per month to rent an apartment.
- Five hundred people are willing to spend $600 or more to rent an apartment.

• If the rent were zero, 1,500 families would want to occupy apartments

The evidence provided by the survey is plotted in Figure 3.1. The rent (price) is on the ordinate and the number of apartments that are rented is plotted on the abscissa.[1] The three observations provided by the survey are plotted at $\langle 0, \$900\rangle$, $\langle 500, \$600\rangle$, and $\langle 1500, 0\rangle$. The straight line connecting the hypothetical data points provides a reasonable guess as to how the number of apartments people want to rent — the demand for housing — is related to the price, everything else constant. This is the demand curve.

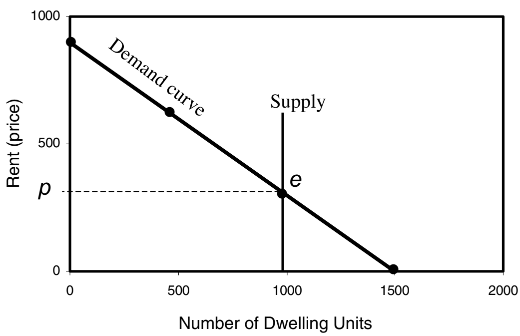


Fig. 3.1. The market for rental housing
The demand curve shows the quantity that people want to buy as a function of price. The equilibrium point, where demand equals supply, is marked with an *e*. This intersection point determines both price and quantity.

This relationship between quantity and price is captured by the following *demand function*,

$$q = 1500 - \frac{5}{3}p\,, \tag{1}$$

for p in the range $0 \leqslant p \leqslant 900$. Note that at a price of $p = \$900$ no one will rent an apartment — this point on the graph, the y-intercept, is known as the choke point.

[1]Economists customarily plot price on the ordinate and quantity on the abscissa. It was not always so: pioneering French mathematical economist Antoine A. Cournot [1783–1850] plotted quantity on the ordinate and price on the abscissa in his *Principes Mathématiques de la Théorie des Richesses*. But rightly or wrongly, the reverse convention is now firmly established.

The fixed supply of only 1,000 housing units is indicated by the vertical line on our graph. What price will prevail in this market?

- Could the rent be \$450 per month? Equation (1) tells us that at this rent consumers would only want to rent 750 of the identical housing units. Landlords would be stuck with 250 empty apartments. Rather than let an empty apartment earn zero rent, any landlord who did not succeed in renting an apartment for \$450 will have every incentive to cut the price, say to \$440 per month. This lower price will either attract a new tenant or steal a renter from another landlord who was getting \$450.
- Even if all landlords succumb to the market pressure and lower the price to \$440, there will still be excess supply. At that price only about 767 apartments will be rented — there will be 233 vacant apartments earning zero rent. There will be pressure for still more price concessions.

The downward spiral will continue, but not indefinitely.

- At a price of \$300 the resulting demand for 1,000 housing units will precisely equal the supply.

The tenants may hope that the price will be pushed below \$300. To see why this cannot happen, let us suppose the contrary. If, for example, the prevailing price were \$150, the graph tells us that there would be demand for 1,250 apartments; but only 1,000 are available. Some of these 1,250 potential renters, finding that they cannot rent for \$150, will offer to pay a higher rent. Landlords, tempted by the higher offer, will evict their present tenants, unless they agree to match the outsider's offer. This holds for any price below that which equates demand with supply, because such a price generates excess demand. Thus the only price that can prevail in this market is \$300. At this rent, the demand for housing precisely equals the supply. We say that a *market equilibrium* has been established when the quantity demanded equals supply because at this unique price there is no inherent tendency for the price to change.

Urban Renewal

The equilibrium price may fluctuate as a result of changes external to the market. Suppose, for example, that the City Planners decide that the time is ripe for urban renewal. Suppose that the government purchases 250 apartments from their owners and tears them down in order to convert the vacated land into a public park. How will this affect the housing market? Figure 3.2 shows what happens. The supply curve is shifted over to the left to 1,000 − 250 = 750 dwelling units. As a result, there would be a shortage of housing if the price of $300 were to prevail. The market is no longer in equilibrium with a price of $300. The market forces respond, rents are bid up, and eventually equilibrium is restored to the housing market at point e' on the graph where demand once more equals supply, and this yields a higher rent of $450.

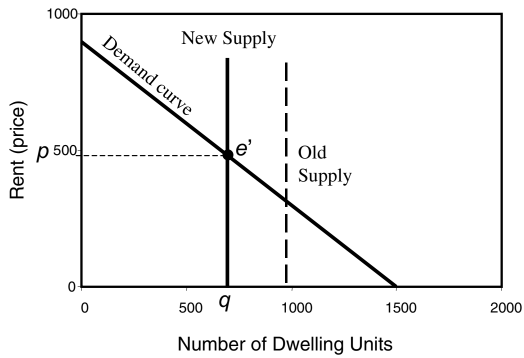


Fig. 3.2. Reduced supply
When supply is reduced to 750 dwelling units, the equilibrium point slides up the demand curve. Price rises to the point where demand is reduced to 750 dwelling units at new equilibrium point e'.

Who gains and who loses from the reduction in housing? The landlords who sold out were compensated when the town paid for the apartments. The landlords who retained their apartments are earning more rent. But the poor tenants are paying $450 rent on 750 apartments, for a total of $337,500. Before they were paying $300 rent for 1000 apartments, or $300,000! They are paying more money for less housing!

Studying the revenue rectangle demarcated by points p, e', q, 0 on Figure 3.2 will show why this happens. The area of this rectangle is height times width, or rent of $450 times quantity of 750; thus the area of this revenue rectangle represents the total rent that landlords collect after the reduction in supply. It is obviously larger than the area of the revenue rectangle generated by the old equilibrium point e with coordinates (1000, $300) on Figure 3.1. Note that a further reduction in the number of apartments, say to 500, would cause rent to climb to $600 but would reduce the total rent collected by landlords to $300,000.

Rent Control

Landlords are never popular. Suppose that the tenants persuade the town planners that the landlords are profiteering from urban renewal by charging exorbitant rents. Suppose that rent controls are enacted rolling rents back to the old level of $300. Now there are 1,000 people trying to rent 750 apartments. There is a housing shortage. Those who are lucky (or have connections or pay bribes) will get an apartment; others must do without.[2]

While this parable helps to provide an intuitive introduction to how markets work, it is obviously a gross simplification of housing markets in practice. In order to focus on the essential features of the market it was appropriate to assume that all apartments are the same.[3] Our analysis did not allow for the fact that supply of housing will eventually respond to changes in price. If rents rise, landlords may find that it has become profitable to construct new housing. If rents are kept too low, either by rent control or by adverse market conditions, landlords will lose money. They would like to sell out, but will find that no one else will want to take over their money losing property. Stuck with their loses, the landlords are all too likely to let their apartments decay. They fall behind in paying their property taxes to the city. If things get too bad the landlord will miss payments on the mortgage, and the bank will threaten to foreclose on the delinquent landlord and take possession of the building. Eventually the building may be abandoned. Thus housing deterioration may at times be the unintended consequence of rent control.

Key Concepts

While the story of the rental market is a simple one, it does serve to illustrate certain basic concepts that are of fundamental importance in understanding how markets work:

---

[2]Rent regulation has at times been a major issue in New York City politics. Tenants who lived in an apartment constructed before 1947 were protected by the regulation if they or a close relative had occupied the apartment since 1971. One of the beneficiaries was film star Mia Farrow, who paid only a fraction of the market rent for a ten room apartment on Central Park West. She had inherited the lease on the rent controlled apartment from her mother.

[3]The model of Monopolistic Competition presented in Chapter 6 will consider markets characterized by product differentiation, such as a rental market in which apartments have their own individualizing features.

- The *demand function* shows how much customers will buy as a function of price. The plot of this relationship on a graph is the *demand curve*.
- The *equilibrium price* equates demand with supply.
- If the price is above the equilibrium price, the quantity supplied will exceed demand; i.e., there will be *excess supply*, and landlords will complain that there is a housing surplus.
- If the price is below the equilibrium price, the quantity supplied will fall short of demand; i.e., there will be *excess demand*; and tenants will complain that there is a housing shortage.
- *Price ceilings* (e.g., rent control) will generate a shortage, supply falling short of demand.

The Law of Demand: Our conclusion that a reduction in supply generated by urban renewal would lead to an increase in rents relied on the assumption that the demand curve was downward sloping; otherwise, the reduction in supply might have led to a fall in rents! Economists generally believe that this is almost always the case. Indeed, the proposition that demand curves are downward sloping, that demand is a negative function of price, is regarded as so fundamental that it is often referred to as the *Law of Demand*. While this proposition is accepted as a general rule, there may be exceptions:

1. Price is sometimes regarded as an indicator of quality. If the price of something is so low as to sound "almost too good to be true," it probably isn't. Demand will drop off if an extremely low price signals, rightly or wrongly, that the item may be shoddy, defective or stolen.
2. A price increase today may be a harbinger of more in the future. If prices are going to rise even higher, it is best to purchase more today.
3. A high price may generate snob appeal. Who, other than an accountant or an economist, would want to be seen wearing a $10 watch!

All of these examples are exceptions to the rule that demand curves almost always have a negative slope.[4]

[4]In addition to these three reasons we will find in Chapter 4 that under certain circumstances a rational consumer might in rare circumstances respond to rising prices by increasing consumption.

3.3 The Econoland corn market

A second hypothetical example will provide additional insight into how markets work. The Econoland corn market is composed of 100 sellers (farmers) and 200 buyers (consumers). All 200 consumers are identical. The quantity of corn (number of bushels) that each consumer will buy, given the price of other commodities, is specified by the following demand schedule:

Demand:

price/bushel	\$18	\$16	\$14	\$12	\$10	\$8	\$6	\$4	\$2	\$0
quantity demanded	0	1	2	3	4	5	6	7	8	9

The behavior of the typical consumer (let us say the ith consumer) is described by the demand function

$$q_i(p) = 9 - \frac{p}{2}, \quad 0 \le p \le \$18\,. \tag{2}$$

Suppose also that all 100 producers are identical. The quantity that the representative producer is willing to market, given the price of seed and other inputs and the price of alternative uses of his corn (e.g., hog production), is given by the following supply schedule:

Supply:

price/bushel	\$0	\$2	\$4	\$6	\$8	\$10	\$12	\$14	\$16	\$18
quantity supplied	0	0	6	12	18	24	30	36	42	48

The typical producer (let us say the jth producer) has supply function

$$s_j(p) = 3p - 6\,, \quad p \ge 2\,. \tag{3}$$

In order to determine the price that will prevail in this market it is necessary to construct *market* demand and supply schedules. To see how, observe that if the price were \$8, each of the 200 consumers would want to purchase 5 units, so total demand for the entire market would be 1,000; thus the point $(1000, \$8)$ is on the market demand curve. Also, at the price of \$8 each of the 100 suppliers would produce 18 bushels of corn, so the point $(1800, \$8)$ is on the market supply curve. Successive rows of the following schedule report market demand, market supply and excess demand, given that there are 200 consumers and 100 farmers:

The market:

Price	0	\$2	\$4	\$6	\$8	\$10	\$12	\$14	\$16	\$18
Market Demand	1800	1600	1400	1200	1000	800	600	400	200	0
Market Supply	0	0	600	1200	1800	2400	3000	3600	4200	4800
Excess Demand	1800	1600	800	0	−800	−1600	−2400	−3200	−4000	−4800

It is clear from the schedule that the equilibrium price is \$6.00. At any other price, demand does not equal supply; i.e., excess demand is not zero. Observe also that excess supply (quantity supplied exceeding quantity demanded) is the same thing as negative excess demand (quantity demanded exceeding quantity supplied) and vice versa.

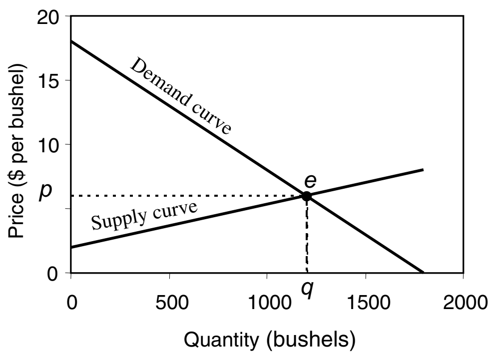


Fig. 3.3. The corn market
The equilibrium price p and quantity q are determined by equilibrium point e where the demand and supply curve intersect.

The same story is illustrated graphically in Figure 3.3. Point e, where the market demand and supply curves intersect, reveals the equilibrium price and quantity. This is the only price at which excess demand is zero. A higher price could not prevail because it would elicit the production of more corn than consumers would be willing to buy — excess supply would tend to push the price down. A lower price could not prevail because it would encourage consumers to try to buy more than producers would be willing to bring to market — excess demand would put upward pressure on prices.

It is easy to solve for the market solution algebraically rather than graphically. Since there are 200 consumers, each with demand function

$q_i(p) = 9 - p/2$, the market demand function must be

$$q(p) = 200q_i(p) = 1800 - 100p\,, \quad 0 \leqslant p \leqslant \$18\,. \tag{4}$$

Notation convention: q_i refers to the consumption of the typical (ith) consumer; q without the subscript refers to demand for the entire market.

Similarly, since supply is $s_i(p) = 3p - 6$, $p \geq \$2$, for each of 100 suppliers, the market supply function must be

$$s(p) = 100s_i(p) = 300p - 600\,, \quad p \geq \$2\,. \tag{5}$$

In equilibrium, $s(p) = q(p)$ or $300p - 600 = 1800 - 100p$. Therefore, the equilibrium price is $p = \$6$ because $q(\$6) = s(\$6) = 1,200$ bushels.

This corny example was simplified by assuming that all consumers have the same demand function. More generally, suppose there are n buyers and m sellers. Let $q_i(p)$ denote the amount of corn that the ith consumer will want to purchase as a function of its price; then *market demand* will be

$$q(p) = q_1(p) + q_2(p) + \cdots + q_n(p)\,. \tag{6}$$

Similarly, if there are m sellers with supply functions $s_i(p)$, *market supply* will be

$$s(p) = s_1(p) + s_2(p) + \cdots + s_m(p)\,. \tag{7}$$

Clearly, at the price for which $q(p) = s(p)$, we must have

$$q_1(p) + q_2(p) + \cdots + q_n(p) = s_1(p) + \cdots + s_m(p)\,. \tag{8}$$

While it is usually most convenient to think of quantity sold as a function of price, it is sometimes useful to think of the *inverse demand function*, $p(q) = q^{-1}(p)$. For our corn example we had $q(p) = 1800 - 100p$ and so the inverse demand function is

$$p(q) = q^{-1}(p) = 18 - \frac{q}{100}\,. \tag{9}$$

Summary

The equilibrium market price has been determined by the free play of market forces. Neither the buyers nor the sellers have been able to dictate the price. They are price takers rather than price setters. They have to take the price offered by the market.

The *market equilibrium* has the following features:

- All buyers are purchasing the quantity they wish to purchase at the prevailing market price.
- All sellers are selling the quantity they wish to sell at the prevailing market price.
- Both buyers and sellers are *price takers.* Price takers must accept the price determined in the market. Later we will study *price setters* — a buyer or seller who is big enough to be able to influence the market price.

If the farmers could organize to keep some of their product off the market, they might be able to push the price up. Attempts by groups of farmers to act as price setters have been made from time to time, but except when supported by government intervention, attempts by a large group of suppliers to fight the market are almost always overwhelmed by the forces of competition.

If the government had intervened, as with price controls, things would obviously have worked out differently. As we shall see in a later chapter, the outcome will be quite different if there are a small number of buyers or a small number of sellers, for then they may be able to work deliberately together to affect the price.

3.3.1 *Crop failure*

Prices fluctuate in response to a variety of market forces. As our first example, suppose that in the spring each farmer, in anticipation of a price of \$6, plants 12 bushels of corn. But then as a result of a July blight the crop falls short and each of the 100 farmers will be able to bring only 7.5 bushels to market at harvest time. Since it is much too late in the season to even think about replanting, only 750 bushels of corn will be available in the fall. What happens is illustrated in Figure 3.4. There is a new "short run" supply curve showing that, regardless of price, only 750 bushels of corn will arrive on the market. As a result of this "negative supply shock," the new market equilibrium price will climb to \$10.50 per bushel, as indicated by point e' on the graph. Alternatively, substituting the supply of 750 into our inverse demand function, equation (9), yields $p = \$10.50$.

It may not be surprising to find that the limited supply of corn has caused the price to go up, but note what has happened to revenue. Before the farmers were selling 1,200 bushels at a price of \$6.00, taking home $\$6 \times 1200 = \$7{,}200$ from the market. Thanks to the crop failure, they now

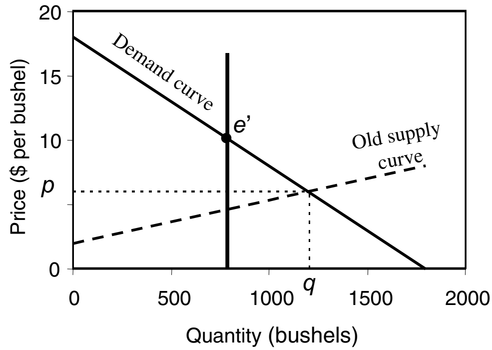


Fig. 3.4. Crop failure
When crops fail and only 750 bushels can be brought to market, the price rises to $10.50 in order to equate demand with the reduced supply at the new equilibrium point.

bring $750 \times \$10.50 = \$7{,}875$ home from the market, $675 more than before! The seeming paradox that a general crop failure can benefit farmers will be clarified later in this chapter when we discuss the concept of demand elasticity.

3.3.2 *Consumer surplus*

How much have the consumers lost as a result of the crop failure? To find out the dollar value of the loss imposed on consumers, we must introduce the concept of consumer surplus. The basic notion of consumer surplus is that the purchasers of a commodity often pay less for that commodity than it is worth to them. An individual buyer's *consumer surplus* is the excess of what that customer would have been willing to pay for a commodity over what it actually costs. For the entire market, we must add up the consumer surplus of each of the individual buyers of the commodity.

The consumer surplus concept is illustrated in Figure 3.5, which shows the hypothetical demand curve for a good book. There are 20 potential purchasers, but some value the book much more than others. The staircase form of the hypothetical demand curve arises because one customer is willing to pay as much as $100 for the book, the second is willing to pay $95 but no more, the third $90 and so on. To determine how much consumer surplus will be realized if the book sells for $60, note that the book loving customer who would have been willing to pay as much as $100 only had to pay $60. The excess of $40 that this happy consumer would have been

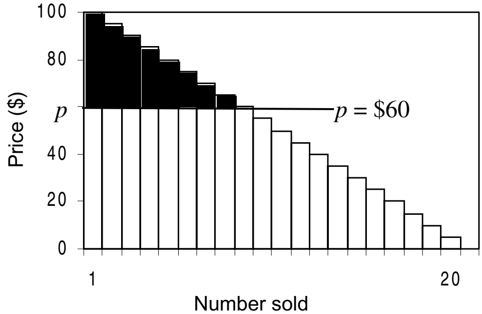


Fig. 3.5. Consumer surplus
Consumer surplus from purchasing a good book for \$60 is measured by the area between the demand curve and the \$60 price line.

willing to pay over what she actually paid for it is her consumer surplus. The graph also shows that two copies of the book would have been sold if the price were \$95, one to our first consumer and one to a second consumer for whom the book is worth \$95 but not \$100. If this second consumer is able to buy the book for \$60 he will realize consumer surplus of $\$95 - \$60 = \$45$. The graph indicates that the first eight buyers of the book all enjoy some consumer surplus, but in decreasing amounts. The buyer of the ninth copy of the book does not enjoy any consumer surplus because this consumer would not be willing to pay a penny more than the \$60 price for the book. The total consumer surplus realized by all the purchasers, represented by the crosshatched area on the graph under the demand curve staircase and above the horizontal price line, is \$180.

When there are a large number of consumers, each buying different quantities, it is natural to draw the demand curve as a downward sloping curve or line rather than as a series of steps, but the logic of consumer surplus remains the same. In Figure 3.6, which reproduces the corn market demand and supply curves from Figure 3.4, the area of the large shaded triangle with vertices d, e, p^e represents the consumer surplus when $p = \$6$. The area is obviously $(\$18 - \$6) \times 1200/2 = \$7{,}200$.

Quite generally, whether the demand curve is linear or not, we can calculate consumer surplus by integration. Let $p(q) = q^{-1}(p)$ denote the inverse demand function yielding the market price as a function of quantity sold. Then the consumer surplus at price p is

$$S_c(p) = \int_0^{q(p)} [\,p(q) - p]dq\,. \tag{10}$$

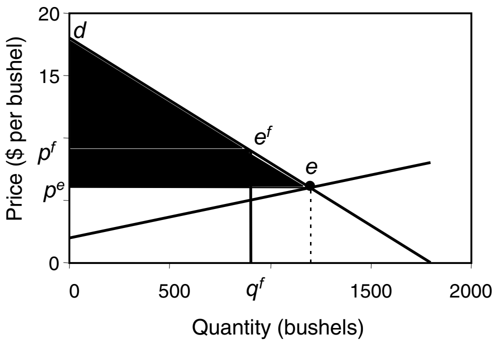


Fig. 3.6. Consumer surplus lost: Crop failure
In a good crop year, $p = \$6$ and consumer surplus is the large shaded triangle with vertexes d, e, p . When crop failure reduces the supply to 900, the price increases to \$9.00 and consumer surplus is only the smaller triangle with vertices at d, e , and p . Therefore, the consumer surplus lost as a result of the crop failure is the trapezoid with coordinates p , e , e, p .

For our hypothetical corn market example we had inverse demand function $p(q) = 18 - q/100$ and $p^e = \$6$; therefore, the surplus that would be realized if there were no crop failure is

$$S_{c(\$6.00)} = \int_0^{1200} \left(18 - \frac{q}{100} - 6\right) dq = 12q - \frac{q^2}{200}\bigg|_0^{1200} = \$7{,}200\,, \qquad (11)$$

just as the graph suggested.

Now consider the effect of the crop failure that was displayed in Figure 3.4. With output of only 750, the price must rise to \$10.50 in order for demand to equal the reduced supply. As a result of the price rise, the consumer surplus triangle is reduced to $(18 - 10.50) \times 750/2 = \$2{,}812.50$. Or by integration,

$$S_{c(\$10.50)} = \int_0^{750} \left(18 - \frac{q}{100} - 10.5\right) dq = 7.5q - \frac{q^2}{200}\bigg|_0^{750} = \$2{,}812{,}50\,. \qquad (12)$$

The consumers lose \$7,200 − \$2,812.50 = \$4,387.50 as a result of the crop failure. Since producers gained an offsetting \$675, the net dollar loss from the crop failure is \$4,387.50 − \$675 = \$3,712.50.

3.3.3 *Price controls and rationing*

Price hikes are never popular. Suppose the government decides to help consumers by imposing a price ceiling of $6.00 per bushel. This will not lead to any more corn being brought to market. Worse, the controls may discourage farmers from planting much corn next year.

At the price of $6.00, consumers will want to buy 1,200 bushels but supply will still only be 750. The price cap has generated a shortage of 300 bushels, just as rent controls generate a shortage of housing. Who will get the corn? The grocer may attempt to fairly allocate the corn to regular customers, friends and family. Or the grocer may accept a side payment under the table. Alternatively, the government may ration corn by issuing coupons to each household. In order to buy corn you have to surrender a ration coupon to the seller.

During World War II sugar, gasoline, coffee, meat and a number of other commodities were rationed in the United States. Just counting the coupons was a lot of work. Before too long, sizable portions of the rationed commodities were illegally diverted to the *black market* to be sold at the highest attainable price. Black markets are illegal, but they seldom lack customers. The system of price controls and rationing may explain why there was not more inflation during World War II, but prices soared when the controls were removed shortly after the end of the war. Perhaps the controls only postponed the inflation that was the inevitable consequence of an all out war effort.

3.4 Demand and supply curve shifters

A crop failure is only one type of disturbance that can affect the market equilibrium. We shall consider other factors, some of which shift the demand curve while others shift the supply curve.

3.4.1 *Some demand curve shifters*

First we shall consider factors that shift the demand curve:

1. When the economy moves into a boom, consumers have more income to spend. They buy more. This will shift the demand curve for most commodities to the right.

So called *inferior goods* are the exception to this rule. If the income increase induces consumers to purchase less corn (perhaps they now buy more meat instead), then corn is an *inferior* good.

2. A farmer's cooperative might stimulate sales by advertising its product.
3. An increase in the price of barley might encourage consumers to *substitute* corn for barley; i.e., consumers may buy more corn instead of barley when the price of barley goes up. This means that the increase in the price of barley has shifted the demand curve for corn upward and to the right, contributing to a rise in the price of corn.
4. An increase in the price of butter might discourage the consumption of corn. If so the demand curve for corn would shift downward and to the left, contributing to a decline in the price of corn.

Substitutes and Complements

Barley is a substitute for corn while butter is a complement.

- The price of *complementary commodities* (like corn and butter or strawberries and cream) tend to move in opposite directions. Thus a poor strawberry harvest may contribute to a decline in the demand for cream.
- The prices of *substitute commodities* (like corn and barley or raspberries and strawberries) tend to move together.

Sliding Along the Supply curve: Demand curve shifters cause the equilibrium point to slide along a stable supply curve.

- If the demand curve shifts up or to the right, the equilibrium point will slide up the supply curve, which causes both quantity and price to increase.
- If the demand curve shifts to the left, perhaps because the economy has moved into a recession, the equilibrium point will slide down the supply curve and both quantity and price will decline.

3.4.2 *Some supply curve shifters*

Some types of disturbances shift the supply curve rather than the demand curve:

- A fall in the price of hogs may encourage farmers to bring more corn to market rather than feed it to their pigs. This will push the market supply curve of corn to the right, the equilibrium point will slide down the stable demand curve for corn. As a result, the price of corn falls but the quantity sold increases.

- An increase in the price of diesel fuel will make it more expensive for farmers to run their trucks and tractors. At any given price the farmer is likely to bring less to market. Some farmers may go out of business. All this means that the supply curve shifts upwards. The equilibrium point will slide up the demand curve and quantity sold will drop off.

3.4.3 *Recapitulation*

The demand and supply curve apparatus provides a convenient framework for examining the effect of a variety of market disturbances. The trick in putting this apparatus to work is to decide whether the disturbance affects the position of the demand curve, the position of the supply curve, or both.

There is a key difference between demand and supply curve shifters: Demand curve shifters cause price and quantity to move in the same direction (both up or both down) because the equilibrium point slides along the *positively* sloped supply curve. Supply curve shifters cause price and quantity to move in opposite directions (if price goes up, quantity goes down, or vice versa) because the equilibrium point slides along the *negatively* sloped demand curve.

Sometimes the demand and supply curves will shift simultaneously. For example, the increase in the price of barley may not only cause consumers to shift to consuming more corn; it may also induce some farmers to shift time and land from the production of corn into the production of the more expensive barley, pushing the supply curve for corn to the left.

3.4.4 *Functions of several variables and partial derivatives*

When looking at two dimensional graphs of demand and supply curves, it is natural to think of them as shifting in response to changes in other prices, income, and so forth. When working with demand and supply functions, it makes more sense to incorporate such factors directly into the analysis.

Simple example: The effect of changes in the price of pork and diesel fuel on the supply of corn might be captured by the following elaboration of equation (5):

$$s_c^*(p_c, p_p, p_d) = -543 + 300p_c - 10p_p - 7p_d\,. \tag{13}$$

The domain of this function involves three variables: p_c is the price of corn, p_p is the price of pork and p_d is the price of diesel fuel. This equation is a generalization of equation (5): $S_c(p_c) = 300p - 600$. That is to say, as a special case, $s_c^*(p_c, p_f, p_p)$ will yield the same relationship between p_c and s_c that was captured by equation (5) and plotted in Figure 3.3 Specifically, if $p_p = \$5.00$ and $p_d = \$1.00$, then $s_c^*(p_c, 5, 1) = -543 + 300p_c - 50 - 7 = 300p_c - 600$; i.e., $s_c^*(p_c, 5, 1) = s_c(p_c)$.

Demand may also depend on more than price. Income may be said to be a *demand curve shifter* when we are thinking in terms of a two dimensional graph. The relationship might be represented mathematically by the function

$$q(p, y) = 10 - 100p + \frac{y}{2}\,, \tag{14}$$

where q is quantity measured in bushels, p is price in dollars, and y is dollars of income. Then for the special case of $y = 3{,}580$, this is identical to equation (4) and obviously a \$1.00 increase in p (given y), will again change quantity demanded by 100 bushels.

It is clear from equation (14), since it happens to be linear, that at any given level of income a dollar increase in price will reduce quantity demanded by 100 bushels. And that is exactly what the partial derivative is all about. The *partial derivative* of q with respect p, written $\partial q/\partial p$, looks at the change in q if p changes while the other variable(s) (e.g., income) remain constant — it is like a *controlled experiment*. Similarly, the partial derivative of q with respect to y, written $\partial q/\partial y$ looks at the change in q with response to a change in y, holding p constant. We have *two* partials for demand equation (14) because q depends on two variables, p and y.

Basic rule: The partial derivative is calculated according to the usual rules of differentiation, but with the control variable (or variables) treated as constant.

Thus for equation (14) above we have two partial derivatives:

$$\frac{\partial q}{\partial p} = -100 \quad \text{and} \quad \frac{\partial q}{\partial y} = \frac{1}{2}\,. \tag{15}$$

Note that the partial symbol ∂ (sometimes pronounced "deb'ba") is substituted for d in order to indicate that this is a partial derivative.

Similarly, from equation (13) we have for the effects on the supply of corn of changes in the price of pork and diesel fuel:

$$\frac{\partial s_c}{\partial p_c} = 300\,, \quad \frac{\partial s_c}{\partial p_p} = -10\,, \quad \text{and} \quad \frac{\partial s_c}{\partial p_d} = -7\,. \tag{16}$$

The non-linear case is slightly more interesting. First consider the demand equation

$$q(p) = 10p^{-0.5}\,; \tag{17}$$

then obviously, $dq/dp = -5p^{-1.5}$; also, the reciprocal of dq/dp is the slope of the demand curve; i.e., $dp/dq = -p^{1.5}/5$. We can add income y into equation (17), obtaining

$$q(p, y) = 10p^{-0.5} + \frac{y}{2}\,. \tag{18}$$

Now $\partial q/\partial p = -5p^{-1.5}$, just as before; also, $\partial q/\partial y = 1/2$. Note that in this instance the slope of the demand curve depends on price but not income. A slightly richer case is

$$q = 10p^{-0.5}y^{0.3} \tag{19}$$

with partial derivatives

$$\frac{\partial q}{\partial p} = -5p^{-1.5}y^{0.3} = -\frac{1}{2}\frac{(10p^{-0.5}y^{0.3})}{p} = -\frac{1}{2}\frac{q}{p} \tag{20}$$

(obtained by treating y as a constant, as in a controlled experiment) and

$$\frac{\partial q}{\partial y} = 3p^{-0.5}y^{-0.7} = 0.3\frac{q}{y} \tag{21}$$

(obtained by treating p as a constant).[5]

3.5 Elasticity

The consumption of some commodities is quite sensitive to price; others less so. If the price of strawberries climbs when supply falls off near the end of the season, we can always eat cherries or watermelon instead. Because of the availability of such close substitutes, an increase in the price of strawberries may lead to a substantial reduction in the demand because consumers will

[5] We could write $\partial q(p, y)/\partial y = 3p^{-0.5}y^{-0.7}$ if necessary in order to make clear which variable(s) is being held constant.

not find it all that painful to switch to the alternatives. Cigarettes and coffee are another story. Those who are addicted to nicotine will cut back on their intake of many other goods before an increase in the price of cigarettes keeps them from indulging their habit. Coffee drinkers are reluctant to switch to tea or to give up their morning jolt of caffeine.

How should we measure the sensitivity of the quantity sold of a commodity to a change in its price? One obvious measure of the sensitivity of demand to changes in price is the slope of the demand curve. Doesn't a flatter demand curve imply that demand is more responsive to changes in price? Yes indeed, but there is a serious problem with using the slope of a demand curve, or its reciprocal dq/dp, as a measure of demand responsiveness. An example will reveal that the demand curve's slope does not reliably measure the sensitivity of quantity demanded to price because the slope measure is not "unit free."

Example of the Units of Measure problem: Suppose that we wish to determine whether the demand for cigarettes is more price sensitive than the demand for coffee. It is reported that a $1.00 increase in the price of a pack of cigarettes leads a consumer to cut cigarette consumption by 1/10th of a pack of cigarettes (2 out of 20 coffin nails); that is, $dq_c/dp_c = -1/10$th and the slope of the demand function for cigarettes, $dp_c/dq_c = -10$. It is also reported that a $1.00 increase in the price of a cup of Starbuck's coffee reduced this consumer's consumption by one cup; i.e., the slope of the demand curve for coffee is -1. Thus the demand for coffee appears to be much more sensitive to price changes than the demand for cigarettes.

To see why this is an invalid comparison, suppose we had happened to measured smoking in terms of number of cigarettes rather than packs. Since there are 20 cigarettes in a pack, we will find that our consumer will reduce consumption by two cigarettes if the price per cigarettes goes up by five cents. That is to say, $dq_c/dp_c = -2/.05 = -40$ and $dp_c/dq_c = -1/40$, which is less than the slope of the coffee demand curve. The problem is that the slope of a demand curve depends on the *unit of measure* — it is not unit free.

Moral: The slopes of demand curves for commodities measured in different units (e.g., packs of cigarettes versus cups of coffee) are not commensurable.

For example, with such a measure an artful statistician could make the demand for cigarettes appear either more or less sensitive to movements in its price than coffee just by changing the unit of measure.

3.5.1 *Arc-elasticity*

The price elasticity of demand is a "unit-free" measure of the responsiveness of demand to a change in price. First we will consider the concept of *arc-elasticity* and then move on to a discussion of *point elasticity.*

Definition: *Arc-elasticity* is the percentage change in quantity divided by the percentage change in price.

This form of the elasticity concept, which is frequently applied when only a pair of data points is available, is best explained by a couple of examples:

Example #1: An economist has estimated that a 10% increase in the price of cigarettes would cause a 12% reduction in smoking by high school seniors. This implies that the elasticity of high-school student demand for cigarettes is $12\% \div 10\% = 1.2$.

Unit Free: Note that both the percentage change in quantity and the percentage change in price are unit free. Therefore, the elasticity, which is just the ratio of these two percentage changes, is also unit free. We calculate the same elasticity whether we measure cigarettes in packs or cartons.

Example #2: One year farmers harvested about 4.6 billion bushels of corn. Prices averaged \$2.18. The next year the crop was 10 percent smaller because of a drought. Harvest prices averaged \$2.66.

To find the arc-elasticity of demand for corn we first note that:

$$\text{Average price} = \overline{p} = \frac{(2.66 + 2.18)}{2} = \$2.42$$

Also, the price increased by $2.66 - 2.18 = 0.48$. Therefore, the percentage increase in price is:[6]

$$\frac{\Delta p}{\overline{p}} = \frac{0.48}{2.42} = 0.20 = 20\%\,.$$

[6] There is more than one way to compute percentage changes. It is common practice to use the figure from the earlier period as the denominator and calculate $(2.66 - 2.18)/2.18 = 22\%$. But then, if prices were to return to \$2.18, the percentage change would be $(2.18 - 2.66)/2.66 = -18\%$ rather than -22%, which erroneously suggests that there has been a net increase of 4% over the two year period. An advantage of using the average in the denominator is that it makes the percentage change the same (apart from sign) regardless of the direction of movement.

Since the reported percentage change in quantity is -10%, the elasticity of demand for corn is estimated to be 1/2.

Note #1: The Greek letter Δ is often used to denote change; thus, Δp is the change in the price of corn from the preceding year.

Note #2: The formula for arc elasticity is

$$\eta = -\frac{\Delta q}{\overline{q}} \div \frac{\Delta p}{\overline{p}}\,. \tag{22}$$

Sign Ambiguity: Many but not all authors include the minus sign in the equation defining price elasticity in order that the price elasticity of demand measure will normally yield a positive number. For example, when Table 3.1 reports that the price elasticity of demand for shoes is -0.9135 you should assume that the authors are using this alternative concept, equation (22) without the minus sign, rather than intending to suggest that the demand curve for shoes has a positive slope.

3.5.2 *Point elasticity*

A second form of the elasticity concept, point elasticity, is relevant when q is measured as a continuous variable and the demand function is differentiable. The concept of point elasticity can be developed from the concept of arc elasticity: First note that the expression for arc-elasticity (22) can be manipulated to yield:

$$\eta = -\frac{\Delta q}{q} \div \frac{\Delta p}{p} = -\frac{\Delta q}{\Delta p}\frac{p}{q}\,. \tag{23}$$

Then shifting from finite changes to the infinitesimal changes of calculus, it naturally follows that we should define elasticity in the continuous case as follows:

$$\eta = -\frac{dq}{dp}\frac{p}{q}\,. \tag{24}$$

That is the definition of *point-elasticity.* Example: Suppose the demand for coal (tons per year) is given by the equation

$$q = 100 - \frac{p}{3}\,.$$

With this simple demand curve, we have $dq/dp = -1/3$ and the slope of the demand curve is $dp/dq = -3$ (provided we follow the convention of plotting q on the abscissa and p on the ordinate).

- If $p = 100$, $q = 66\ 2/3$, and the price elasticity of demand is $\eta = 1/3 \times 100/66.6 = 1/2$.
- If $p = 150$, $q = 50$, and the price elasticity of demand is $\eta = 1/3 \times 150/50 = 1$.
- If $p = 200$, $q = 33\ 1/3$ and the price elasticity of demand is $\eta = 1/3 \times 200/33.3 = 2$.

As these examples make clear, a peculiar feature of demand elasticity is that it is not the same at every point on a linear demand curve. Figure 3.7 illustrates this not particularly attractive feature of the demand elasticity concept, but it is the penalty we must incur if we are to have a unit-free measure of the responsiveness of demand for a commodity to changes in its price.

A simple graphical technique that can be used to quickly determine the elasticity at any point on a *linear* demand curve is demonstrated in Figure 3.8. Consider any point e on the demand curve. To find the elasticity at that point one takes the ratio of the increase in quantity $q(0) - q(p)$ that would be sold if the price were lowered to 0, which is distance B on the graph, with actual sales $q(p)$, which is distance A on the graph. Then

$$\eta \equiv -\frac{dq}{dp}\frac{p}{q} = \frac{q(0) - q(p)}{q(p)} = \frac{B}{A}, \tag{25}$$

provided the demand curve is linear. To see why, note that $dp/dq = -p/[q(0) - q(p)]$ and so $dq/dp = -[q(0) - q(p)]/p$; therefore, $\eta = \{[q(0) - q(p)]/p\}p/q(p) = [q(0) - q(p)]/q(p) = B/A$. If the demand curve is not linear, simply draw a line tangent to the demand curve at point e and apply (25) to measure the elasticity on the tangent line at point e. This works because two demand curves always have the same elasticity at the point of tangency.

Key Concepts

- We say that demand is *inelastic* if $\eta < 1$.
- Demand is *elastic* if $\eta > 1$.
- It is *unit elastic* if $\eta = 1$.

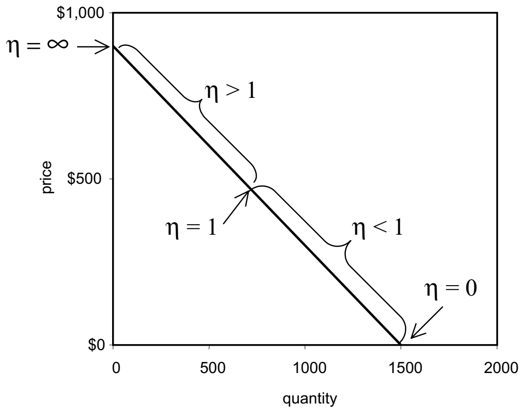


Fig. 3.7. Point elasticity — Linear demand curve
On a linear demand curve, elasticity ranges from zero when $p = 0$ to infinity when $q = 0$.

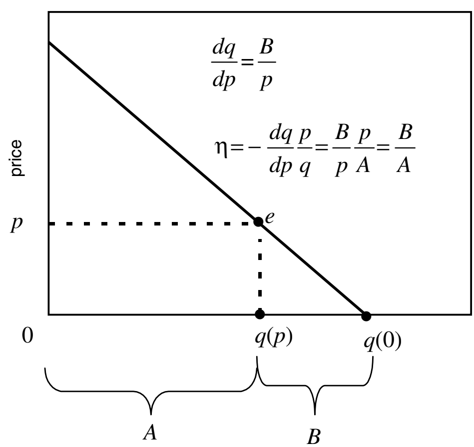


Fig. 3.8. Graphical determination of elasticity
Point elasticity on a linear demand curve is easily determined by calculating the ratio $\eta = B/A = [q(0) - q(p)]/q(p)$.

Constant Elasticity Demand Curves

Elasticity may be constant at every point on a suitably constructed demand curve. Several examples appear in Figure 3.9.

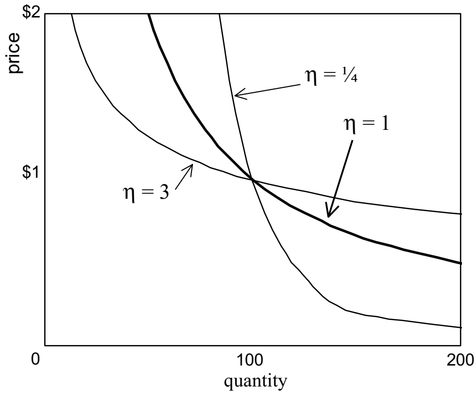


Fig. 3.9. Constant elasticity demand curves

- The heavy demand curve labeled $\eta = 1$ plots the demand function $q = 100/p$.
  To verify that this demand curve does indeed have constant elasticity, observe that $dq/dp = -100/p^2 = -(100/p)/p = -q/p$; therefore, $\eta = -dq/dp \times p/q = 1$.
- The demand curve labeled $\eta = 3$ in Figure 3.9 is the plot of equation $q = 100/p^3$.
  Note that $dq/dp = -300/p^4$ and so $\eta = -300/p^3 \times p/(100/p^3) = -3$.
- In general, if for any coefficient b and demand function of the form $q = ap^{-b}$,
  $dq/dp = -bap^{-(b+1)}$ and $\eta = bap^{-(b+1)}p/q = b$.

Trick question: If the demand curve is $\log_e q = a - b\log_e p$, what is the price elasticity of demand?[7]

Serious questions: Consider *any* inverse demand function p(q).
What is the elasticity of demand at the revenue maximizing price?

[7] This is a trick question because $\log\ q = a - b\log\ p$ can be obtained by taking the logs of both sides of $q = ap^{-}$, the equation used in the preceding example. Since it is the same relationship it must have the same elasticity; i.e., $\eta = b$. Alternatively, you can derive the result directly by recalling from elementary calculus that $d\log\ q/dp = dq/dpq^{-1}$.

To find out, note that $R(q) = p(q) \times q$. If revenue is at its maximum, we must have

$$\frac{dR(q)}{dq} = \frac{dp}{dq}q + p(q) = 0\,. \tag{26}$$

This establishes an important proposition that the price elasticity of demand is unity,

$$\eta = -\frac{dp}{dq}\frac{q}{p} = 1\,, \tag{27}$$

when the revenue maximizing quantity is marketed.

3.5.3 *Price versus income elasticity*

Income elasticity measures the responsiveness of quantity supplied to changes in income. It is very similar to price elasticity:[8]

$$\textit{Price elasticity of demand} = \eta_p = -\frac{\partial q}{\partial p}\frac{p}{q} \tag{28}$$

$$\textit{Income elasticity of demand} = \eta_y = \frac{\partial q}{\partial y}\frac{y}{q} \tag{29}$$

To illustrate, let us recall equation (19):

$$q = 10p^{-0.5}y^{0.3}$$

For the price elasticity of demand we calculate

$$\eta_p = -\frac{\partial q}{\partial p}\frac{p}{q} = 5p^{-1.5}y^{0.3}\frac{p}{q} = \frac{5p^{-0.5}y^{0.3}}{(10p^{-0.5}y^{0.3})} = 0.5\,. \tag{30}$$

The income elasticity of demand is calculated in parallel fashion:

$$\eta_y = -\frac{\partial q}{\partial y}\frac{y}{q} = 3p^{-0.5}y^{-0.7}\frac{y}{q} = \frac{3p^{-0.5}y^{0.3}}{(10p^{-0.5}y^{0.3})} = 0.3\,. \tag{31}$$

A number of constant income elasticity demand functions are plotted, given p, in Figure 3.10. For some luxury goods the income elasticity may be greater than unity because only the well-to-do can afford to buy them. For

[8]In contrast to equation (24), we use the partial derivative sign in defining η because we are working with a function of two variables, income as well as price.

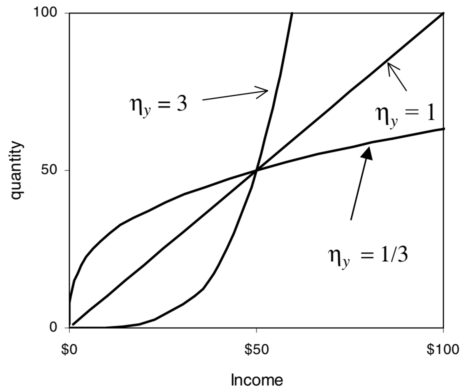


Fig. 3.10. Constant income elasticity of demand

the necessities of life, income elasticity is likely to be less than unity. Indeed, statistician Ernst Engel [1821–1896] argued, on the basis of intensive empirical research, that the proportion of a consumer's budget spent on food tends to decline as the consumer's income increases. This proposition, known as *Engel's Law*, may well be the first quantitative economic regularity established on the basis of empirical research. Numerous surveys of consumer behavior from many different countries have supported this proposition.

3.5.4 *Some elasticity estimates*

Table 3.1 presents price and income elasticities for a large number of commodities that were estimated by econometricians Houthakker and Taylor.[9] For most of the commodities the authors present both short-run and long-run estimates in recognition of the fact that for many commodities it takes time for consumers to adjust fully to a price change. Several points should be noted about these estimates:

- The authors are reporting the price elasticities with a negative sign; i.e., they are leaving out the minus sign in equation (24).

[9]Technically, total expenditure instead of income elasticities are reported on the table, where expenditures are defined as income less saving. But since total expenditure is roughly proportional to income, the income and expenditure elasticities are essentially the same.

Table 3.1. Some price and income elasticity estimates.

	Expenditure elasticities		Price elasticities	
	Short-run	Long-run	Short-run	Long-run
Purchased meals	1.6126		−2.2703	
Tobacco products	.2075	.8615	−.4556	−1.8919
Shoes & other footwear	.9433		−.9135	
Jewelry & watches	1.0025	1.6447	−.4100	−.6726
Owner occupied housing	.0707	2.4495	−.0351	−1.2150
Electricity	.1319	1.9364	−.1289	−1.8926
Legal Services	.4264		−.3707	
Taxicabs	1.1460		−.6299	
Newspapers & magazines	.3841		−.4185	
Motion pictures	.8126	3.4075	−.8748	−3.6685
Foreign travel (by U.S. residents)	.2355	3.0873	−.1351	−1.7707

Source: H. S. Houthakker and Lester D. Taylor, *Consumer Demand in the United States: Analyses and Projections*, Harvard University Press, 1970, 2nd ed., pp. 166–7.

- The income elasticity estimates are based on total expenditure rather than total income, but this probably does not make much difference.
- The long-run elasticities are generally larger than the short run, reflecting the fact that time is required to adapt to changes.

  For example, a family that heats the house with electricity may not immediately switch from electricity to gas or oil when its price rises. They may be slow to add extra insulation. But over the long haul they will make substantial adjustments.

  In the case of motion pictures, it may take time for a price reduction to attract new customers; but once they start watching the flicks, they may get addicted and go quite frequently.
- Since the estimates are all from the same source, they are likely to be more or less consistent with each other. But some of them are questionable, such as the rather high long-run price elasticity of demand for tobacco products.
- The investigators reported their estimates to four decimal places, but in truth there is probably less than two digits of precision in the estimates.

Supply Elasticity

The concept of elasticity has a number of other applications, including the task of measuring the responsiveness of supply to changes in price. If $s(p)$ is the supply curve, it is a natural extension to define the supply elasticity, as

$$\eta_s = \frac{ds}{dp}\frac{p}{s}\,. \tag{32}$$

3.6 Some applications

3.6.1 *Paradoxes resolved*

Housing Paradox Revisited: When we looked at the effect of a reduction in the stock of housing in Middletown at the beginning of this chapter, we observed in Figure 3.2 that the revenue collected by landlords might actually increase, but only up to a point. Specifically, the reduction in supply from 1000 to 750 led to a revenue increase, but when supply was reduced to 500, revenue declined. The concept of elasticity helps in explaining what happened. Recall demand equation (1):

$$q = 1500 - \frac{5}{3}p\,.$$

The elasticity of demand will be unity when 750 units are available, as can most easily be seen with the shortcut rule presented on Figure 3.8. This is the point of maximum revenue, as explained by the proposition on page 82, equation (27). That is why total rents paid by tenants will decline if the supply of housing is reduced below 750 units.

Crop Failures Revisited: A similar paradox was encountered when we looked at the consequences of a crop failure in section 3.3.1. The crop failure certainly hurt consumers, but farmers on average realized more income because the price increased substantially as a result of the crop failure. This will happen whenever the demand for a commodity is inelastic ($\eta < 1$), as it generally is for most agricultural commodities.

3.6.2 *The minimum wage*

The minimum wage is a floor on wages. In 1938, during the Great Depression, the Federal Government of the United States established a national minimum wage of 25 cents per hour. The act was of only limited impact initially, in part because the act excluded many sectors of the economy, including agriculture, from coverage and in part because the law was not effectively enforced. Over the years the minimum has been gradually raised in a series of steps from the 25¢ per hour in 1938 to a floor of $5.15 per hour by year 2000. While no act of Congress has ever lowered the minimum wage, its purchasing power has with great frequency been reduced by inflation. In terms of purchasing power, the minimum wage peaked out in 1968. That year's $1.58 minimum wage had 50% more purchasing power than the $5.15 minimum wage of year 2000. Congress and the President had let inflation dramatically reduce the strength of the minimum wage, but some say this was just as well.

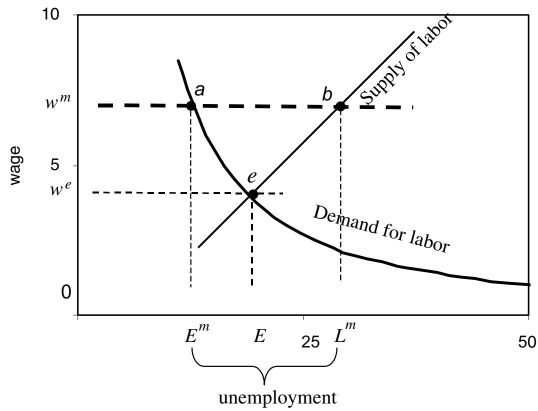


Fig. 3.11. Minimum wage
When minimum wage legislation forces employers to pay substantially more than the equilibrium wage, the quantity of labor demanded is reduced to E but the supply increases to L. Thus the minimum wage generates unemployment equal to $L - E$.
The total income earned by workers after the minimum wage is introduced is $w \times E$, which will be more than before the minimum wage was enacted if and only if the demand curve is inelastic.

Many economists object that the minimum wage creates unemployment by pricing teenagers and unskilled workers out of the labor market. Figure 3.11 explains the argument. In the absence of the minimum wage, the market for teenage labor would equilibrate with a wage w^e and level of employment E. With a minimum wage, only E^m workers (generated by point a on the demand for labor curve) will be employed; but because of the potentially higher reward, the supply of workers will increase to L^m (determined by point b on the supply curve for labor). Hence the minimum wage will generate unemployment of $U^m = L^m — E^m$. The excess supply of workers may encourage employers to pick and choose and perhaps discriminate in hiring. Critics argue that eliminating the minimum wage would reduce teenage unemployment and job-market discrimination. It is interesting to note that while the model predicts that a hike in the minimum wage causes a reduction in teenage employment that does not mean that it will reduce the income of teenagers. If the demand for teenage labor is inelastic, as drawn on the graph, an increase in the minimum wage will lead to an increase in teenage income (wage $\times$ hours worked) — but that will be only cold comfort for those who cannot find a job.

Although there have been a tremendous number of econometric studies of the minimum wage, they offer no clear or decisive answer to the question of whether the demand for teenage labor is elastic or inelastic. Indeed, two distinguished scholars, David Card and Alan B. Krueger[10] surprised economists with an empirical study suggesting that the minimum wage may not create unemployment after all! We will find, in Chapter 7.3.2, that if the labor market is not competitive the imposition of a minimum wage may lead profit-maximizing employer to hire more workers and increase employment. Thus the issue of whether the minimum wage causes unemployment is still somewhat up in the air.

It is worth noting in conclusion that labor unions may generate unemployment in much the same way as the minimum wage. If the union pushes up the wage that employers must pay, it will result in a reduction in the demand for labor, assuming that the labor market would be competitive in the absence of union intervention. However, the total income received by the union members will increase if the demand for labor is inelastic.

3.6.3 *Farm price supports*

No politician likes inflation, but politicians often find themselves pressured into supporting policies that push up prices. Farmers have been particularly adept at lobbying their representatives to guarantee them a "fair" price for their product, although the high price that farmers regard as fair for them might be regarded as most unfair by the majority of consumers if they took the time to think about it.

As explained in Chapter 1, the 20th century was an age of tremendous progress in agricultural productivity. Technological advance has shifted the supply curve for agricultural products outward to the right faster than the growth in population has shifted the demand curve. Because the demand

[10] Card, David and Alan B. Krueger, *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton University Press, 1955.

for agricultural products is price inelastic, farm revenue declined when technological progress shifted the supply curve outward. The problem has been compounded because the demand for food is income inelastic, which means that the food component of the typical household's budget has not kept pace with rising income levels over the years. During much of the 20th century, declining farm income relative to what could be earned in manufacturing, encouraged major migrations from idyllic farm life to factory work. It also generated sympathy in Washington for the plight of the farmer.

More than 100 years ago, Populist orator Mary Elizabeth Lease implored farmers to "raise less corn and more hell." Restricting production to offset the rightward shifts in the supply curve for agricultural commodities generated by technological progress might have prevented the fall in prices that farmers found so ruinous, but with rare exception farmers have not been successful in their attempts to mount a coordinated effort to stem production. Indeed, when prices fall, farmers may find themselves compelled to work all the harder in an effort to make the fixed mortgage payments to the bank on the funds they had borrowed to buy land and equipment. But if the farmers could not successfully unite on their own to restrict production, they could find help in Washington.

Over the years a complex web of legislation has been enacted and revised and revised again in an effort to help the farmer. Price supports are a favorite technique for achieving a "fair price" for the farmer. Here is a simplified explanation of how they work: Suppose the government makes a standing offer to purchase wheat at a specific support price p^s that is above equilibrium price p^e. The effect of this policy is to augment private sector demand, $q(p)$ with demand from the support program. Because the government stands ready to buy however much is required to support the price, the demand curve shifts as shown by the horizontal segment corresponding to p^s on Figure 3.12. As can be seen from the graph, maintaining the support price will require that the government purchase the "surplus" of $s(p^s) - q(p^s)$. The government can stockpile the wheat as a reserve against future crop failures. If the support price were not set too high, the surplus purchased in years of bountiful harvest could be sold off in years of crop failure. Those arguing for price stabilization point out that it is only prudent to put aside for use in future droughts part of the output obtained in years of bounty.[11] But in practice, the additions to the stockpiles

[11] In Chapter 11 we will investigate the role of speculators and government price support programs in offsetting the effects of crop failure.

in good years exceed on average the withdrawals in years of meager yield because the farm lobby successfully pressures for a "fair" support price that is above the average equilibrium price. That is why price support programs almost always lead to the accumulation of surpluses in larger and larger stockpiles.

In times when people in many sectors of the globe have suffered from extreme hunger, farm surpluses can be a tremendous embarrassment to Uncle Sam. Sometimes surpluses have been burned or buried. When in the 1954, President Eisenhower established the "Food for Peace" program to export the surplus to feed the world's hungry, this gift made no friends among farmers outside the United States. They complained that the "dumping" of wheat on the market pushed down the world price. During much of the cold war, Uncle Sam was selling surplus wheat to the Russians at a discount below the prevailing support price consumers in the United States had to pay. It may not be too cynical to say that Midwestern congressmen and senators had the farm problem on their minds as well as the needs of hungry citizens in the inner city when they cast their votes in support of federal funding for school lunches and the food stamp program.

In addition to encouraging demand, from time to time a variety of restrictions on production have been imposed. For example, starting in the 1930s production quotas were established to restrict the output of dairy farmers, A dairy's quota specifies the number of gallons that the farmer can market as fresh ("Grade A") milk, with over quota-milk bringing a

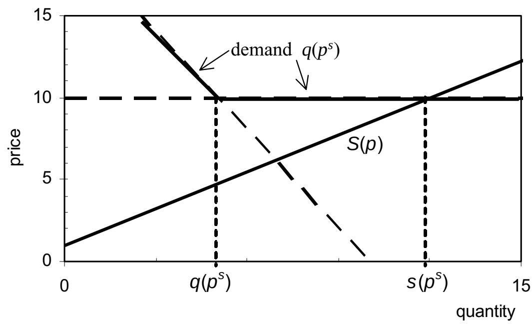


Fig. 3.12. Farm price supports
Because of the government's support price of $p = 10$, market equilibrium cannot exist at point e. Instead, at the support price, supply is $s(p)$ but only $q(p)$ is demanded. Thus there is a surplus (excess supply) equal to $s(p) - q(p)$ that must be purchased by the government.

lower price because it can only be used to make processed milk products, such as powdered milk, cheese and ice-cream. As another example, the "Soil Bank" program of the 1950s attempted to shift the supply curve to the left by paying farmers a subsidy for *not* farming part of their land, but it was only partially successful in restricting production because the farmers set aside their least productive land and farmed their remaining acres more intensively.

Farmers in many countries — England, France, Germany and Japan as well as the United States — have proved remarkably adapt at winning supportive legislation from their politicians. But price supports, acreage restrictions and export subsidies belong to the class of economic medicines that ease the pain by prolonging the illness. To the extent that a nation's farm policy eases the downward pressure on farm incomes generated by rapid technological progress, it also reduces the economic incentives for farmers to shift to other occupations. Easing the plight of the farmers slows the rate at which the economy adjusts to the fact that greatly increased farm productivity means that fewer farmers are required to feed the world. More than this, when developed countries dump surplus farm output on the world market, it artificially depresses prices to the point where farmers in poor countries find they cannot compete.

3.7 Foreign exchange rates

The *foreign exchange rate* is the price of foreign currency. For example, an American student planning to visit Japan would need to purchase yen with dollars. If the exchange rate were $1 = 120 yen, our student would be able to purchase 120 yen with each dollar taken to Japan. The transaction might be executed at a "Change" booth at the airport or at an ATM in Japan. If the exchange rate had been only $1 = 90 yen, the dollar would have been less valuable because only 90 yen could be purchased with each dollar. Naturally, the American student would like the dollar to be "strong"; i.e., the more yen that the American student can acquire for each dollar the better. Conversely, a Japanese tourist coming to the United States would prefer to have to pay only 90 yen rather than 120 yen for each dollar to be spent in the United States. That is to say, the Japanese tourist would like the yen to be strong (i.e., the visitor would like the dollar to be weak). American firms importing goods from abroad like a strong dollar because it means they pay a lower price, measured in dollars, for the goods they purchase overseas. Contrariwise, American exporters like a weak dollar because it stimulates sales abroad by reducing the price that foreigners have to pay in their currency for the goods.

The *Wall Street Journal* and other major newspapers report exchange rates on a daily basis. Information on hour to hour fluctuations in the exchange rate is posted on the Web.[12] Figure 1.6 demonstrated that at times foreign exchange rates can move quite abruptly from one year to the next while in other periods they have been remarkably stable. Foreign exchange rates are determined in the marketplace, but the outcome is often influenced by government intervention. The exchange rate picture is complicated because central banks — such as the Bank of England, the Bank of Japan and the U.S. Federal Reserve Board — at times intervene actively in the market by buying or selling foreign currencies in an effort to influence the market. As a first step toward learning how all this works, we must investigate the factors entering into the demand and supply of foreign exchange.

3.7.1 *Demand and supply*

The demand and supply for the dollar in the foreign exchange market during 2000 is summarized on Table 3.2.

Goods: The first item on the table is United States goods exports, which amounted to $772 billion in 2000. The countries buying these goods needed dollars to pay for them — thus U.S. exports generate demand for the dollar in the foreign exchange market. But our goods imports were even larger, $1,224 billion, and we supplied dollars in paying for them. Line 4 shows that the United States imported $452 billion more goods than she exported.

Services: The United States supplied services worth $293 billion, which include $82 billion in travel services and $38 billion in royalties and license fees — dollars were demanded to pay for these expenditures. It is clear from the table that the U.S. provided much more in the way of services to foreigners than she purchased from abroad. Nevertheless, she had an overall deficit on the combined Goods and Services category recorded on line 20. How did the U.S. pay for this excess of imports over exports?

---

[12]Professor Werner Antweiler of the University of British Columbia's Sauder School of Business reports exchange rates for more than 200 currencies: http://fx.sauder.ubc.ca/.

Table 3.2. United States international transactions, 2000.

Line	(Billions of dollars)	Demand for $	Supply of $	Balance
1	Goods			
2	Exports	772		
3	Imports		1,224	
4	Balance			(452)
5	Services			
6	Exports	293		
7	Imports		217	
8	Balance			76
9	Investment income			
10	From U.S. investments abroad	353		
11	On foreign assets invested in the U.S.		368	
12	Balance			(15)
13	Unilateral transfers, net		54	(54)
14	Investments			
15	Increase in U.S. assets invested abroad		581	
16	Increase in foreign assets invested in U.S.	1,024		
17	Balance			443
18	Miscellaneous		(1)	1
19	Column Sum	2,443	2,443	0
Addendum:				
20	Goods and services	1,066	1,224	(376)
21	Goods services and investment income	1,419	1,441	(391)

Source: *Survey of Current Business*, July, 2002
Note: Negative numbers are in parentheses, in accordance with the accounting convention.

Investment Income: Funds that Americans have invested overseas earn profits, interest and dividends. Foreign currency must be converted into dollars when these earning are brought home to the United States — here is a source of demand for the dollar. But foreigners have huge investments in the United States, and as can be seen from line 12, on balance investment income contributes toward a further excess in the supply of dollars over the demand.

Unilateral Transfers: This strange sounding category includes U.S. Government foreign aid grants, gifts of emigrants to those who remained behind and pensions to those who have retired abroad. On net, unilateral transfers add to the supply of dollars to the foreign exchange market.

Investments: When an American invests funds in Hong Kong, that's foreign investment, and in order to obtain the funds required to make that overseas purchase, dollars must be supplied to the foreign exchange market. Conversely, if a Japanese investor decides to purchase Manhattan real-estate, the dollars required to complete the transaction must be purchased in the foreign exchange market.

Demand equals supply

The data about the demand and supply of the dollar recorded on Table 3.2 for 2000 are history now, but what the magnitude of the various flows turned out to be was influenced by the dynamics of exchange rate movements. The demand and supply columns on Table 3.2 sum to precisely the same number (see line 19) — the exchange rate adjusts in order to establish this equality of demand with supply. The variety of market forces are summarized by the demand and supply curves for the dollar plotted on Figure 3.13. The graph focuses on exchanges of dollars and yen. It shows how the forces of supply and demand determine the value of the dollar, as measured by how many yen one dollar will buy:

- The demand curve shows the number of dollars that market participants wish to buy with yen. This demand curve has the conventional negative slope because a higher value of the dollar would discourage the conversion of yen into dollars in order to buy American goods and services, invest in American financial markets, etc.
- The supply of dollars is drawn with the conventional positive slope — a high number of yen to the dollar makes it cheaper to import Japanese electronics and autos into the United States, which will encourage Americans to spend additional dollars importing more.

The exchange rate adjusts so as to equate the demand for dollars with supply in the foreign exchange market at equilibrium point e.

In the absence of intervention by central banks and governments in the foreign exchange market, the value of currencies fluctuates from year to year and indeed from hour to hour. What the exchange rate turns out to be depends upon a variety of market forces. An economic boom may stimulate a country's imports, which will shift outward the supply curve of the country's currency on the foreign exchange market and contribute to a decline in its exchange rate. A rise in interest rates in the United States, relative to rates in other countries, may encourage foreign investors to move their funds to the U.S., which causes an upward shift in the demand curve for the dollar, and a resulting strengthening of the dollar in the foreign exchange market.

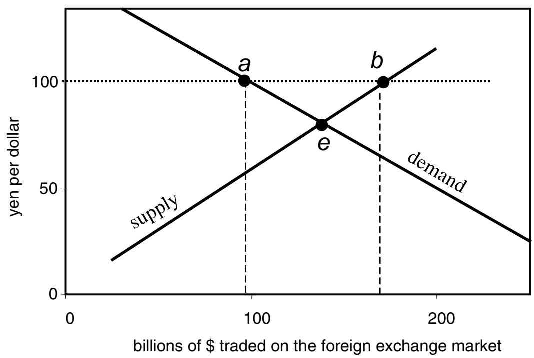


Fig. 3.13. Demand and supply of foreign exchange
The equilibrium exchange rate, established by the intersection of the demand and supply curves at point e, is about 85 yen to the dollar.

The Bank of Japan could push the exchange rate up to 100 yen to the dollar by purchasing dollars with yen on the foreign exchange market. The Federal Reserve Board (the central bank of the United States) could achieve the same objective by using some of its holdings of yen to purchase dollars. The distance b–a indicates the number of dollars that would have to be purchased in order to fix the exchange rate at 100 yen to the dollar.

If a country experiences more inflation than its trading partners, its currency is likely to deteriorate in the foreign exchange market. To illustrate, suppose that prices double in Country A but are stable in Country B. If the exchange rate remains stable, goods from Country A will be twice as expensive in Country B, which will lead to a cutback in shipments from Country A to Country B and a reduction in the demand for Country A's currency. At the same time, residents of Country A will find that goods from Country B are a bargain relative to the inflated price of goods produced in Country A, which will lead to increased imports and a greater supply of Country A's currency on the foreign exchange market. Both forces put pressure on the value of Country A's currency. Purchasing power parity will be restored, in the absence of other changes, if the value of Country A's currency falls by the same percentage as the rise in prices so as to restore the initial equilibrium.

3.7.2 *Fixed versus floating exchange rates*

At times countries have kept the exchange rate fixed. For example, the yen/dollar exchange rate was fixed at 360 yen equal to \$1.00 from 1953 to 1970, as can be seen on Figure 1.6. In recent years the rate has been allowed to "float"; that is to say, the exchange rate has been allowed to fluctuate in response to market forces instead of being fixed, although both the U.S. and the Japanese central banks do intervene from time to time in an effort to smooth out the more abrupt fluctuations.

To see how central bankers can try to stabilize exchange rates, suppose that the Bank of Japan (the Japanese central bank) wished for whatever reason to stabilize the exchange rate at 100 yen equals $1.00. In the situation represented on Figure 3.13, the current equilibrium rate is 80 yen to the dollar — the yen is said to be "strong" (and therefore the dollar is "weak") because it takes only 80 yen to buy a dollar. To lower the value of the yen (i.e., raise the value of the dollar), the Bank of Japan can augment the supply of yen by purchasing dollars with yen in the foreign exchange market, much as the purchase of agricultural commodities by a government pulls up their price (Recall Figure 3.12).

Suppose that in future years changes in market forces — i.e., shifts in the demand and supply curves — lead to a fall in the value of the yen (i.e., a rise in the equilibrium value of the dollar relative to the yen). For example, the equilibrium exchange rate might move toward 120 yen equal to $1.00. The Bank of Japan, if it thought it appropriate, could now use the dollars it had acquired over the years to purchase yen in order to move the exchange rate back toward 100 yen equal $1.00. However, no central bank has unlimited holdings of foreign currency and other reserves, such as securities issued by the U.S. government and gold. This means that the Bank of Japan can support the value of its currency for only so long. Of course, the Federal Reserve Bank of the United States can, if it is so disposed, also play a stabilizing role by purchasing yen for dollars when the value of the yen falls below 100 yen to the dollar and selling yen for dollars when the value rises above 100 yen.

How much in the way of intervention will be required to support a country's currency depends on the gap between demand and supply at the support price. If the gap is too large, the country will find that its attempt to preserve the value of its currency in the foreign exchange market leads to the rapid depletion of its international reserves, such as foreign currency and gold. When the country's reserves are finally exhausted, the central bank will be powerless to support the value of the nation's currency above the market price. Before this happens, speculators who anticipate the crisis will have make every effort to move funds out of the deteriorating currency into a more secure asset, perhaps by purchasing Euros or dollars. This increases the supply of the distressed currency, thereby enlarging the gap between demand and supply — that is how speculative activity can hasten the collapse of a currency. The question of whether and how vigorously central bankers should move to stabilize foreign exchange rates has been debated by generations of economists.

There are several other techniques, in addition to directly participating in the foreign exchange market, by which governments and central banks can attempt to influence the exchange rate. As we shall see in Chapter 10, the central bank may attempt to support a deteriorating currency by raising interest rates in order to encourage foreign investors to move their funds into the country. But raising interest rates may lead to recession. Governments have at times imposed direct controls limiting access to the foreign exchange market — citizens wishing to travel abroad or import unessential goods may find that they will not be allowed to purchase foreign currencies. But direct controls may be circumvented by black marketeers and contribute to an inefficient allocation of resources.

The Euro

Instead of just fixing the exchange rates, countries may adopt a common currency. In January 2002, twelve European countries replaced their own currency with the Euro, which means that the inconvenience involved in converting one currency into another, as in traveling from say Germany to France, has at last been eliminated. Perhaps it is ironic that this long sought goal is being achieved at a time when the automatic teller machine, the pocket calculator, and the computer are minimizing the inconvenience involved in having multiple currencies. The common currency does facilitate the full integration of markets but requires the coordination of monetary policy.

3.7.3 *The United States, A debtor nation*

As Table 3.2 makes clear, in year 2000 the United States imported substantially more in the way of goods and services than she exported (line 20). Further, we earned less on our investments abroad (line 10) than foreign investors earned in the United States (line 11). The table shows (line 17) that this deficit was financed in large measure by an excess of foreign investment in the United States over the investments that Americans were making abroad. In year 2000 foreigners chose to invest $443 billion more in the U.S. than U.S. citizens and corporations invested abroad. It is the willingness of foreigners to invest in American corporate stock and real

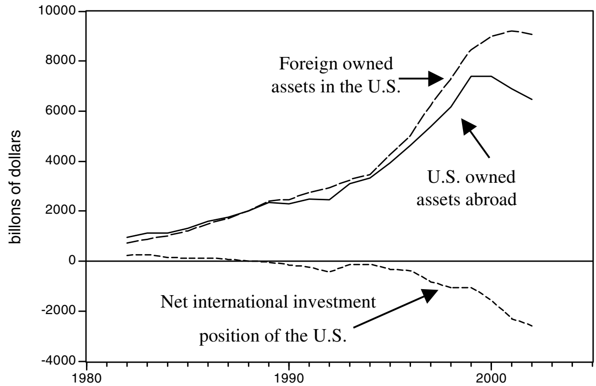


Fig. 3.14. The net international investment position of the United States
The United States is now a debtor nation, for the U.S. government, private citizens and corporations have less invested abroad than foreigners have invested in the United States. At the end of 2002, the value of foreign investments in the United States exceeded the value of U.S. investments abroad by more than $2,380 billion.

estate and U.S. government securities that helps to finance the excess of U.S. imports over exports.

Year 2000 was not exceptionally, for the United States has been running sizable deficits in the balance on goods and services since the early 1970s. How can this be? In a sense the United States has been living beyond its means. Figure 3.14 shows that although U.S. investments aboard have grown over the years they have been surpassed by the more rapid growth of foreign investments in the United States. The result is that the Net International Investment Position of the United States has been negative since the mid 1980s. That is to say, over the years the excess of funds that foreigners have invested in the United States over the investments that Americans have made abroad had cumulated to about $1.5 trillion by year 2000. The United States is a debtor nation — we owe more to foreign countries than they owe to us. We can continue to import more than we export just as long as foreign investors choose to continue investing more and more in American stocks and bonds, real estate, and U.S. government securities.

3.8 Taxes

Governments finance their affairs by levying taxes. But taxes do more than raise revenue. Taxes also affect the price paid by consumers and the quantity sold. We must look at the effects of excise taxes, such as the taxes on gasoline, liquor, cigarettes, gas guzzling cars and airline tickets. When OPEC pushes up the price of oil, consumers in the United States complain about the excise tax the states levy on gasoline, which has ranged in the United States from 8 cents per gallon in Alaska to 33 cents in Connecticut.

To find out how taxes influence prices and output, let us consider how a tax would affect the hypothetical corn market. Remember, this market had demand curve (equation 4)

$$q(p) = 1800 - 100p\,, \quad 0 \leqslant p \leqslant \$18$$

and supply function (equation 5)

$$s(p) = 300p - 600, p \geqslant \$2\,,$$

which yield equilibrium sales $q^e = 1,200$ at a price $p = \$6.00$.

The governor is hoping that a tax of $t = \$4.00$ per bushel will yield tax revenue of $tq^e = \$4 \times 1200 = \$4,800$. However, this cannot happen because quantity will change! Producers would continue to supply 1,200 bushels of wheat only if they could still get $p_s = \$6.00$ per bushel. But if the sellers get $p_s = \$6.00$, the added \$4.00 tax will push the price consumers have to pay up to $p_c = \$10$, and at this high a price consumers would be willing to purchase only 800 bushels. Clearly, quantity cannot remain at 1,200, so there is no way that the governor can realize the predicted tax revenue of \$4,800 with a \$4.00 per bushel tax.

To find out what will happen to prices and quantity as a result of the \$4.00 tax it is necessary to carefully consider the distinction created by the tax between the price paid by the consumer, p_c, and the price received by the seller, p_s. A tax of \$4.00 per bushel will drive a *tax wedge* between p_c and p_s:

$$p_c - p_s = \$4.00 \tag{33}$$

The price paid by consumers, p_c determines the quantity that they will purchase, so we rewrite equation (4) as

$$q(p_c) = 1800 - 100p_c\,, 0 \leqslant p_c \leqslant \$18\,. \tag{4*}$$

In contrast, producers respond to the price they receive, p_s, and so we rewrite equation (5) as

$$s(p_s) = 300p_s - 600\,, \quad p_s \geqslant \$2\,; \quad \text{else} \;\; 0\,. \tag{5*}$$

Substituting for p_s from equation (33) into $(5)^*$ yields a new supply function with the price paid by the consumer as its argument:

$$s^*(p_c) = 300(p_c - 4) - 600\,, \quad p_c \geqslant \$6\,; \quad \text{else} \;\; 0\,. \qquad (5)^{**}$$

Now we equate demand and supply,

$$q(p_c) = 1800 - 100p_c = s^*(p_c) = 300(p_c - 4) - 600\,, \qquad (34)$$

obtaining $p_c = \$9$, $p_s = \$5$, $q(p_c) = s(p_s) = 900$, and $900 \times \$4.00 = \$3{,}600$ in tax revenue for the governor. Figure 3.15 shows this result.

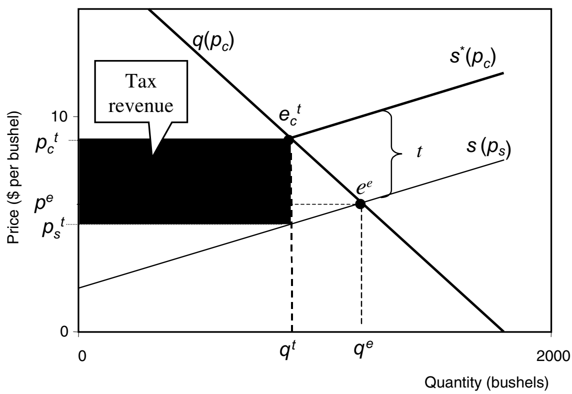


Fig. 3.15. A $4 tax per bushel of corn
Before the tax, the old equilibrium price p and quantity q were determined by the intersection of demand curve $q(p)$ with supply curve $s(p)$ at point e .
The $4 tax drives a wedge between the price p paid by consumers and the price p received by the seller; i.e., $p \;\; - p \;\; = t$. After the $4 per bushel tax is imposed, producers will be willing to supply old quantity q only if they receive the old price, which would mean that consumers would have to pay $p \;\; = p \;\; + t$. Thus supply as a function of p shifts upwards by the amount of the tax to $s^*(p\;\;)$. The new equilibrium is established at point e where $s^*(p\;\;) = q(p\;\;)$.
Tax revenue of tq is represented by the shaded rectangle on the graph.

One might think that the $4 tax would be shared equally by buyers and sellers — wouldn't that be fair? The market determines otherwise. Since the price consumers pay has gone up from $6 to $9, they may object that the $3 increase in price that they are paying is more than their fair share of the $4 excise tax.

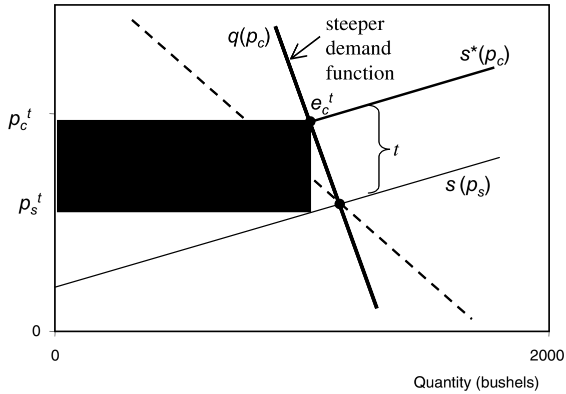


Fig. 3.16. Tax incidence with a steeper demand curve
With a steeper demand curve, the tax generates more revenue because there is a smaller reduction in quantity sold. Also, the price paid by consumers increases more, contributing to a greater loss in consumer surplus.

3.8.1 *Tax incidence*

The *incidence of a tax* is the division of the burden of the tax between buyers and sellers. We shall find that neither the governor nor the legislature has the power to determine how much of the tax increase will be translated into a higher price p_c for the consumer and how much will be imposed on the price p_s received by the seller. The incidence of the tax depends on the forces of supply and demand. To see the truth of this proposition, consider Figure 3.16. Observe that with a steeper demand curve, indicating that demand is less responsive to price change, sales drop off much less, more tax revenue is collected, and even *more* of the \$4.00 tax is shifted onto consumers. Careful contemplation of the demand and supply graph will suggest that if it had been supply that had been less sensitive to price (i.e., the supply curve were steeper), *less* of the tax would have been passed onto consumers.

In order to work out the incidence of the tax, it will be useful to analyze the problem algebraically. Moving away from our specific numerical example, let us write the demand and supply equations as

$$q(p_c) = d_0 - d_1 p_c \tag{35}$$

and

$$s(p_s) = -s_0 + s_1 p_s\,; \tag{36}$$

the corresponding inverse demand and supply functions are

$$p_c q = \frac{d_0 - q}{d_1} \tag{37}$$

and

$$p_s s = \frac{s + s_0}{s_1}\,. \tag{38}$$

Now the tax drives a wedge between the price paid by the consumer and what the producer gets:

$$p_c - p_s = t \tag{39}$$

After substituting from (39) into (36) we must equate demand with supply to solve simultaneously for p_c.

$$q(p_c) = d_0 - d_1 p_c = -s_0 + s_1(p_c - t)\,. \tag{40}$$

Hence, we find that the price paid by the consumer as a function of the tax rate is

$$p_c t = \frac{d_0 + s_0 + s_1 t}{d_1 + s_1}\,. \tag{41}$$

Observe by differentiating (41) that

$$\frac{dp_c}{dt} = \frac{s_1}{d_1 + s_1}\,. \tag{42}$$

Since $p_s = p_c - t$ from (39),

$$\frac{dp_s}{dt} = \frac{dp_c}{dt} - 1 = -\frac{d_1}{d_1 + s_1}\,. \tag{43}$$

This demonstrates that it does not matter whether the law specifies that the tax is incorporated in the quoted price, as with gasoline, or whether it is an add-on, as with the sales tax. It is the law of supply and demand, not the law passed by the legislature, that determines the incidence of an excise tax. More precisely, the incidence of the tax depends on the relative magnitudes of the two slope coefficients indicating how sensitive demand and supply are to price changes; the intercepts of the demand and supply curves do not affect tax incidence. For our numerical example we had $d_1 = 100$, $s_1 = 300$. Therefore, $dp_c/dt = 3/4$. If instead we had $d_1 = 300$ and $s_1 = 100$, then $dp_c/dt = 1/4$ and $dp_s/dt = -3/4$. Only if $d_1 = s_1$ will the tax be evenly divided.

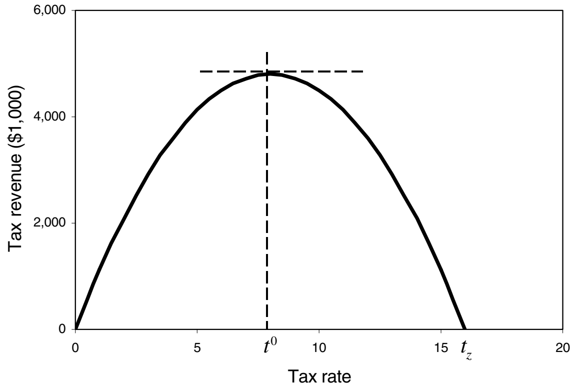


Fig. 3.17. The Laffer Curve
Tax revenue is maximized with a per unit tax of t^0.

3.8.2 *Maximizing tax revenue — The Laffer curve*

What tax rate will maximize tax revenue? While it may be obvious that a low tax rate will yield little revenue, too high a rate will cause such a large reduction in quantity that little or no revenue will be generated. In 1978 Arthur Laffer, an economist who was to serve on President Ronald Reagan's staff, explained over lunch to *Wall Street Journal* reporter Alfred Malabre Jr. that large tax cuts would spur so much economic activity that they would increase tax revenue. He explained his argument with a simple graph that he drew on the restaurant's tablecloth. The next day Laffer's graph, similar to Figure 3.17, appeared in the *Wall Street Journal*. Laffer was asserting that because tax rates were above the revenue maximizing rate t^0, a reduction in taxes would raise additional revenue.

Let us derive the Laffer curve and the revenue-maximizing tax rate for the case of linear demand and supply curves. Substituting (41) back into demand function (35) yields the equilibrium quantity as a function of the tax rate:

$$q^e(t) = q[p_c(t)] = d_0 - d_1 \frac{d_0 + s_0 + s_1 t}{d_1 + s_1}\,. \tag{44}$$

Equation (44) simplifies to

$$q^e(t) = q^e(0) - \frac{d_1 s_1 t}{d_1 + s_1} = q^e(0) - \alpha t\,, \tag{45}$$

where $\alpha = dq^e/dt = d_1 s_1/(d_1 + s_1)$. Since tax revenue is the product of the tax times quantity sold, equation (45) implies that tax revenue as a function of t is

$$T_r(t) = tq^e(0) - \alpha t^2 \,. \tag{46}$$

Three points about this Laffer curve function deserve attention:

1. The first term, $tq^e(0)$, is the product of the tax rate times the quantity that would be sold *if* the tax rate were zero. That is to say, $tq^e(0)$ is the naïve estimate of the revenue the tax will yield calculated under the erroneous assumption that the tax will not affect quantity sold.
2. $T_r(0) = 0$, which makes sense because a tax rate of zero will obviously yield no tax revenue.
3. Tax revenue will also be zero if the tax rate is so high that nothing is sold, which is distance t_z in Figure 3.17.

The critical tax rate t_z can be found graphically on Figure 3.18. As the tax increases the $s^*(p_c)$ curve shifts upwards, the e_t point slides up the demand curve, and quantity declines. For our numerical example, $t_z = 16$ because with this large a tax point e_t reaches the y-intercept, quantity sold is zero and, therefore, so is tax revenue. To find t_z analytically, note from equations (44) and (45) that $q^e(t) = 0$ yields

$$t_z = \frac{q^e(0)}{\alpha} = \frac{d_0}{d_1} - \frac{s_0}{s_1} \,, \tag{47}$$

where d_0/d_1 is the y-intercept of the demand curve and s_0/s_1 is the y-intercept of the supply curve.

To find the revenue maximizing tax rate, call it t^0, we set the first derivative of Laffer function (46) equal to zero:

$$\frac{dT_r(t)}{dt} = q^e(0) - 2\alpha t = 0 \,. \tag{48}$$

We find with the aid of (47) that

$$t^0 = \frac{q^e(0)}{2\alpha} = \left(\frac{d_0}{d_1} - \frac{s_0}{s_1}\right)/2 = \frac{t_z}{2} \,. \tag{49}$$

and

$$q(t^0) = \frac{q^e(0)}{2}\,. \tag{50}$$

The revenue maximizing tax rate is half the rate yielding zero tax revenue, and the resulting equilibrium quantity is half the level that would be sold if the tax were abolished, provided the demand and supply functions are linear.

Equation (49) reveals that imposing the revenue maximizing tax cuts sales to half what they were without a tax. Equation (50) implies that the slopes of the demand and supply curves do not affect the size of the revenue maximizing tax. The revenue maximizing tax is 1/2 the gap between the y intercepts of the demand and supply curves, or $(18 - 2)/2 = \$8$ for the corn example, as shown in Figure 3.18.

The Laffer curve only shows that there is a possibility that taxes may be so high that more tax revenue could be raised by cutting tax rates. The difficult question to answer, and it is an empirical question, is whether the current tax rate is above or below the critical revenue maximizing value t^0. Laffer thought that tax rates were above the level yielding maximum tax revenue, and his argument was the cornerstone of President Reagan's *Supply-Side Economics* program. The hope was that cutting taxes of the

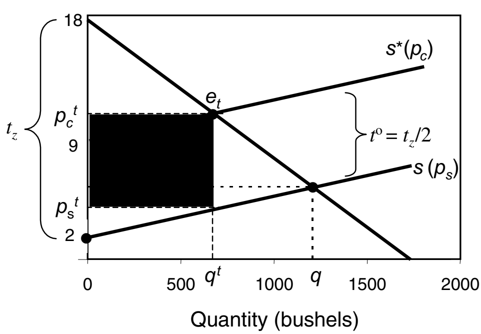


Fig. 3.18. The Revenue maximizing tax
When the demand and supply curves are linear, the revenue maximizing tax rate t^0 is one half of the tax rate t yielding zero tax revenue. This means that the revenue maximizing tax rate depends only on the y-intercepts of the demand and supply curves. In contrast, the division of the tax between the buyer and the seller depends only on the relative slopes of the demand and supply curves and not their intercepts.

Note that the revenue maximizing tax cuts in half the quantity sold from q to q .

well to do would unleash so much investment that the resulting surge of output would generate enough revenue to finance the tax cut. Critics objected that economists had known, at least since Ricardo wrote in the early 19th century, that it was theoretically possible for tax rates to be above the level yielding maximum tax revenue. The problem with the case Laffer built for supply-side economics was that he did not provide convincing empirical evidence that tax rates were indeed above the critical level yielding maximum tax revenue.

3.8.3 *Consumer surplus, producer surplus, and the excess burden*

The pain inflicted on consumers by a tax is measured by the reduction in consumer surplus, shown on Figure 3.19. Before the tax was imposed, consumers enjoyed the surplus indicted by the large triangle with vertices as points D, e and p, or a total of $(18-6)\times 1200/2 =$ \$7,200. After the \$8 tax is imposed, the price paid by consumers rises by \$6 to $p_c^t =$ \$12 and consumer surplus is the smaller triangle D, e_c^t, p_c with area $(18 - 12) \times 600/2 =$ \$1,800. Thus the lost consumer surplus is \$7,200 – \$1,800 = \$5,400, which is represented on the graph by the area of trapezoid p_c^t, e_c^t, e, p.

Suppliers are also hurt by the tax. Estimating their loss will be tedious, but we break the argument into parts to ease the pain.

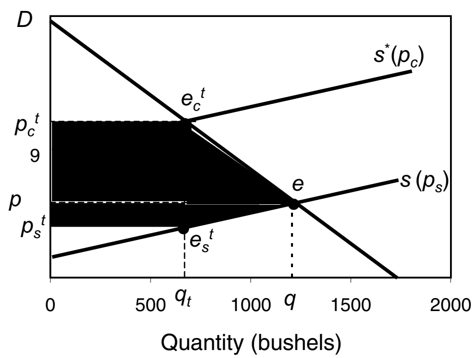


Fig. 3.19. Tax reduces both producer and consumer surplus
The lost consumer surplus resulting from the tax is measured by trapezoid p , e , e, p. The lost producer surplus is measured by trapezoid p, e, e , p . Their combined loss is the sum of these two trapezoids p , e , e, e , p .

1. Suppliers lose sales revenue. Before the tax they enjoyed sales revenue of $6 \times 1{,}200 = \$7{,}200$, which is represented on Figure 3.19 by the area of rectangle p, e, q, 0. After the tax is imposed their sales revenue is only $4 \times 600 = \$2{,}400$, represented by the area of rectangle p_s^t, e_s^t, q_t, 0. Thus, their sales revenue is down by \$7,200 – \$2,400 = \$4,800.
2. Fortunately, this loss of sales revenue is partially offset by the fact that the producers, because they have less to produce, will save on production costs. We shall argue that the reduced production cost is represented by the trapezoidal area below the supply curve: e_s^t, e, q, q_t; i.e., $(6+4)(1200-600)/2 = \$3{,}000$. The point is that the upward slope of the supply curve reflects the fact that the cost of producing an extra unit of output increases as output rises: The cost of producing the 600th unit was \$4.00; if it had been less, it would have been profitable to supply it when the price was below \$4.00; if it was more than \$4.00, that 600th unit would not have been supplied at a price of \$4.00; similarly, the added cost from producing 601 rather than 600 units must be $\$4 + 2/600$; otherwise, it would not have been supplied at that price; and so on, up to the 1,200 unit, which is supplied only when the price reaches \$6.00 because that is the added cost of producing it. This all adds up to a cost saving for producers of \$3,000.
3. The net loss experienced by producers as a result of the \$8.00 tax is a reduction in profits equal to the excess of the lost revenue of \$4,800 over the cost saving of \$3000, or \$1,800.

This reduction in profit is often referred to as the loss in *Producer surplus*. Producer surplus, which is the producer's analogue of consumer surplus, is represented by the trapezoidal on the graph that is above the supply curve and below the horizontal price line. The reduction in producer surplus generated by the tax cut is represented on Figure 3.19 by the trapezoid p, e, e_s^t, p_s^t with area $(6-4)(1200+600)/2 = \$1,800$.

The total pain generated by the tax is the sum of the cut in consumer surplus plus the cut in producer surplus, or \$5,400 + 1,800 = \$6,200; i.e., the polygon p_c^t, e_c^t, e, e_s^t, p_s^t.

The excess burden triangle (aka dead weight loss)

The fundamental point is that the magnitude of the combined loss suffered by producers and consumers is greater than the tax revenue generated by the tax. This is made clear in Figure 3.20. The combined losses inflicted on consumers and producers, reproduced from Figure 3.19, is the polygon p_c^t, e_c^t, e, e_s^t, p_s^t. The tax revenue, $t \times q_t$, is the rectangle p_c^t, e_c^t, e_s^t, p_s^t. The shaded triangle, e_c^t, e, e_s^t is the *excess burden of the tax.* It is the excess of the loss inflicted on producers and consumers over the amount of tax revenue collected, where the loss is measured by the reduction in consumer and producer surplus.

Figure 3.20 shows that the excess burden, call it E_b is

$$E_b = -\frac{t[q^e(t) - q^e(0)]}{2}, \tag{51}$$

or with the help of equation (45),

$$E_b = \frac{\alpha t^2}{2}; \tag{52}$$

that is to say, the excess burden goes up with the square of the tax rate.

If the revenue maximizing tax is imposed, the excess burden will equal one half the tax revenue. This is clear from Figure 3.18 — since $q_t = q - q_t$, the excess burden triangle and the tax revenue rectangle must have the same base and altitude, which means that the excess-burden triangle has half the area of the tax-revenue rectangle.

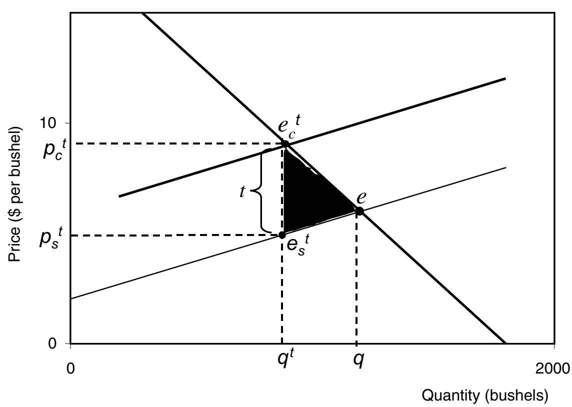


Fig. 3.20. Excess burden (aka deadweight loss) inflicted by the tax
The area of the excess burden triangle is the excess of the value of the loss imposed on consumers and producers as a result of the tax over the value of the tax revenue collected.

Note the following:

- No one gets the excess burden. It is a loss generated because the tax causes an inefficient reduction in production. It is an unintended consequence of imposing the tax.
- Our analysis shows that the total burden of a tax can be much greater than the dollar magnitude of the tax bill, and this is so quite apart from the time, accounting and legal costs of tax compliance.
- In considering whether a government activity or program is worthwhile, it is not enough to establish that the value of the program's benefits is at least equal to the program costs that the government will have to finance with taxes. The benefit should at least equal the dollar cost of the project *plus* the excess burden generated by the tax.

Our analysis of the excess burden relied on the assumption that the objective is to maximize the sum of producer surplus (profits) plus consumer surplus. There are at least three reasons why this may at times not be the primary objective:

1. Consider the case of gasoline: The tax on gasoline not only raises revenue. It also discourages the use of automobiles, thereby reducing pollution and traffic congestion. The benefit of the tax includes the value to the general public of the reduced pollution and traffic congestion as well as the tax revenue.
2. The argument also assumed that consumers are fully cognizant of the benefits they would realize from consuming the commodity. Consider the case of cigarettes. It is argued that a higher tax on cigarettes should be imposed, but not only to raise revenue. Many advocates of a higher tax on cigarettes argue that it would raise their price and save lives by discouraging teenagers from starting to smoke. This argument, sometimes disparaged as "paternalistic," assumes that, contrary to Bentham, the consumers are not the best judges of their own welfare.
3. The analysis rested on the unrealistic assumption that a dollar is equally valuable to the producers, the consumers, and those who will benefit from the tax revenue. We should worry about distributional issues: Will the tax target the poor because it is based on a commodity that is more heavily consumed by those with low incomes? Will the benefits of the activity financed by the tax go primarily to the wealthy or primarily to the poor?

3.8.4 *Tariffs, quotas and the gains from trade*

Tariffs are taxes imposed on goods imported from abroad. They have traditionally been a primary source of revenue for developing nations. Thus it was that prior to World War I the United States relied on tariffs as the primary source of government revenue. It is easier to collect taxes in a nation's harbors rather than in the countryside. But tariffs can and have been used to protect domestic producers from competition as well as to raise revenue.

How tariffs work is shown in Figure 3.21 for the simplified case in which the importing country is so small that its imports have no significant effect on the world price, p_w. We consider a particular commodity, say shoes. In the absence of tariffs the price p of shoes in the importing country would equal the world price, $p_w = \$60$ (we neglect shipping costs, an unnecessary complication). Domestic production is determined by point h on domestic supply curve $s_d(p)$. At this price consumers will want to consume $q_d(p_w)$, as indicated by point d on the demand curve. Imports equal to the distance

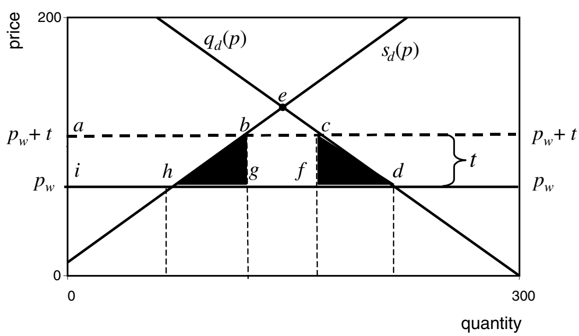


Fig. 3.21. A tariff on imported shoes
In the absence of foreign trade, equilibrium would be established at point e, where the demand and supply curves intersect.

If the world price is p (including shipping costs), the domestic price will be beaten down to p, domestic producers will supply at h, consumers will purchase at point d, and imports will be equal to the gap d–h between demand and supply.

When a tariff t is imposed the price rises to $p + t$, domestic production is determined at point b, demand at point c and imports are reduced to c–b.

The tariff yields revenue equal to the area of rectangle b, c, g, f. Domestic producers gain the a, b, h, i trapezoid from the tariff. But customers lose consumer surplus represented by trapezoid a, c, d, i. The losses of the consumer exceed the gains of government and producers by the area of two cross-hatched excess burden triangles: h, b, g and c, d, f. This is the excess burden (deadweight loss) induced by the tariff.

h–d make up the excess of domestic consumption over domestic production. Now suppose shoe producers successfully lobby the government to impose a tariff of \$50, which pushes the domestic price \$50 above the world price to \$110. Point b indicates the increased domestic production but consumption is cut to c and imports are reduced to the distance c–b. There are gainers and losers:

- Producers surplus increases by the trapezoid a, b, h, i.
- Consumers surplus declines by the trapezoid a, c, d, i.
- The government gains tax revenue rectangle b, c, g, f.
- The excess burden (deadweight loss) from imposing the tariff is given by the sum of the two remaining triangles, g, b, h and c, d, f.[13]

The important point to note is that the country imposing the tariff is hurting itself because the gains enjoyed by producers plus the tariff revenue add up to less than the lost of consumer surplus.

When President Hoover signed into law the Smoot-Hawley Tariff in 1930, it was not to raise revenue. He signed this tariff, the highest in the modern history of the United States, in a desperate attempt to move the economy out of the Great Depression by protecting domestic produces from foreign competition. This massive tariff did restrict imports and it did impose a deadweight loss; but it was no remedy for unemployment. Other countries, also suffering from massive unemployment, retaliated by raising their tariffs on imports from the United States, which generated unemployment in the United States by reducing American exports. The Smoot-Hawley Tariff contributed to the disruption of world trade but it was no remedy for the Great Depression.

Quotas: Import quotas are an alternative to protective tariffs. If instead of a tariff the government had ruled that only the quantity b–c on Figure 3.21 could be imported into the country, the result would have been essentially the same. Domestic production, supplemented by the quota of imports, would have equaled demand at the same equilibrium price. While the respective gains and losses of producers and consumers would be the same, the government would lose the tariff revenue!

While quotas, because they do not yield revenue to the government, may seem even worse than tariffs, that does not keep quotas from being

[13]The details are slightly more complicated but the results are essentially the same in the "large country" case in which the tariff has an impact on the world price.

enacted. As but one example, the quota on importing sugar into the United States means that its price in the U.S. is often twice and sometimes three times the price that sugar sells for on world markets. This helps American sugarcane and sugar beet farmers at the expense of the consumer. It also helps corn farmers because bakers and soft drink manufacturers are led by the high price of sugar to use corn syrup instead. It is left to influential politicians to decide how much each country gets of the sugar quota, which is the privilege of shipping sugar to the United States at substantially above the world price.

In the 1970s and much of the 1980s the United States encouraged Japan to impose Voluntary Export Restraints on their car manufacturers, which constituted an indirect quota on the shipment of Japanese cars into the United States. This may have increased employment in the American auto plants, but it also raised the price of cars for the American consumer. Restricting imports is an expensive technique for remedying unemployment.[14]

3.9 Markets versus the ballot box

In a democracy we like to think that major decisions are made through the voting process. Sovereignty belongs to the voters, but tempered by the operation of political parties, tilted by campaign contributions and influenced by lobbyists. In a market economy, many decisions are made by consumers through their decisions about what they will consume, but tempered by a variety of influences, including advertising and marketing as well as the limits of financial resources.

Some use the concept of "consumer sovereignty" in talking about how the market system works. This concept emphasizes the role of the consumer in influencing what is produced. While we sometimes speak of consumers casting their dollar (or their Euro or their yen) votes in the marketplace, there are obvious differences between the political process and the market. Some decisions are made by the political process while others are made in the marketplace.

[14] An American tariff on Japanese cars might have been better than the Voluntary Export Restraints imposed by the Japanese. The tariff, like the Voluntary Export Restraints, would also have raised the prices of cars, but at least it would have yielded revenue for Uncle Sam. The Voluntary Export Restraint policy imposed by the Japanese government may have helped the Japanese car makers charge a higher price by inhibiting competition.

- In a democracy each citizen usually has precisely one vote. In the marketplace, there is tremendous inequality in purchasing power — those who inherit special skills or a fortune and those who invest more in education and training get to bring more dollar votes to market.
- Some types of decisions are obviously political. We cannot determine whether there shall be war or peace in the marketplace. I cannot buy a little war at the supermarket and you cannot purchase a moment's peace at the gas station. Also, lobbyists and politicians are not supposed to pay for votes. Other decisions obviously belong in the marketplace. We allow the individual consumer to decide whether to eat brown bread or white and what color shirt to wear. In part, what goods are allocated by the political process and what goods are allocated by the market place depend on the nature of the commodities. War or peace is an all or nothing decision affecting every citizen. What one wears and what books one reads we generally regard as the personal business of the individual consumer.
- Governments regulate markets. Markets will not function well if the right to private property is not protected or if contracts are not enforced. In part because a democracy requires an educated and informed electorate, the state has an interest in making sure that its citizens receive a proper education by providing for universal public education and requiring children to attend school.
- Some cynics object that campaign advertising and propaganda mislead voters. Similarly, some object that marketers mislead consumers into buying useless gadgets. Others argue that advertising is educational, contributing to an informed electorate and knowledgeable consumers.

There is no end of controversy about where to draw the line between the political and the economic sectors. What type of restrictions should the government place on gun ownership? Should those under 21 be allowed to buy and consume alcoholic beverages? Should the sale of marijuana be legalized? Should you be allowed to sell one of your kidneys? Should your university sell academic degrees or charge higher tuition for courses that are oversubscribed? The precise roles assigned to the government and to markets vary not only among countries but also from one generation to the next.

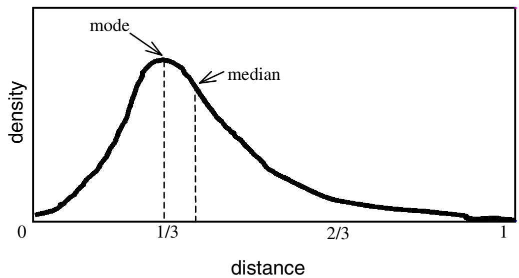


Fig. 3.22. Hotelling's Beach
More customers are located 1/3rd of the way along the beach at the mode than at any other point on the beach.

Half of the customers are located to the left of the median and half are located to the right. If the first ice-cream vendor to arrive at the beach locates at the median, the other vendor will get less than half of the ice-cream customers regardless of whether she locates to the left or the right of the first vendor. If the first vendor locates anywhere else, the second vendor will get at least one half the customers.

Hotelling's model

In the 1920s, famed economist-statistician Harold Hotelling developed a model that explained certain limitations to both political and market processes.[15] Hotelling started his analysis by considering a popular beach. Figure 3.22 shows how bathers are distributed along the beach. They are most densely clustered around the 1/3 mile point, perhaps because it is closer to the parking lot or perhaps because of a scenic attraction. Now suppose you are going to sell ice cream on the beach. Where should you setup your ice-cream stand on the beach, which is one mile long? In making this decision you are aware from past experience that Lovell will arrive at the beach with his ice-cream stand later on in the morning. You also know that customers will go to the ice-cream stand that is closest.

Should you set up your stand at the 1/3-mile point where most of the customers are clustered? That may seem like an obvious answer — statisticians use the term "mode" to refer to the point of greatest density. But if you select the mode for your ice cream stand, Lovell will locate just a little bit to your right. Because the distribution has its longer tail to the right (i.e., "skewed to the right" in statistician's speak), a majority of the customers are to the right of your position at the mode. Lovell has capitalized on this by locating just to your right. He will get all the customers to

[15] Harold Hotelling, "Stability of Competition," *Economic Journal*, 1929, pp. 41–57.

his right plus those to the right of the point halfway between his ice-cream stand and yours. In this business, first-starters do not have an advantage. You had first pick of the locations, but late arrival Lovell comes out ahead.

Your best defensive strategy is to locate your ice-cream stand at the point where 1/2 of the customers are to your left and 1/2 are to your right — statisticians call this point the *median* of the distribution. If you pick this point, you will get at least half the customers regardless of where latecomer Lovell locates his stand. Assuming you do pick the median, the best latecomer Lovell can do is to locate as close to your stand as possible. Locating further away would allow you to claim even more customers. Thus the Hotelling model predicts that there is a tendency for all the sellers to locate at the median of the customer distribution.

Hotelling argued that the tendency for the two ice cream stands to locate next to each other at the median constituted a form of market failure; that is to say, the market mechanism fails to yield an efficient allocation of resources. The tendency for the stands to locate side by side increased the distance consumers have to travel to purchase their ice cream. From the point of view of efficiency, it would be better to have the ice-cream stands at the 1/4th and 3/4th mile point so that no one would have to walk more than a quarter mile to get ice-cream. And Hotelling argued that this tendency toward excessive clustering is not only true of ice cream. Furniture stores and camera stores often cluster in particular areas of the city. You may have noticed that Home Depot and Lowe's often set up stores close to each other, contributing to longer trips for those buying building supplies. Hotelling argued that radio station programming suffered from the same problem. There is a tendency for stations to cluster their programming around the median of the distribution of listener tastes, focusing on popular music and neglecting to service the minority of listeners who prefer classical. Here is a case for Public Radio and Public Television — they cater to the interests of minority listeners whose interests are under-served on private broadcast TV.

Hotelling claimed that his analysis also explains why political parties are so much alike. Suppose that Figure 3.22 represents not locations on a beach but the spectrum of political opinion. Then the incumbent politician seeking reelection had best locate his political platform to reflect the preferences of the median voter — half the voters will want a candidate that is further to the left and the other half will want a candidate further to the right. But if the incumbent takes her stand at any point but the median, the challenger will capture the hearts of more than half the voters by locating between the incumbents' platform and the median voter. Hotelling's model of the *median voter* explains why two party political systems often fail to provide voters with a broader range of genuine choices.

Hotelling's classic contribution explained certain limitations of both market and political processes. His contribution led to the development of Public Choice, the academic discipline for those who seek to model the way in which the political process works. While much progress has been made in this area of research since Hotelling developed his simple model, his contribution is important because it suffices to show that neither political nor economic processes always yield ideal solutions.

Summary

This chapter developed the model of supply and demand in order to explain how prices are determined.

1. Prices adjust so as to equate the forces of demand and supply, represented by the demand and supply functions. The equilibrium price is that price at which quantity demanded is equal to quantity supplied.
2. The demand and supply apparatus is used to analyze the determination of foreign exchange rates, the effects of the minimum wage on unemployment and the impact of excise and sales taxes on prices.
3. Elasticity is a unit-free measure of the responsiveness of demand and supply to price changes. If the demand for a commodity is inelastic, a decrease in quantity supplied will lead to such a large increase in price that consumers will end up spending more money for a smaller supply!
4. The analysis developed in this chapter assumed that markets are competitive. There are so many buyers and sellers that no one can influence the price, both buyers and sellers being price takers. Left for Chapter 6 is a discussion of monopoly and other forms of market structure.
5. Partial derivatives, a key calculus concept that we will be using repeatedly in subsequent chapters, were explained within the context of demand functions. If consumption depends on both income and price, $q(p, y)$, then $\partial q/\partial p$ is the change in quantity in response to a change in price, given income; similarly, $\partial q/\partial y$ is the response of quantity to a change in income, given prices.
6. Consumer surplus, the excess of the value of a commodity to consumers over what they pay for it, can be measured by the area under the demand curve. This concept was used to explain why taxes generate an excess burden (aka deadweight loss).

7. We compared the workings of the market with democracy. One can say that in the market consumers cast dollar votes when they decide what to consume. In a democracy its one-person one-vote while in the marketplace those who are the most industrious, who have greater skills, or inherit the most get to stuff the ballot box. Harold Hotelling's model of the political process explains why in a democracy the median voter may dominate the ballot box.

Key Concepts

arc-elasticity, 77
black market, 71
competition, 58
complementary commodities, 72
consumer surplus, 68
debtor nation, 97
demand function, 59, 63
dollar, strong, 90
elastic demand, 79
Engel's Law, 83
equilibrium price, 63
Euro, 96
excess burden of the tax, 107
excess demand, 63
excess supply, 63
exchange rate
  fixed versus floating, 95
foreign exchange rates, 90
income elasticity, 82
inelastic demand, 79
inferior goods, 72
inverse demand function, 66
Laffer curve, 102
Law of Demand, 63
market demand, 66
market equilibrium, 60
market supply, 66
median voter, 115
net International Investment Position, 97
partial derivative, 74
point-elasticity, 78
price ceiling, 63
price setter, 67
price taker, 67
producer surplus, 106
quotas, 110
substitute commodities, 72
supply elasticity, 85
supply side economics, 104
tariffs, 109
tax incidence, 100
unit elastic, 79

Exercises

1. There are 400 widgit consumers and 100 widget producers in Never-Never Land. Each of the 400 consumers has demand curve

$$q_i = 100 - p/4\,.$$

Each producer has supply curve

$$s_i = 4(p - 5)\,, p > 5\,.$$

a. Determine the equations for *market* demand and *market* supply. Plot them on a neat graph.
b. Is the equilibrium price \$76, \$84, \$316, or \$____? Solve for the equilibrium quantity that will be sold in this market.
c. How much excess supply would there be if the government enforced a law prohibiting the sale of widgets at any price below \$100?
d. How many widgets would the government have to buy in order to push the market price up to \$100?
e. What is the (point) elasticity of demand at the equilibrium price you determined in *b*?

2. Based on a sample survey of 17,592 college students, economists Frank Chaloupka and Henry Wechsler concluded that a "10 percent increase in price would reduce cigarette consumption among college smokers by 7 to 8%."[16] What does this imply about the price elasticity of smoker's demand for cigarettes?
3. According to a study by Michael Grossman and Sara Markowitz, a 10% increase in the price of beer would reduce the number of college and university students involved in various kinds of violence by 4%.[17] What does this imply about the elasticity of campus violence with respect to the price of beer?
4. Suppose the demand function for eggs is $q = 100 - 10p$.

   a. What is the price elasticity of demand when $p = 5$?
   b. What is the price elasticity of demand when $p = 10$?

5. In Never-Never Land the annual demand for oil is

$$q = 500,000p^{-1/2}M^{2/3}\,,$$

where p is the price per barrel of oil and M is income.

   a. What is the price elasticity of demand?
   b. What is the income elasticity of demand?

---

[16] Frank Chaloupka and Henry Wechsler, "The Economics of Smoking," NBER Working Paper #7262, July 1999.
[17] Michael Grossman and Sara Markowitz, "Alcohol Regulation and Violence on College Campuses," NBER Working Paper #7129, May 1999.

6. Taxing widgits (question 1 continued)

   a. What would happen to quantity sold and the price paid by consumers if the governor imposes a \$100 tax on each widgit sold in Never-Never Land?

   b. Find the inverse demand and supply functions for the widgit industry. Then determine how large an excise tax would just suffice to drive all the widgit manufactures out of business?

   c. Determine the excise tax that would maximize the total tax revenue collected by the governor.
   Hint: Consider equations (49) and (50).

   d. How much tax revenue will be collected if the revenue maximizing tax is imposed? Show on your graph the tax revenue, the reduction in consumer surplus, the reduction in producer surplus and the dead weight loss resulting from the tax.

7. Trading widgets (Build on question 1. Forget about taxes and price supports.)

   a. Suppose that importers bring widgets in from aboard and sell them for \$50, which is the world price plus shipping costs. What will happen to the widget market (domestic production, total consumption, imports, consumer surplus)? Explain.
   Suggestion: redraw your original graph and show what happens.

   b. How would trade be affected if the government placed a \$20 tariff on widgets? How much revenue would the government collect?

8. Price controls: Suppose that the government imposes a price cap of \$4.00 on the corn market graphed on Figure 3.3. The demand function was specified by equation (4) and the supply function by equation (5).

   a. Determine how large a shortage of corn (excess of demand over supply) will be generated by the price ceiling.

   b. Calculate the change in consumer surplus resulting from the price ceiling. Explain who gains and who loses from this price ceiling.

   c. Would consumers gain or lose more from the price ceiling if the supply function had been less elastic? Explain.

9. Elasticity

   a. Verify that the demand curve $q = 1000/p$ has unit elasticity ($\eta = 1$) at *every* price!

b. The demand for oranges is $q = 100 - 10p + 20y$, where p is price and y is income. Determine $\partial q/\partial p$ and $\partial q/\partial y$. Then solve for the price elasticity and the income elasticity of demand as functions of p and y.

c. The demand for apples is $q = 100y^{0.5}p^{-2}$. Determine $\partial q/\partial p$ and $\partial q/\partial y$. Then solve for the price and income elasticities of demand as a function of price.

d. For each of the demand curves on Figure 3.23 determine the demand elasticity when the price is \$500. Then determine on Figure 3.24 the elasticity of demand on each demand curve when the quantity sold is 500. Explain.
Hint: Use the quick trick explained on Figure 3.8.

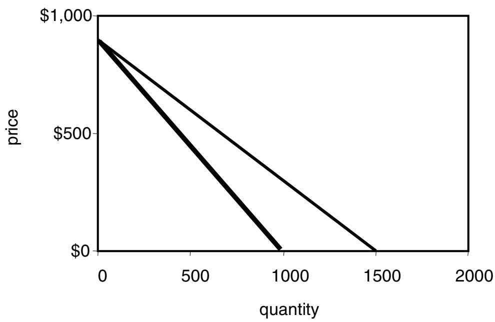


Fig. 3.23. Elasticity exercise

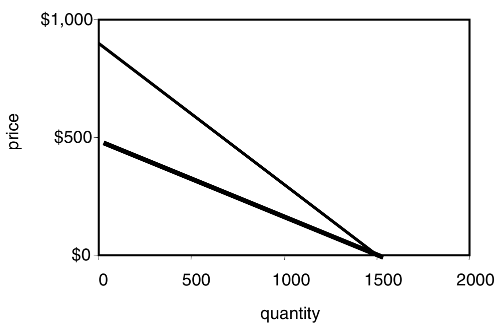


Fig. 3.24. Elasticity exercise (Continued)

e. Consider the demand function $q = 3p^{-2}$. Now correct any errors in the following calculations:

$$dq/dp = -6p^{-3} = -2q/p\,;$$

$$dp/dq = -p/2q = -p^{-3}/6\,.$$

10.* At the meeting of your company's Board of Directors the President reports that the demand elasticity for their major project is $\eta = -1/2$. One of the Directors objects that the President should be fired because the firm cannot be maximizing profits. How did the Board Member know that profits were not being maximized?

4

Maximizing Satisfaction

4.1 Introduction

Economists often try to distinguish between positive and normative economics. *Positive economics* has to do with the world as it is and how it might be affected by different types of government policy, such as a change in the tax laws or government regulations. *Normative economics* has to do with the world as it ought to be. It involves ethical principles and norms of fairness and equity, such as the fundamental question of how income *should* be distributed — who should get how much? This chapter considers both positive and normative issues in developing a theory of consumer behavior.

First we will consider the question of how income ought to be distributed, which is a normative issue. Utilitarian philosopher Jeremy Bentham's [1748–1832] claim that income should be distributed so as to maximize the sum total of happiness will be contrasted with John Rawls' [1921–2002] argument that income should be distributed to maximize the position of the least advantaged.

Then we will derive some positive implications from the assumption that consumers spend their income so as to maximize their personal satisfaction. Here are two rather surprising results:

- A maximizing consumer might buy *more* of a commodity when its price goes up, which violates the Law of Demand. (Recall from Chapter 3 that the Law of Demand is the proposition that when a good's price rises consumption must decline).
- If inflation causes all prices and wages to increase by the same percentage (e.g., 75%), a maximizing consumer will purchase precisely the same amount of each good as before the inflation.

Finally, we shall return to the issue of free trade, which was considered in Chapter 2, where the problem was to make efficient decisions about who should produce how much of what good without worrying about how the increased output would be divided. In this chapter we shall elaborate on the concept of economic efficiency.

4.2 One good and two consumers

How should a fixed quantity of a scarce resource be divided between two individuals? This question, not nearly as simple as it looks, focuses on fundamental issues of *equity* or *fairness*.

Example

Consider an elderly couple writing a will dividing the family estate between their two children. In ancient times the usual practice was to follow the principle of "primogeniture," passing the entire estate to the oldest son. It is said that the rule of primogeniture was functional, for it prevented the partitioning of large family land-holdings into smaller and smaller units that were too small to farm efficiently. If it were not for primogeniture, royal families might have subdivided their country among their children, which, over generations, would have led to the gradual division of once powerful nations into smaller and smaller principalities.[1]

When the American colonies freed themselves from British rule, primogeniture was cast aside in favor of an alternative principle — equal division of the estate among all children. This rule is likely to be judged fairer, at least by all but first-born males. But suppose that one of the children, perhaps because of a serious disability, may need life long continuing support. Should the disadvantaged receive a larger share of the estate?

Obviously, such issues of equity and fairness are not easily resolved. And equity issues are not confined to individual families but arise as matters of national concern, as in the perennial debate about the extent to which the more well-to-do should be taxed to help provide a safety-net to support those judged to be in need. The principle of utilitarianism, which provides a guidepost for resolving questions of equity, has been of special interest to economists.

4.2.1 *Utilitarianism*

The English philosopher *Jeremy Bentham* [1748–1832] founded his philosophy of *utilitarianism* on the argument that society should strive to maximize the "sum total of satisfaction." Bentham argued that his rule would provide useful guidance in practical situations if only one would, as a mental exercise, stand aside like a *disinterested spectator* when called upon to make judgments on equity issues. Some find his argument appealing; others have their doubts. One problem with utilitarian doctrine is that putting it into practice presumes that one can measure satisfaction or happiness. But even if psychologists could provide us with a meter for measuring happiness,

[1] Of course, estates can be preserved intact without favoring the first born son. Indeed, German National Socialist Law under Adolph Hitler provided that the estate would go to the youngest son.

there are other problems with the theory. For the moment we shall pretend that happiness can be measured and compared among individuals. This will help us focus on certain surprising implications of Bentham's principle by considering a grossly simplified example.

Suppose that we are responsible for allocating a fixed quantity of a single commodity, call it X, between Dick and Jane. If X_D denotes the quantity of the good that Dick consumes and X_J the quantity consumed by Jane, then we must obviously have

$$X = X_J + X_D\,, \quad X_J \geq 0\,, \quad X_D \geq 0\,. \tag{1}$$

Now we shall suppose that the satisfaction or "utility" enjoyed by Jane depends only on what she consumes — this dependence of satisfaction on the quantity of X consumed is represented by a *utility function*, denoted $U_J(X_J)$. Similarly, $U_D(X_D)$ is the function showing how Dick's utility depends on the quantity of good X that he consumes. Now Jeremy Bentham would have us maximize the sum total of satisfaction

$$U_\Sigma(X_J, X_D) = U_J(X_J) + U_D(X_D)\,, \tag{2}$$

subject to the constraint

$$X_J + X_D \leq X\,. \tag{3}$$

We must consider this maximization problem in detail.

4.2.2 *Marginal utility*

Before evaluating the solution to Bentham's maximization problem, let us digress to consider two basic properties that economists believe utility functions can be expected to satisfy. First, it seems reasonable to presume that *more is better*; in particular, the more of good X that Jane gets to consume the better off she will be. Assuming that her utility function is differentiable, this implies

$$\frac{dU_J(X_J)}{dX_J} > 0\,. \tag{4}$$

This derivative, called *marginal utility*, is assumed to be positive — more is better.[2] Second, it is often assumed that utility functions must satisfy

[2]In order to work with derivatives it will be assumed that quantities can be measured in continuous units, like gasoline, rather than in discrete units, like cookies that crumble.

the *Law of Diminishing Marginal Utility*; i.e.,

$$\frac{dU_J^2(X_J)}{dX_J^2} < 0\,. \tag{5}$$

This second property means that the increase in utility enjoyed by Jane from gaining an additional unit of X will be smaller if she already has a large quantity of the commodity. For example, that first apple you eat today may be delicious but the fifth apple of the day will provide a much smaller incremental gain in total satisfaction.

To be concrete, assume for the time being that Jane's utility depends on the log of the quantity of X that she consumes;[3] i.e.,

$$U_J(X_J) = \ln(X_J)\,, \tag{6}$$

where "ln" denotes the logarithm to the base $e \doteq 2.718$. This yields:

$$\frac{dU_J(X_J)}{dX_J} = \frac{d\ln(X_J)}{dX_J} = \frac{1}{X_J} > 0\,; \tag{7}$$

i.e., marginal utility is positive, satisfying condition (4). Furthermore,

$$\frac{d^2U_J(X_J)}{dX_J^2} = -\frac{1}{X_J^2} < 0\,, \tag{8}$$

which means that the Law of Diminishing Marginal Utility is also satisfied by this utility function.

4.2.3 *Maximizing total satisfaction*

Now as a special case of Bentham's maximization problem, suppose that $X = 100$ cookies and that Dick's utility function is $U_D(X_D) = \ln(X_D)$, which is identical to Jane's. Then the problem of allocating resources so as to maximize the sum total of happiness as specified by equation 2 reduces to the task of maximizing

$$U_\Sigma(X_J, X_D) = \ln(X_J) + \ln(X_D) \tag{9}$$

[3]The assumption that utility is proportional to the log of income has a long history, having been invoked by Swiss mathematician Daniel Bernoulli [1700–1782] in his studies of the economics of gambling and insurance. The assumption is adopted here solely for expository convenience, although it was at one time thought to be justified by the Weber-Fechner law of "psycho-physics", the proposition that sensations are proportional to the log of the stimulus. Such phenomena as sound (decibels) and earthquakes (Richter scale) are measured with a logarithmic scale.

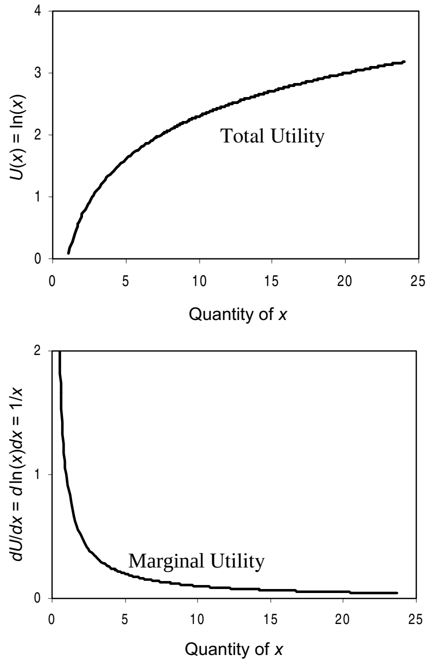


Fig. 4.1. Total and marginal utility
If total utility is $U = \ln(x)$, then marginal utility is $dU/dx = d\ln(x)/dx = 1/x$. Observe that $d^2U/dx^2 = -x^{-2} < 0$, as required by the Law of Diminishing Marginal Utility.

subject to the constraint

$$X_J + X_D = X = 100\,, \tag{10}$$

where $X = 100$ is the total supply of the scarce commodity. Possible allocations of the 100 cookies are plotted on Figure 4.2. Table 4.1 evaluates the utility generated by a few alternative ways of allocating the 100 units of X. Figure 4.3 shows the tradeoff between the utility enjoyed by Jane and that enjoyed by Dick — the curve is dubbed the "utility possibility frontier." The *utility possibility frontier* reveals the painful truth that given the limited supply of resources we can only increase the utility of one individual, say Dick, by transferring cookies to him from Jane, which would reduce her utility.[4]

[4]Note that $U(X) = \ln X$ implies that $X = e^{(\quad)}$. Therefore, substitution reveals $U \quad = \ln X \quad = \ln(100 - X \quad) = \ln(100 - e^{\ J})$.

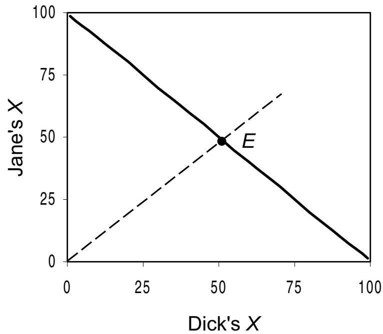


Fig. 4.2. Feasible allocations
If there are 100 units of X to be allocated between Dick and Jane, the feasible allocations are indicated by the negatively sloped line. All points on the dotted line involve equal division. Point E indicates the equal division of the 100 available units of X, each receiving 50.

The *method of substitution* is one way to solve this constrained maximization problem.[5] The first step, recalling $X_D + X_J = 100$ from equation (10) is to note

$$X_D = 100 - X_J\,, \tag{11}$$

which can be used to eliminate the unknown variables X_D in equation (9). Substituting into (2) yields a new function, call it $U^*_\Sigma(X_J)$, that explains the total satisfaction enjoyed by Dick and Jane with a single variable, X_J:

$$\begin{aligned} U^*_\Sigma(X_J) &= U_J(X_J) + U_D(100 - X_J) & (12)\\ &= \ln(X_J) + \ln(100 - X_J)\,. & (13) \end{aligned}$$

Note that this function differs fundamentally from equation (9) in that it implicitly takes into account both the direct and the indirect effect on utility when X_J increases: an increase in X_J tends to increase directly Jane's utility, but it also leads to a reduction in Dick's utility because the supply constraint dictates that his consumption must be reduced if Jane's is to increase.

[5]Instead of employing the method of substitution, one can use the technique of Lagrangian multipliers to find the maximum of a function of several variables subject to an equality constraint. This procedure, explained in Appendix 4.1, translates the problem of maximizing a function in two variables subject to one constraint into a problem involving three variables but no constraint.

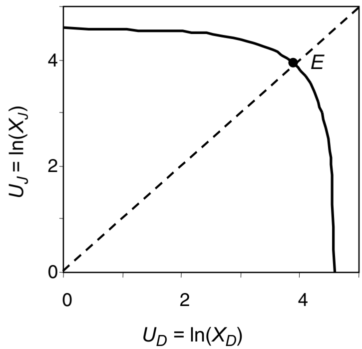


Fig. 4.3. Utility Possibility Frontier, Case 1
The Utility possibility frontier shows possible utility levels for Dick and Jane, given that $U\ \ (X\ \ ) = \ln(X\ \ )$, $U\ \ (X\ \ ) = \ln(X\ \ )$, and $X\ \ + X\ \ = 100$. Point E indicates that Dick and Jane will each enjoy about 3.9 units of satisfaction if each receives 50 units of X.

Table 4.1. Maximizing total satisfaction, Case 1.

X	X	$U\ \ (X\ \ )$	$U\ \ (X\ \ )$	$U_\Sigma(X\ \ , X\ \ ) = U\ \ (X\ \ ) + U\ \ (X\ \ )$
99	1	4.60	0	4.60
75	25	4.32	3.22	7.54
50	50	3.91	3.91	7.82
25	75	3.22	4.32	7.54
1	99	0	4.60	4.60

The next step is to find the maximum of the transformed problem by calculating the derivative of (13) with respect to the unknown X_J:

$$\begin{aligned}\frac{dU^*_\Sigma(X_J)}{dX_J} &= \frac{dU_J(X_J)}{dX_J} + \frac{dU_D(100 - X_J)}{dX_J} \\ &= \frac{d\ln(X_J)}{dX_J} + \frac{d(\ln(100 - X_J)}{dX_J} \\ &= \frac{1}{X_J} - \frac{1}{(100 - X_J)}.^{6}\end{aligned} \tag{14}$$

[6]In calculating $d(\ln(100 - X\ \ )/dX\ \ $ we can make use of the chain rule, $dz/dx = dz/dy\, dy/dx$, by letting $\ln(100 - X\ \ ) = z[y(X)]$, where $z = \ln(y)$ and $y = 100 - X\ \ $. Then $dz/dX\ \ = (100 - X\ \ )^{-1}(-1)$.

Setting this derivative equal to zero, a necessary condition for a maximum, yields

$$\frac{1}{X_J} = \frac{1}{(100 - X_J)} \tag{15}$$

or

$$100 - X_J = X_J\,. \tag{16}$$

Hence, $X_J = X_D = 50$, $U_J = U_D = \ln(50) = 3.91$ and $U_\Sigma = 7.82$.

In this case, maximizing the sum total of satisfaction yields equal shares and equal utility, which seems eminently reasonable.

4.2.4 *Amartya Sen's complaint*

Amartya Sen, in the lecture he presented on the occasion of his being awarded the Nobel Prize in 1998, made a fundamental complaint about Bentham's theory of utilitarianism:

> Bentham's focus — and that of utilitarianism in general — was with the *total utility* of a community. This was irrespective of the distribution of that total, and in this there is an informational limitation of considerable ethical and political importance. For example, a person who is unlucky enough to have a uniformly lower capacity to generate enjoyment and utility out of income (say because of a handicap) would also be given a *lower* share of a given total. This is a consequence of the single-minded pursuit of maximizing the sum-total of utilities.

A slight modification of our simple two-person example will illustrate Sen's objection. Suppose that while Jane still has utility function $U_J(X_J) = \ln(X_J)$, Dick now has utility function

$$U_D = 2\ln(X_D)\,. \tag{17}$$

This says that Dick gets twice as much utility as Jane if we divide the available X evenly between the two, which would yield $U_J(50) = \ln(50) = 3.9$, $U_D(50) = 2\ln(50) = 7.8$, and $U_\Sigma = 11.7$. Is this fair? No, asserts Jane, who may argue that she should receive more since Dick needs less of scarce commodity X to make him happy.

Bentham also objects, but his solution will seem even less fair. With Dick's new utility function, the sum total of satisfaction is

$$U_\Sigma(X_J, X_D) = \ln(X_J) + 2\ln(X_D)\,, \tag{18}$$

which Bentham wishes to maximize subject to the constraint

$$X_J + X_D = 100\,. \tag{19}$$

The resulting utility possibility frontier is plotted on Figure 4.4. Because the tangent line at point B has slope -1, all points on this line yield the same utility sum. All points below it, and this includes all points other than B on the utility possibility frontier, yield a lower sum than B. So B identifies the point Bentham would select for maximizing the sum total of happiness.

To solve this maximizing problem analytically, we substitute as before to obtain

$$U^*_\Sigma(X_J) = \ln(X_J) + 2\ln(100 - X_J)\,. \tag{20}$$

This function shows total utility as a function of X_J, given that X_D adjusts so that all the goodies are used up. To maximize we take the derivative

$$\frac{dU^*_\Sigma(X_J)}{dX_J} = \frac{1}{X_J} - \frac{2}{(100 - X_J)} = 0\,. \tag{21}$$

Hence

$$X_J = 33.3 \quad \text{and} \quad X_D = 100 - 33.3 = 66.7 = 2X_J\,. \tag{22}$$

with

$$U_J(33.3) = \ln(33.3) = 3.5\,, \quad U_D(66.6) = 8.4\,, \quad \text{and} \quad U_\Sigma = 11.9\,. \tag{23}$$

Bentham's dictum that we should strive to maximize the sum total of satisfaction implies, with these utility functions, that Dick should receive more than twice the utility of Jane because of his greater capacity for enjoyment!

What is happening is revealed in part by Table 4.2: because of Dick's greater ability to reap satisfaction from additional units of X, Jeremy Bentham would have us reduce Jane's share to about 1/3 of the available supply of the scarce commodity, which means that her utility is less than half that enjoyed by Dick.

The bizarre allocation generated by this simple example suggests that Bentham's proposition that society should attempt to maximize the sum total of satisfaction is not as appealing as it initially appeared. Maximizing the sum total of satisfaction leads to an allocation that seems grossly unfair. We must consider an alternative ethical guideline.

Table 4.2. Maximizing total satisfaction, Case 2.

X	X	$U\ (X\ )$	$U\ (X\ ) = 2\ln X$	$U_\Sigma(X\ ,X\ ) = U\ (X\ ) + U\ (X\ )$
99	1	4.6	0	4.6
90.5	9.5	4.5	4.5	9.0
75	25	4.32	6.44	10.76
50	50	3.91	7.82	11.73
33	67	3.50	8.41	11.9
25	75	3.22	8.64	11.86
1	99	0	9.2	9.2

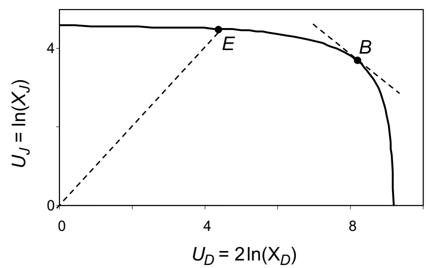


Fig. 4.4. Utility possibility frontier, Case 2
If $U\ (X\ ) = 2\ln(X\ )$, $U\ (X\ ) = \ln(X\ )$, and $X\ + X\ = 100$, the utility possibility curve stretches to the right in recognition of the alternative specification of Dick's utility function. Point E, where the 45^0 ray intersects the utility possibility frontier, yields equal utility by giving 9.5 units to Dick and 90.5 to Jane; specifically $U\ (9.5) = U\ (90.5) = 4.5$. Dick is much better off at Bentham's point B where the sum total of utility is maximized in accordance with the theory of utilitarianism, but Jane is at a tremendous disadvantage.

4.2.5 *John Rawls — The position of the least advantaged*

Harvard philosopher *John Rawls* has advanced the proposition that society should attempt to "maximize the position of the *least advantaged*."[7] Clearly, in the above example Jane is the least advantaged because of her lesser capacity to reap enjoyment. Rawls would have us redistribute the available supply of good X from Dick to Jane as long as she continues to be the least advantaged; i.e., as long as her utility is less than Dick's.

The argument is clarified by looking more closely at the utility possibility frontier calculated on Table 4.2 and plotted on Figure 4.4. Point E indicates the result of equal utility for all. Bentham would maximize total satisfaction at point B. We will find that point E is Rawls' preferred allo-

[7] *A Theory of Justice*, Harvard University Press, 1971.

cation in that it maximizes the position of the "Least Advantaged." To see why, note that all allocations to the right of the 45^0 equality ray emanating from the origin place Jane at a disadvantage (i.e., Dick's utility is higher than hers); all allocations to the left of that line disadvantage Dick (i.e., his utility is less than Jane's). At any point to the right of the 45^0 line we can improve the position of disadvantaged Jane by giving her some more goodies at the expense of Dick. At any point to the left of the 45^0 line we can improve the position of disadvantaged Dick by taking some goodies from Jane. Either way, we are driven by Rawls proposition to the point where the 45^0 line intersects the utility possibility curve. If we give 90.5 units to Jane, her utility will be 4.5 while Dick will enjoy 4.5 units of satisfaction from the 9.5 units of X that are left for him. Here we have unequal shares, approximately equal utility, and a total utility of $U_\Sigma = 9.0$, which is far below both the $U_\Sigma = 11.73$ that would be obtained when the available supply is shared equally or the maximum of 11.9 which Bentham advocates.

It might seem that Rawls principle of maximizing the position of the least advantaged implies that we should always strive to make everyone equally well off. After all, if one individual, say Dick, receives a higher level of utility than another, say Jane, then Jane is the least advantaged and redistributing goods from Dick to Jane can increase her utility. But this is the case only in the simplest possible situation in which there is a fixed quantity of the commodity to be allocated. Consider instead two hypothetical individuals, industrious citizen Hard Working and unemployable Down and Out. Suppose that taxing Hard Working to help Down and Out leads to a cutback in work effort and a resulting reduction in total output. Then the utility possibility frontier might conceivably take on the shape displayed on Figure 4.5. This would be the case if the 50% tax rate required to achieve equal consumption is beyond the peak of the Laffer curve plotted on Figure 3.16; i.e., $t_0 < 50\%$. Rawls would select point R on Figure 4.5 because it maximizes the position of least advantaged Down and Out even though utility is not equalized. Bentham would give Hard Working even more.[8]

Rawls attempted to muster support for his concept of fairness by appealing to a concept which he called the *Veil of Ignorance*. To illustrate this mental exercise, suppose that in advance of entering the world you are

[8]While at first blush it might appear that Rawls and Bentham would not worry about whether Down and Out was unemployable as a result of a war injury or because he had ingested control substances in his youth, this is not the case if the allocation rule would affect the behavior of future generations.

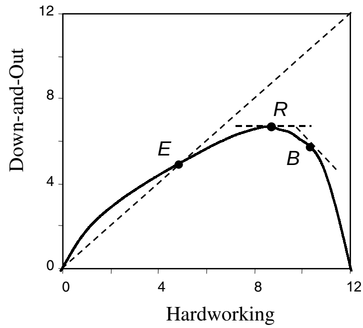


Fig. 4.5. Utility possibility frontier, Case 3
Because a tax imposed on Hardworking to finance a welfare payment to Down-and-Out leads to a reduction in work effort, total output falls. If the bigger the tax the larger the reduction in output, both citizens may be worse off when a 50% tax is imposed in order to achieve equal shares of the total output — compare equal-utility point E on the graph with R and B.

capable of making rational choices but do not know who you are going to be with certainty; e.g., you may be Dick or you may be Jane. Suppose that you have a 50% probability of being born with Jane's position in life and a 50% probability of being born with Dick's; then, according to Rawls, you would allocate goods between the two according to his principle of maximizing the position of the least advantaged — it would minimize the regret you would feel if the coin came up the wrong way.

In presenting his veil of ignorance argument, philosopher Rawls acknowledge that he was drawing on an argument presented by economist Nobel Laureate William Vickrey [1914–1996], who had suggested that in deciding how income should be distributed you might consider an immigrant trying to choose between two countries when both countries have the same average income but one has much less inequality than the other. In contrast to Rawls, Vickrey concluded that in the face of total uncertainty about the position that would be attained in the new country, the immigrant would maximum expected utility by going to the country that yields the highest sum total of happiness.[9] While the mental exercises

[9]William Vickrey, "Utility, Strategy and Social Decision Rules," *Quarterly Journal of Economics*, 1960, pp. 507–535. The differing conclusions arise because Vickrey assumed that in making decisions under uncertainty individuals maximize "expected utility," while Rawls thought that one should select the alternative that would minimize regret if it should turn out that one had made the wrong decision. For the Dick and Jane example, the expected utility is $E(U) = 1/2U \;\; + 1/2U \;\; = 1/2U_{\Sigma}$.

that Rawls and Vickrey proposed for judging the fundamental ethical issue of how income ought to be distributed are quite similar, the fact that they reached quite different conclusions about the appropriate distribution of income is most disconcerting, to say the least. No wonder questions of how income and wealth should be distributed are so heatedly debated.[10]

The seemingly simple problem of fairly allocating a single good between two individuals has turned out to be excruciatingly difficult. The problem was complicated even though we assumed that we knew precisely the utility functions revealing how much satisfaction each individual would derive from any given quantity of the scarce commodity. That is a most unsatisfactory assumption. We certainly cannot simply ask the individuals themselves to inform us on the basis of introspection because they would have every incentive to lie (although the best lie to tell would depend upon whether the allocator was a follower of Bentham or of Rawls). Is it possible to determine the utility functions from observed behavior? In the next section we will find why it may be impossible to learn enough about the nature of utility functions from studying the behavior of individuals in the market place or even by conducting experiments in an economics laboratory.

4.3 Two goods and one consumer

Now we shall consider a quite different problem:

> How will a consumer allocate a fixed income between two goods, assuming that actions are guided by the desire to maximize personal satisfaction?

In focusing on a single consumer we will be leaving, for the time being, the complex issues of equity; we will not be considering value judgments about how the world ought to be. Instead we shall focus, but only briefly, on the behavior of an individual interested in maximizing satisfaction or "utility."

4.3.1 *Consumer graphics — Indifference curves*

As a first step toward understanding the implications of the assumption that consumers maximize utility we must invest time in developing a rather cumbersome but very useful analytical apparatus. Let us begin with the

[10]Chapter 7, which is devoted to the topic of income distribution, presents data on income inequality in different countries, examines factors generating inequality, and considers alternative government policy for influencing the distribution of income.

assumption that the amount of satisfaction or "utility" that a consumer obtains from the consumption of specific quantities of two goods, X and Y, can be summarized by a utility function involving two variables: $U(X, Y)$. For an example, suppose that the utility function is:

$$U(X,Y) = X^{\frac{3}{4}} Y^{\frac{1}{4}}, \quad X \geq 0, \quad Y \geq 0. \tag{24}$$

Figure 4.6 plots in two dimensions certain properties of this function. The graph, called an *indifference map*, is similar to the weather map most daily newspapers publish in order to show how temperatures vary over the country. While the coordinates of each point on the weather map correspond to the longitude and latitude of a particularly location, on the indifference map each point corresponds to particular quantities consumed of the two goods. The contour curves plotted by the meteorologist on the weather map, called "isotherm lines," show the set of all points (locations) corresponding to a particular temperature; thus Miami may be on the 75 degree isotherm line while on the same day Minneapolis is on the -10 degree isotherm line. Similarly, an *indifference curve* plotted on an indifference map shows the combinations of the two goods yielding the same level of satisfaction to our consumer. For example, points $(6, 6)$ and $(8, 2.5)$ are both on the indifference curve labeled U=6 on our indifference map because both yield 6 units of satisfaction; i.e., $U(6, 6) = U(8, 2.5) = 6$.

Is equation (24) a reasonable functional form for a utility function? Is Figure 4.6 a reasonable picture summarizing an individual's preferences for varying combinations of two goods, X and Y? It seems reasonable to

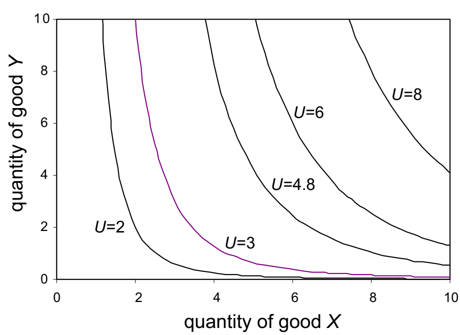


Fig. 4.6. An Indifference Map
Each indifference curve on the map shows the loci of all points yielding the indicated level of utility. The plotted indifference curves are for utility function $U = X^{3\ \ 4}Y^{1\ \ 4}$.

suppose that *more is better*, so that a larger quantity of one good with no reduction in the other should yield a higher degree of satisfaction or utility. The graph displays this property, for utility does indeed increase as we move up or to the right. Differentiation of the utility function confirms the impression conveyed by the graph: The marginal utility of X is $\partial U/\partial X = (3/4)X^{-1/4}Y^{1/4} = (3/4)(Y/X)^{1/4} = (3/4)U(X,Y)/X > 0$; similarly the marginal utility of Y is $\partial U/\partial Y = (1/4)(X/Y)^{3/4} = (1/4)U(X,Y)/Y > 0$.

The assumption that *more is better* has two implications for the shape of indifference curves. Indeed, we can conclude immediately that neither Figure 4.7 nor Figure 4.8 constitutes a legitimate indifference map.

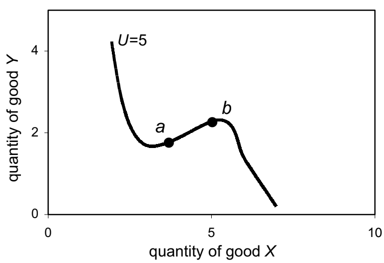


Fig. 4.7. Illegitimate indifference curves, Case 1
Indifference curve $U = 5$ has a positive slope at point a, which implies that there exits a point b involving more of both goods than point a but yielding the same level of satisfaction. This contradicts the assumption that more is better. Therefore, the assumption that more is better implies that the slope of indifference curves must be negative.

Fig. 4.8. Illegitimate indifference curves, Case 2
Suppose two indifference curves cross at point a. Now $U(b) = U(a)$ because they are on the same indifference curve; similarly, $U(c) = U(a)$ because they are on the same indifference curve. Therefore, $U(c) = U(b)$. But this is impossible because c involves more of both goods than b, which means that it must yield greater satisfaction. Therefore, the assumption that more is better implies that indifference curves do not cross.

1. The indifference curve on Figure 4.7 has a positive slope! To see why this is incompatible with the assumption that more is better, consider any two distinct points on the positively sloped region of that indifference curve, call them point a and point b. Since they are both on the same indifference curve they must both yield the same level of utility. But one, say b, involves more of both goods, which must yield more utility than the other point if more is indeed better. This argument demonstrates that the assumption that more is better implies that indifference curves (unlike the meteorologist's isotherm lines) must always have negative slopes.
2. Two indifference curves on Figure 4.8 intersect! A proof by contradiction (indirect proof) will establish that this is incompatible with the assumption that more is better. Consider points a, b, and c on the graph. Since b is on the same indifference curve as a, $U(b) = U(a)$; similarly, since point c is on the same indifference curve as a, $U(c) = U(a)$; therefore, $U(c) = U(b)$ (transitivity). But this is impossible because c involves more of both goods than b. Thus the assumption that more is better implies that indifference curves cannot intersect.

Now we are prepared to begin working out certain behavioral implications of the assumption that consumers are utility maximizers. Our consumer wishes to maximize $U(X, Y)$, but the consumption of the two goods is limited by income. If our consumer has income M to spend each week on the two commodities costing p_x and p_y, then the amount that can be consumed of each good is subject to the budget constraint:

$$p_x X + p_y Y \leq M\,.^{11} \tag{25}$$

For example, suppose our consumer has \$100 to spend, that the price of X is $p_x = \$10$, and that the price of the other good is $p_y = \$20$; then the individual will strive to maximize utility function (24) subject to the budget constraint

$$10X + 20Y \leq 100\,. \tag{26}$$

Since we obviously require $X \geq 0$ and $Y \geq 0$, the consumer's feasible choices are limited by the budget constraint to the triangular region with

[11] We are assuming for the time being that the consumer can neither borrow nor lend. Such complications are considered in Section 4.4.5.

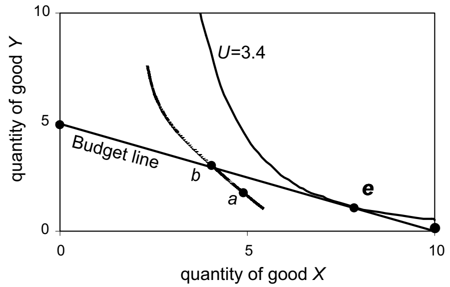


Fig. 4.9. Consumer equilibrium
Our consumer endowed with income of $M = 100$ can buy any combination of goods X and Y that lie on or below the budget line, given that p = 10 and p = 20.

The utility maximizing point cannot be below the budget line, like a, because then there would be other feasible points involving more of both goods. Points on an indifference curve that crosses the budget line, like b, cannot yield maximum utility because then there must be a point on that indifference curve that is below the budget line, like a. Therefore, a consumer can maximize satisfaction subject to a budget constraint only by purchasing at a point where an indifference curve is tangent to the budget line, like e.

vertexes at points $\langle 0,0\rangle$, $\langle 10,0\rangle$ and $\langle 0,5\rangle$ on Figure 4.9. More than this, the consumer will never consume at a point inside the triangle, such as point a, because more is better; the consumer who fails to spend all income is not maximizing satisfaction.[12] Thus we can conclude that a utility maximizing consumer with the specified income and prices will end up somewhere on the budget line defined by the equation

$$10X + 20Y = 100\,, \tag{27}$$

with $X \geq 0$ and $Y \geq 0$. Or more generally for prices p_x, p_y, and income M,

$$p_x X + p_y Y = M\,, \quad X \geq 0 \text{ and } Y \geq 0\,. \tag{28}$$

Therefore, the equation for the budget line is, in general,

$$Y = \frac{M}{p_y} - \frac{p_x X}{p_y}\,, \tag{29}$$

with slope $\partial Y/\partial X = -p_x/p_y$.

[12]This conclusion would have to be modified if there were a possibility of saving for the future, a complication that we will consider later in this chapter.

We can also conclude that the consumer will not consume at a point on the budget line like b, which is on an indifference curve that crosses the budget line. Since that indifference curve crosses the budget line it must have a point which is below the budget line, such as a, with $U(a) = U(b)$. Since a is below the budget line, there must be many other feasible consumption possibilities involving more of both goods than a, which means that they yield more utility than a or b. Thus b does not yield maximum utility.

All this implies a basic proposition:

> A utility maximizer will select a consumption point where the indifference curve is tangent to the budget line, such as e on the graph. Point e will be said to represent *consumer equilibrium* because there is no incentive for the consumer to move away from this utility maximizing point.

This proposition is subject to two qualifications:

- The consumer maximizes utility at point a on Figure 4.10 where the budget line is flatter than the indifference curve. This is a "corner tangency" where the best attainable consumption bundle involves only one of the two goods, just X, being consumed.
- Figure 4.11 shows two points of tangency. But point b is clearly inferior to a — the second order condition is violated at b. Tangency is only a *necessary* and not a *sufficient* condition for utility maximization.

Fig. 4.10. Corner tangency
A maximizing consumer may select a point where the slope of the budget line differs from the slope of the indifference curve if only one of the two commodities is purchased, as at point a. Point a is called a "corner tangency."

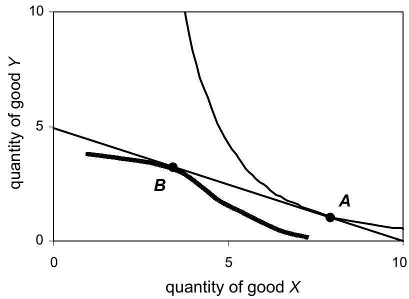


Fig. 4.11. Second order condition
Because our consumer can enjoy more utility at point A than at point B, our consumer would never select point B even though the budget line is tangent to the indifference curve at B. The second order condition for a maximum is violated at B. Tangency is not a sufficient condition for a maximum.

The top panel of Figure 4.12 shows how the utility maximizing consumer will respond if the price of X rises to $p_x = \$20$. The shift in the budget line reveals the new set of options now that the price of good X has increased. The graph indicates that this utility maximizing consumer will move to point e^*, enjoying less satisfaction than before. The bottom panel summarizes the evidence on how the change in the p_x affects the demand for X, given income and the price of other goods. The two points on the bottom panel, point d^* with coordinates $\langle 7.5, \$10\rangle$ and point d at $\langle 3.75, \$20\rangle$, must be on the demand curve for they are derived from the indifference map. The curve on the graph, based on only these two points, is a reasonable guess about the shape of the true demand curve. While additional points on the demand curve could be determined graphically by plotting more budget lines, we shall find that the equation for the demand curve can be derived analytically with the aid of the calculus.

4.3.2 *Analytical solution*

The consumer's problem is to maximize a utility function $U(X, Y)$, subject to budget constraint (28). In illustrating the procedure we shall sacrifice generality by working with the particularly simple functional form of the utility function, that of equation (24).

As a first step in solving this problem of maximizing a function of two variables subject to a constraint, let us transform it by substitution into an equivalent problem of maximizing an unconstrained function in but one

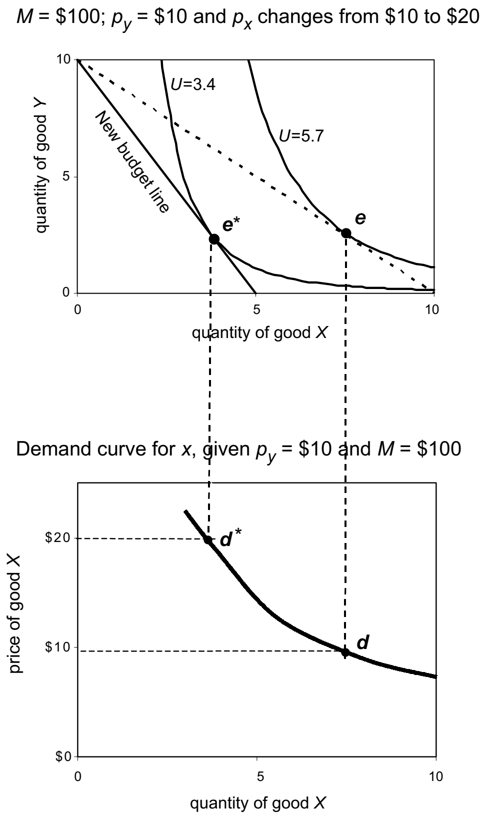


Fig. 4.12. Derivation of the demand curve
Initially our consumer is in equilibrium at point e on the top panel, given income of \$100 and $p \;\; = p \;\; = \$10$. The initial budget constraint, represented by the dotted line, is $100 - 10X - 10Y = 0$. Tangency point e indicates the quantities purchased by our utility maximizing consumer.

When the price of X increases to $p \;\; = \$20$ while the price of the other good and income remain unchanged, the budget constraint shifts to $100 - 20X - 10Y$. Thus the budget line rotates through point $\langle 0, 10 \rangle$ because our consumer could still buy ten units of Y with no X. The new budget line is steeper, as indicated on the graph, and the new equilibrium is at tangency point e^*.

The two equilibrium points e and e^* generated by this process yield points d and d^* on the demand curve that is sketched on the bottom panel.

unknown variable.[13] The initial step toward achieving this transformation is to note that income-constraint equation (28) yields

$$Y = \frac{M}{p_y} - \left(\frac{p_x}{p_y}\right) X\,. \tag{30}$$

This equation can be used to eliminate Y from the utility function $U(X, Y)$ in order to obtain a new function explaining utility that involves only one unknown and no constraints. With our specific functional form of the utility function, $U(X, Y) = X^{3/4}Y^{1/4}$, as specified by equation (24), we have on substituting from (30)

$$U^*(X) = X^{3/4}\left[\frac{M}{p_y} - \left(\frac{p_x}{p_y}\right) X\right]^{1/4}. \tag{31}$$

This function reveals how an increase in X will affect utility when the resulting reduction in the consumption of good Y imposed by the budget constraint is implicitly taken into account.

To maximize utility we must set the first derivative of this new function with respect to X equal to zero:[14]

$$\begin{aligned}\frac{dU^*(X)}{dX} &= \left(\frac{3}{4}\right) X^{-1/4}\left[\frac{M}{p_y} - \left(\frac{p_x}{p_y}\right) X\right]^{1/4} \\ &\quad + X^{3/4}\left[\frac{M}{p_y} - \left(\frac{p_x}{p_y}\right) X\right]^{-3/4}\frac{(-p_x/p_y)}{4} = 0\,. \end{aligned} \tag{32}$$

To solve for X, we multiply both sides of this equation by the artfully selected constant $4X^{(1/4)}[M/p_y - (p_x/p_y)X]^{(3/4)}$:

---

[13] Alternatively, we could solve this constrained maximization by the Lagrangian Multiplier technique discussed in the Appendix.

[14] The basic strategy for finding the maximum relies on the chain rule: In its simplest form, the chain rule says that given the functions $z(y)$ and $y(x)$, the composite function $z[y(x)]$ has derivative $dz/dx = dz/dy\, dy/dx$, where dz/dy is evaluated at $y(x)$.

A slight elaboration is required for functions of two or more variables: Suppose $z(x, y)$ and $y(x)$. Then $\partial z/\partial x$ is the change in z keeping y constant while the total derivative $dz/dx = \partial z/\partial x + \partial z/\partial y\ dy/dx$ takes into account both the direct effect of the change in x on z and the indirect effect of the change in x on y and then of that change in y on z.

For the utility example we apply this general rule by setting $z(x, y) = U(x, y)$. Then $\partial z/\partial x = \partial U/\partial x$ (the marginal utility of x), $\partial z/\partial y = \partial U/\partial y$ (the marginal utility of y), and $dy/dx = -p\ \ /p\ \ $. Thus we have $dU/dx = \partial U/\partial x + \partial U/\partial y(-p\ \ /p\ \ )$.

$$4X^{1/4}\left[\frac{M}{p_y}-\left(\frac{p_x}{p_y}\right)X\right]^{3/4}\frac{dU^*(X)}{dX}=3\left[\frac{M}{p_y}-\left(\frac{p_x}{p_y}\right)X\right]-\left(\frac{p_x}{p_y}\right)X=0\,. \tag{33}$$

Hence,

$$\frac{3M}{p_y}=4X\left(\frac{p_x}{p_y}\right), \tag{34}$$

or finally:

$$X=\frac{\left(\frac{3}{4}\right)M}{p_x}\,. \tag{35}$$

That is to say, the demand function for commodity X of a consumer maximizing utility function (24) is

$$X(p_x,p_y,M)=\frac{3M}{4p_x}\,. \tag{36}$$

Two examples:

1. If $M=\$12$ and $p_x=\$2$, then $X(\$2,p_y,\$12)=4.5$.
2. If p_x were to fall to \$1, everything else unchanged, then $X(\$1,p_y,\$12)=9$.

Equation (36) generated the demand curve plotted on the bottom panel of Figure 4.12.

Note that for this particular utility function we have the unexpected result that consumption of X is *not* influenced by the price p_y of the other good. Further, examination of demand equation (35) shows that total expenditure on good X, p_xX, always equals 3/4ths of income. This *constant income share* for good X is dictated by the form of utility function (24). Alternative functional forms for the utility function yield a rich variety of demand curves with quite different properties.

The demand for the other good, Y, which can be found either by symmetry or by substituting the expression for X into the budget constraint (29), is:

$$Y(p_x,p_y,M)=\frac{M}{4p_y}\,. \tag{37}$$

Indirect Utility Function:

Now that we have the demand functions for X and Y we can derive the indirect utility function showing utility as a function of income and prices.

To obtain this relationship we substitute the demand for each good as a function of prices and income into the consumer's utility function:

$$
\begin{aligned}
U^I(M, p_x, p_y) &= \left[\frac{\left(\frac{3}{4}\right)M}{p_x}\right]^{3/4} \left(\frac{\frac{1}{4}M}{p_y}\right)^{1/4} \\
&= \left(\frac{3}{p_x}\right)^{3/4} \left(\frac{1}{p_y}\right)^{1/4} \frac{M}{4}\,. \qquad (38)
\end{aligned}
$$

Example: If $M = \$12$, $p_X = \$2$ and $p_y = \$1$, then $U^I(\$12, \$2, \$1) = 4.07$.

This indirect utility function shows the highest attainable level of utility as a function of income and prices. It shows the best that the consumer can do, given the specified income and prices. It shows our consumer's utility, assuming that the consumer has indeed succeeded in maximizing the level of satisfaction attainable from the given income. The indirect utility function might be used to appraise how changes in prices and money income affect our consumer's well being.

Expenditure Function:

It is useful to derive from (38) the expenditure function showing the income required to obtain a specified level of utility:

$$
M(U, p_x, p_y) = 4\left(\frac{3}{p_x}\right)^{-3/4} \left(\frac{1}{p_y}\right)^{-1/4} U\,. \qquad (39)
$$

For example, when inflation pushes up the price of consumer goods, a benevolent employer could use the expenditure function to calculate the adjustment of wages that would allow the workers to preserve their level of real income (i.e., utility). It provides for the maintenance of purchasing power.

Compensated Demand Function:

How much X will a consumer purchase who is so fortunate as to have his income adjusted so that he can enjoy the same utility as before the price changes? This question is answered by the compensated demand function. If we substitute the expenditure function $M(U, p_x, p_y)$ into demand equation $X(p_x, p_y, M)$ we obtain a new function $X^c(p_x, p_y, U)$ showing how the demand for X responds to changes in its price *if* the consumer is *compensated* so as to be able to attain precisely the same level of utility as before.

For example, if we again have $U(X,Y) = X^{3/4}Y^{1/4}$, as specified by equation (24), we obtain by substituting expenditure function (39) into demand equation (35)

$$\begin{aligned} X^c(p_x, p_y, U) &= 3\left(\frac{3}{p_x}\right)^{-3/4} \left(\frac{1}{p_y}\right)^{-1/4} \frac{U}{p_x} \\ &= \left(\frac{3p_y}{p_x}\right)^{1/4} U \,. \end{aligned} \tag{40}$$

To distinguish the two types of demand functions, we write $X^c(p_x, p_y, U)$ when referring to the compensated demand function.

4.3.3 *Application — Compensating for inflation*

Consider a consumer with utility function (24) who enjoyed an income of $M = \$12$ when the prices were $p_x = \$2$, $p_y = \$1$. Then from (35), $X(\$2, \$1, \$12) = 4.5$; from (37), $Y(\$1, \$1, \$12) = 3$; and by (38), $U^I(\$12, \$2, \$1) = 4.07$. Suppose that the price of X doubles from \$2 to \$4 but the price of Y remains unchanged at \$1. By how much would our consumer's money income have to increase in order to be able to enjoy the same level of utility as before?

The naïve answer is to note that the same quantities of the two goods ($X = 4.5$ and $Y = 3$) could be purchased only if income increased to $\$4 \times 4.5 + \$1 \times 3 = \$21$, or an increase of 75%. But substitution of \$21 into the indirect utility function (38) reveals that $U^I(\$21, \$4, \$1) = 4.23$, substantially above the $U(\$12, \$2, \$1) = 4.07$ enjoyed before the inflation. The gain in utility is achieved by consuming $X = 3.78$ and $Y = 5.05$ — our utility maximizing consumer has substituted away from X because it is now more expensive relative to Y. The point of this example is that increasing a consumer's income to the point where the same quantities of the two goods can be purchased as before the price hike *overcompensates* for inflation!

Expenditure function (39) reveals the income that will just suffice to yield the same level of utility is $M(4.07, \$4, \$1) = \$20.18$, or a 68% raise rather than 75% raise above the \$12 income received before the price hike. To verify, substituting \$20.18 into the two demand functions yields $X(\$4, \$20.18) = 3.78$ and $Y(\$1, \$20.18) = 5.05$. Hence $U = 4.07$, precisely the level that prevailed before the inflation.

The naïve computations had suggested that a 75% increase in money income was required to compensate for inflation when in fact an increase from \$12 to \$20.18 or 68% would suffice. What was wrong with the naïve calculation? The naïve procedure was based on the erroneous assumption that the consumer would have to be able to purchase precisely the same quantities of the two goods as before the inflation in order to enjoy the same level of utility. This is more than enough because the increase in the relative price of X induces the consumer to substitute away from that commodity and consume more of the commodity that had not suffered a price increase.

Although this numerical example relied on a particularly simple functional form for the utility function, it illustrates a quite general proposition. Workers are overcompensated for inflation if their income increases by enough to purchase the same quantities of each good as they purchased before the inflation, unless the prices of all goods rise by the same percentage. The same proposition applies to retirees. When the time comes in Chapter 8 to examine how the Consumer Price Index is constructed we will find that the Social Security System used to overcompensate retirees in precisely this way.

4.3.4 *General statement of the two commodity problem*

Equation (24) is a very special form of the utility function. More generally, we might suppose that the consumer wishes to maximize a more general form of the utility function $U(X, Y)$ subject to budget constraint (25). It shall be supposed that the utility function is continuous and differentiable and, in addition, that more is better; i.e., $\partial U/\partial X > 0$ and $\partial U/\partial Y > 0$. The first of these partial derivatives, $\partial U/\partial X$ — the derivative of U with respect to X, given the magnitude of Y — is called the *marginal utility* of X; it shows how the total utility that our consumer enjoys will respond to an increase in X, given that the quantity of Y made available to the consumer remains fixed. Similarly, $\partial U/\partial Y$ is the marginal utility of Y.[15]

Now our consumer's problem is to maximize $U(X, Y)$ subject to the budget constraint, equation (28). By substituting the budget constraint into the utility function we simplify the consumer's problem of maximizing

[15]For utility function (24), the marginal utility of X was $\partial U/\partial X = (3/4)X^{-1\ 4}Y^{1\ 4} = (3/4)(Y/X)^{1\ 4}$. For this particular utility function, but not generally, marginal utility depends only on the ratio X/Y in which the two commodities are consumed.

the utility function in two variables subject to the budget constraint to the simpler task of maximizing an unconstrained function in but one unknown

$$U^*(X) = U\left(X, \frac{M}{p_y} - \frac{p_x X}{p_y}\right). \tag{41}$$

Note that this equation captures both the direct effect of a change in X on utility and the indirect effect on utility of the reduction in Y that necessary results when X increases, thanks to the budget constraint. Let us take both of these effects into account by totally differentiating equation (41) with respect to the unknown X, which yields as a necessary condition for a maximum

$$\frac{dU^*}{dX} = \frac{\partial U}{\partial X} - \frac{\partial U}{\partial Y}\left(\frac{p_x}{p_y}\right) = 0\,; \tag{42}$$

i.e.,

$$\frac{\partial U/\partial X}{\partial U/\partial Y} = \frac{p_x}{p_y}\,. \tag{43}$$

That is to say, a utility maximizing consumer will adjust consumption so as to equate the ratio of the marginal utility of X divided by the marginal utility of Y to the ratio of the price of X to the price of Y.

Now the slope of an indifference curve, such as that passing through point e on the top panel of Figure 4.12, is intimately related to the concept of marginal utility. To visualize the link, first consider how utility would change as a result of infinitesimal changes of dX and dY in the quantity of the two goods consumed:

$$dU = \frac{\partial U}{\partial X}dX + \frac{\partial U}{\partial Y}dY\,. \tag{44}$$

But on any indifference curve $dU = 0$; otherwise, utility would change; therefore, if we are to stay on the same indifference curve the changes of dX and dY cannot be taken at random but must satisfy:

$$dU = \frac{\partial U}{\partial X}dX + \frac{\partial U}{\partial Y}dY = 0\,. \tag{45}$$

Solving, we find that

$$\left.\frac{-dY}{dX}\right|_U = \frac{\partial U/\partial X}{\partial U/\partial Y} = \frac{p_x}{p_y}\,, \tag{46}$$

where the second equality follows from equation (43). Now we call the fundamental concept $-dY/dX\big|_U$, which is minus the slope of the indifference curve, the *marginal rate of substitution*. Using this terminology, the first equality of equation (46) is telling us that the marginal rate of substitution is equal to the ratio of the marginal utilities of the two commodities. The equality of the marginal rate of substitution with the ratio p_x/p_y is equivalent to the tangency property at equilibrium point e on Figure 4.9.

4.3.5 *Diminishing marginal utility and the law of demand*

Does the *Law of Demand* imply *diminishing marginal utility*? That is to say, if a demand function slopes downward does that imply that marginal utility declines if more of the commodity is consumed? To see that this proposition is not true in general, consider the utility function

$$U^+ = X^3 Y\,. \tag{47}$$

Now the marginal utility of Y is $\partial U^+/\partial Y = X^3$, which clearly does not diminish when Y increases, in contradiction to the principle of diminishing marginal utility. Worse, $\partial U^+/\partial X = 3X^2Y$, which means that the marginal utility of X increases when more of that good is consumed; more than this, $\partial U^{+2}/\partial^2 X = 6XY > 0$. However, it is easy to show that a utility maximizing consumer with this utility function will have demand function $(3/4)M/p_x$, which is identical to that obtained with utility function $U(X,Y) = X^{3/4}Y^{1/4}$. This is disconcerting for two reasons:

1. It means that diminishing marginal utility is not necessary in order to have a demand function satisfy the Law of Demand.[16]
2. It also means that there is no easy way to measure utility empirically. Even if empirical research were to yield an accurate estimate of a consumer's demand function showing how the quantity purchased responds to changes in prices and income, we would still not be able to determine from such evidence whether or not the consumer's utility function satisfied the law of diminishing marginal utility.

Example: Even if we knew that a consumer had demand function $X = (3/4)M/p_x$, we could not tell whether the utility function was $U^+(X,Y) = X^3Y$ or $U(X,Y) = X^{3/4}Y^{1/4}$.

[16]Later we will demonstrate that a rational consumer may consume more of a commodity when its price increases. And this may happen even when there is diminishing marginal utility!

Moral: We may well despair at ever obtaining by objective methods an estimate of marginal utility, but without that type of information it would not be possible for even Jeremy Bentham's disinterested spectator to make the type of welfare judgments required to put utilitarianism into practice! The immeasurability of utility creates a fundamental problem for the doctrine of utilitarianism.

4.3.6 *Inflation, consumption, and homogeneity*

Suppose that in a time of inflation the prices of all goods, including wages, double. How will a utility maximizing consumer respond? It may not be intuitively obvious as first glance, but the assumption that consumers are utility maximizers implies that there will be no change in consumption if all prices and income change by the same proportion. To see why, recall that in the discussion of the equation for the budget line we considered a consumer who has $M = \$100$ to spend during a week when the price of X was $p_x = \$10$, and the price of the other good is $p_y = \$20$. This consumer strives to maximize the utility function subject to the budget constraint

$$10X + 20Y = 100\,. \tag{48}$$

This is the equation for the budget line plotted on Figure 4.9. Now when prices and income all double, $M = 200$, $p_x = 20$ and $p_y = 40$ and the budget constraint is

$$20X + 40Y = 200\,; \tag{49}$$

but this is precisely equivalent to the pre-inflation constraint. Graphically, we still have exactly the same budget line as before the inflation, which means that the consumer's equilibrium is unchanged. Of course, it may be human nature for everyone to complain about the rise in the prices of the goods they buy without considering the extent to which the inflation has been offset by rising money income. It is only human nature for our consumer to believe that the money wage increase was earned and indeed long overdue and object, perhaps vehemently, that the inflation is wiping out the wage increase. But the fact is that our aggrieved consumer is enjoying exactly the same utility as before inflation pushed up both prices and wages proportionately.

The proposition that a doubling of incomes and prices will lead to no change in consumption must hold quite generally, regardless of the precise form of the demand function, provided only that it is obtained by maximizing utility subject to the budget constraint. So in general we can state

$$q(M, p_x, p_y) = q(2M, 2p_x, 2p_y)\,. \tag{50}$$

More than this, the proposition must hold if prices and income go up three fold rather than just doubling. Indeed, for any positive number ρ we must have:

$$q(M, p_x, p_y) = q(\rho M, \rho p_x, \rho p_y)\,, \tag{51}$$

and we say that the demand function is *homogeneous of degree zero* in income and prices.[17] In particular, if we set $\rho = 1/p_y$, we have

$$q(M, p_x, p_y) = q\left(\frac{M}{p_y}, \frac{p_x}{p_y}, 1\right), \tag{52}$$

That is to say, it is only *relative prices* p_x/p_y (the price of X relative to the price of Y) and income adjusted for inflation (M/p_y) that matter in determining demand for a commodity purchased by a utility maximizing consumer.[18]

Example: The demand function $X(M, p_x, p_y) = 4 + 1/4M - 1/2p_x - 3/4p_y$ appears to be reasonable in that consumption is positively related to income but falls if the price of the good increases. But since it is *not* homogeneous of degree zero in income and prices, it can *not* be the demand function of a utility maximizer!

To verify that this demand function is indeed not homogeneous of degree zero it is only necessary to produce a single counterexample. Here is such a counterexample:

If $M = \$4$, $p_x = \$1$ and $p_y = \$2$ we have

$$X(\$4, \$1, \$2) = 4 + 1 - \frac{1}{2} - \frac{3}{2} = 3.$$

---

[17] More generally, we say that a function $f(X, Y)$ is homogeneous of degree k if and only if for any X, Y and λ we have $\lambda\ f(X, Y) = f(\lambda X, \lambda Y)$.

[18] In Chapter 2 we found that the supply of a commodity depended only on relative prices.

But doubling income and prices (i.e., $\rho = 2$), we have

$$M = \$8,\ p_x = \$2 \text{ and } p_y = \$4 \text{ and } X(\$8, \$2, \$4) = 4 + 2 - 1 - 3 = 2 \neq 3.$$

Because the function is not homogeneous of degree zero it is impossible to construct a utility function that will yield this demand function when utility is maximized, subject to a budget constraint.

4.4 Indifference curve applications

The following propositions appear self-evident:

1. An increase in price must lead to a reduction in sales (i.e., demand curves slope downward in accordance with the Law of Demand).
2. A tax on wages will cause a reduction in hours worked.
3. An increase in the interest rate will generate increased saving.

While these propositions may seem self evident, indifference curve analysis will reveal that none of them is implied by the proposition that consumers are rational utility maximizers! To show that the assumption that consumers are utility maximizes does not imply any of these propositions it will only be necessary to present appropriate counter-examples. Each proposition will be considered in turn, but first it is necessary to explain the difference between the "income effect" and the "substitution effect."

4.4.1 *Income effect versus the substitution effect*

The indifference map constituting Figure 4.12 showed how the demand for commodity X by a utility maximizing consumer with utility function $U(X, Y) = X^{3/4}Y^{1/4}$ responds to a change in its price, p_x, given the price of commodity Y and income M. Initially, with $M = 100$ and $p_x = p_y = 10$, we found that the consumer is at point e on indifference curve yielding utility $U = 5.7$. When p_x increases to 20 the best the consumer can do is to shift to point e^* on indifference curve $U = 3.4$. This adjustment involves a reduction in the consumption of good X from 7.5 to 3.75. It proves useful to break this reduction of $7.5 - 3.75 = 3.75$ down into two steps which can be explained with the aid of Figure 4.13. The first step is the movement form e to e' and the second step is from e' to e^*.

Step 1, The substitution effect: The movement from e to e', may be interpreted as the adjustment in consumption that would take place if, simultaneously with the price change, money income were adjusted so as

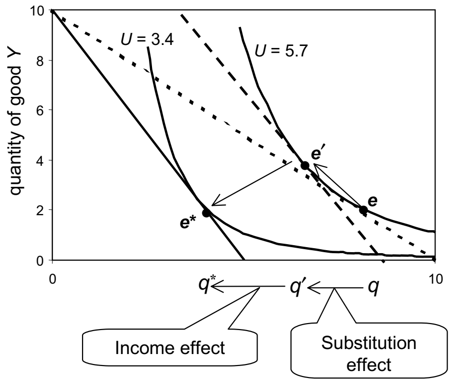


Fig. 4.13. Income vs. substitution effect
The process by which the consumer adjusts to a change in the price of X, given income, can be broken down into two artificial steps:
Step #1: The "substitution effect" involves sliding along the initial indifference curve from point e to point e' in response to the increase in the relative price of X. Due to the substitution effect of the price increase, consumption of good X falls from q to q'.
Step #2: The "income effect" is the movement from point e' to point e^*. The income effect shows the effect of the change in real income occasioned by the price increase. It results in a fall in the quantity of X consumed from q' to q^*.

to leave our utility maximizing consumer exactly as well off as before — precisely the same level of utility would be enjoyed at point $\mathbf{e}'$ as at $\mathbf{e}$. The required increase in income is specified by equation (39). Now that good X is more expensive than before, our consumer will reduce consumption of good X from 7.5 to 6, assuming that money income has been increased to \$170 in order to leave the utility maximizing consumer on the original indifference curve with $U = 5.7$. The precise change in the quantity of X consumed is given by the compensated demand function, (40). The consumer is said to "substitute away" from the good that has increased in price.

Step 2, The income effect: The movement from point e' to point e^*, shows how the consumption of goods X and Y change as a result of the change in utility that our consumer experiences as a result of the price change. Observe that it is equivalent to the effect of a decrease in income, given prices, in moving from point e' on indifference curve $U = 5.7$ to point e^* on indifference curve $U = 3.4$. Note that these two budget lines have the same slope; i.e., the marginal rate of substitution is the same at points e' and e^*. Thus there is no substitution effect in moving from e' to e^*.

To summarize, the increase in the price of X from $p_x = 10$ to $p_x = 20$ led via the *substitution effect* to a reduction in the consumption of good X of 0.75 units $(6.75 - 6.0)$. The *income effect* led to a further reduction of 2.5 units of X $(6.0 - 3.5)$. The combined result of these two effects of the price increase was a drop in demand of $3.25 = 0.75 + 2.5$.

The following facts are of fundamental importance:

- The substitution effect of a price increase is always negative; that is to say, an increase in the price of commodity X necessarily leads to a decrease in consumption of X *if* money income is adjusted so as to keep the consumer on the original indifference curve — the consumer slides to the left along the indifference curve until the slope of the indifference curve (i.e., the marginal rate of substitution), has increased by enough to equal the new price ratio.
- The income effect may be either positive or negative.

4.4.2 *Inferior goods*

Figure 4.14 shows the effect on the consumption of a commodity when there is an increase in income but no price change. Surprisingly, our consumer has reduced consumption of Good X; i.e. Good X is an *inferior good* because the increase in income leads to a reduction in consumption. Observe that the *income effect* on the consumption of an inferior good is negative. Inferior goods are less common than *normal goods* for which the income

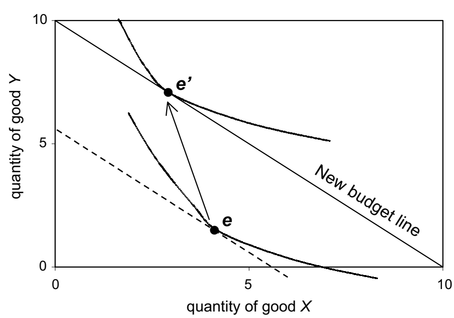


Fig. 4.14. Inferior good
This is the indifference map for an inferior good. When an increase in income leads to an outward shift in the budget line, the consumer shifts from point e to e' and the quantity consumed of good X falls.

effect is positive, but they are far from rare. For example, if a consumer shifts from eating potatoes to eating steak as a result of an increase in income, then potatoes are an inferior good.

4.4.3 *Giffen goods*

Commodities whose consumption increases when the price rises are called *Giffen goods* after Sir Robert Giffen [1837–1910], who is said to have observed the phenomenon at the time of the Irish potato famine. The indifference map on Figure 4.15 is artfully drawn so as to make the income effect of the price increase positive and stronger than the substitution effect. This means that a hike in the price of good X leads to an *increase* in its consumption, which is a violation of the Law of Demand! Yet the indifference curves cannot be ruled out of consideration for they satisfy the customary restrictions: they are negatively sloped and they do not cross. Here we have a legitimate counter-example to the Law of Demand!

Further reflection may suggest that this counter-intuitive result may not be quite as bizarre as it appeared on first blush:

Hypothetical example: A family has weekly income of \$14. When the price of potatoes for one meal was \$1 and the price of meat \$8, they consumed meat only on Sunday and potatoes six days a week. But when the price of

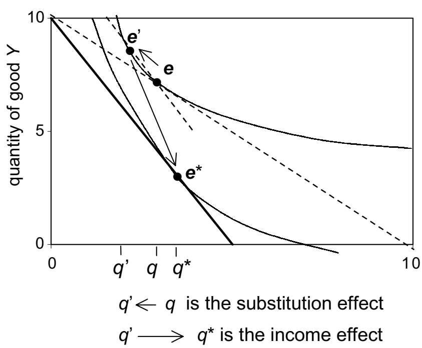


Fig. 4.15. A Giffen Good
The Giffen Good is an extreme type of inferior good. In the case of a "Giffen Good," the substitution effect of the price increase (the movement from e to e') is overwhelmed by a stronger negative income effect (the movement from e' to e^*). As a result the price increase leads to an increase in the consumption of X. Thus the demand curve for this commodity has a positive slope. The Law of Demand does *not* hold for this utility maximizing consumer. Compare with Figure 4.13.

potatoes increased to \$2 as a result of the potato blight, the family decided to eat potatoes seven days a week and do entirely without meat. This family's consumption of potatoes increased in the face of the price hike because the family found it was less painful to consume potatoes seven days a week than to continue to consume meat on Sunday, eat potatoes three days a week, and go hungry the other three days.

This counter-example to the Law of Demand is important for it means that the Law of Demand does not hold as a matter of logical necessity. Nevertheless, unless there is evidence to the contrary, economists usually assume that demand curves have negative slopes.

4.4.4 *Work-leisure tradeoff*

How a worker chooses between work and leisure is illustrated on Figure 4.16. Instead of choosing between two goods, X and Y, our worker chooses between hours of leisure (L) and income (Y). It will be assumed that all income is consumed. Hours of leisure, L, are plotted on the abscissa, moving from left to right; but since there are only 24 hours in the day, the number of hours worked is $H = 24 - L$, which is shown with the second scale below the abscissa. For example, working an 8 hour day, as at point e, leaves 16 hours of leisure.

Assuming the wage is $w = \$5$ an hour our worker could enjoy an income of $Y = \$40$ by working an 8 hour day or an income of zero by enjoying

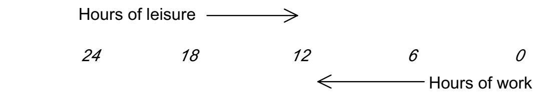


Fig. 4.16. Work-leisure choice
Hours of leisure enjoyed each day are plotted on the abscissa, but because there are exactly 24 hours in the day, the number of hours worked each day can be read from the *backwards scale* running from right to left. The budget line, drawn for a wage of \$5 per hour, indicates that income is zero if the worker decides to enjoy 24 hours of leisure. Our worker maximizes utility at point e by enjoying 16 hours of leisure; i.e., working an 8-hour day and earning an income of \$40.

24 hours of leisure. The full range of options are indicated by the budget line on the graph, which is based on the equation

$$Y = wH = w(24 - L)\,, \tag{53}$$

where Y is income and the hourly wage $w = \$5$. Since the indifference curve at point e is tangent to the budget line, our worker is maximizing her utility.

The next graph shows how the supply curve of labor can be derived from the work-leisure choice indifference map. It adds a second budget line in order to determine how our worker will respond to a reduction in the take-home wage to \$3.00 per hour when the governor imposes a 40% wage tax.

- The new budget line intersects the old at no-work point $(24, 0)$ because regardless of w, $Y = 0$ if $L = 24$; i.e., $H = 0$; but it is flatter because the effective wage rate is lower.
- Our utility maximizing worker has responded to the wage decrease by moving to point $(18, \$18)$, enjoying more leisure but suffering a substantial reduction in money income from \$40 to \$18.

Because of the reduction in hours of work, the tax base is eroded and the governor collects only \$12 (40% of $6 \times \$5$) rather than the anticipated \$16 (40% of $8 \times \$5.00$) in tax revenue! Observe that the resulting supply curve of Figure 4.18 has the expected positive slope.

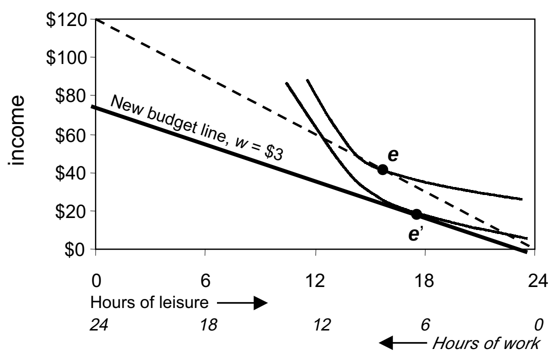


Fig. 4.17. The supply of labor responds to a wage change
Suppose a 40% wage tax is enacted, and as a result the worker's hourly after-tax wage drops from \$5 to \$3. The new budget line reflects the reduction in the after-tax wage, and our worker has chosen to cut back to a 6 hour work day at e'.

Fig. 4.18. Labor supply curve
Here is the labor supply curve generated by the indifference map on Figure 4.17. Points e and e' correspond to the two equilibrium points on Figure 4.17.

It seems intuitively obvious that a worker will reduce the supply of labor if the take-home wage is cut. Indeed, this assumption is a major premise often advanced by tax cut advocates, for it implies that a reduction in taxes will stimulate output. But Figure 4.19 shows that it is quite conceivable that a utility maximizing worker might put in *more* hours when after-tax wages fall. One might say that the effect of the wage cut in this example is to lead the worker to substitute leisure for income but that this effect is swamped by the income effect leading to a more than offsetting increase in hours worked. The result is a *negatively* sloped or *backward bending supply curve of labor*, which is hardly consistent with the conventional wisdom. This example suggests that an increase in taxes might lead to an increase in hours worked! Here is a quite plausible example explaining how such behavior might conceivably arise:

Example: A student works as a research assistant to earn \$1,000 during the year to help meet part of the financial costs of attending college. The faculty employer, pleased with the student's work, offers a raise in the hope that the student will put in more hours, but just the opposite happens. The student works less because fewer hours of work are required to meet the \$1,000 financial obligation.

It may be useful to consider a utility maximizing worker with utility function $U = L^{2/3}Y^{1/3}$. Our rational worker will strive to maximize $U = L^{2/3}Y^{1/3}$ subject to the constraint that $Y = w(24 - L)$. Substitution leads to the unconstrained problem of maximizing $U = L^{2/3}[w(24 - L)]^{1/3}$.

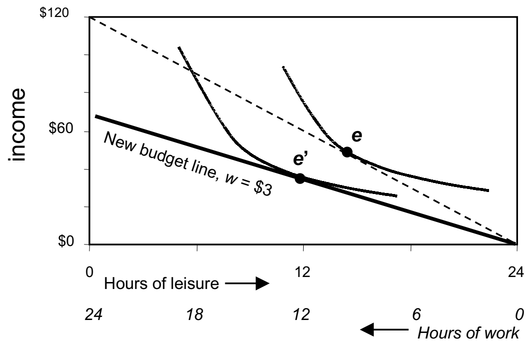


Fig. 4.19. Backward bending supply of labor
Here is a rational utility maximizing worker who decides to put in longer hours when a tax increase reduces her after-tax wage from \$5 to \$3 per hour, the income effect of the reduced wage dominating the substitution effect. By increasing hours worked, the drop in take home pay is limited to a reduction from \$40 to \$36.

Because the resulting supply function of labor for this rational consumer has a negative slope, it is said to be backward bending.

Differentiation yields

$$\frac{dU}{dL} = \frac{2}{3}L^{-\frac{1}{3}}[w(24-L)]^{\frac{1}{3}} - L^{\frac{2}{3}}[w(24-L)]^{-\frac{2}{3}}\frac{w}{3} = 0 \tag{54}$$

as a necessary condition for a maximum. Multiplying by $3L^{1/3}[w(24-L)]^{2/3}$ leads to

$$2w(24-L) - wL = 0\,,$$

but the wage rate w cancels out and we find that the worker puts in $8 = 24-L$ hours on the job and enjoys $L = 16$ hours of leisure out of 24, provided $w > 0$. This worker's utility function yields for any $w > 0$ the *8 hour work day*! Of course, this intriguing result holds only for this particular form of the utility function.

The assumption that workers maximize utility has suggested a broad range of possible outcomes. It is possible that a rational utility maximizing worker would work 8 hours regardless of the wage rate. But it is also conceivable that work effort might actually increase as a result of a reduction in hourly wages. The empirical evidence is still divided on the issue of how prevalent backward bending supply curves of labor actually are, some econometric studies yielding negatively sloped supply curves while others showing the opposite.

The next graph looks at the effect of welfare on work effort. The budget line is drawn under the assumption that the consumer on welfare can receive \$20 per day of support from the state, provided this is the only source of income. However, if the welfare recipient works, the welfare payment will be reduced by \$1.00 for every dollar earned. For example, if our worker works 3 hours at \$5 per hour she will earn \$15 and receive a welfare check of \$5, or a total of \$20. If she works 4 hours she will earn \$20 and receive no welfare payment. If she stops working she will receive a \$20 welfare check. Thus the budget constraint for our welfare recipient is

$$Y = \max(\$20, \$5 \cdot H) = \max(\$20, \$5 \cdot (24 - L)\,. \tag{55}$$

This budget line now has a *Welfare Step.* Clearly, with this welfare augmented budget line there is no payoff from working, other than possible self esteem, unless one puts in more than 4 hours per day (and then only if one neglects commuting, child care costs, other expenses, and the possible loss of welfare health benefits)! For the indifference curve drawn on the graph, the consumer would have worked 8 hours without welfare. If eligible for welfare, utility is maximized by not working at all. This form of welfare benefit amounts to a 100% tax on low income, which clearly discourages work effort. The welfare step keeps workers from stepping off welfare.

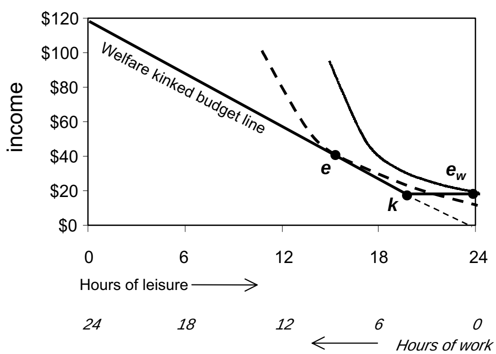


Fig. 4.20. Work or welfare?
In the absence of welfare, a worker earning \$5 per hour would face the budget line with coordinates $\langle 0, \$120\rangle$ and $\langle 24, 0\rangle$, just as before. This worker maximizes utility by working an eight hour day at point e, earning \$40.

When a welfare program is established providing \$20 a day, less \$1 for every \$1 the welfare recipient earns, the budget line kinks at point k to form a welfare step. Once this welfare program is established, this worker decides to maximize utility at point e by not working. This welfare program has the perverse effect of discouraging work effort.

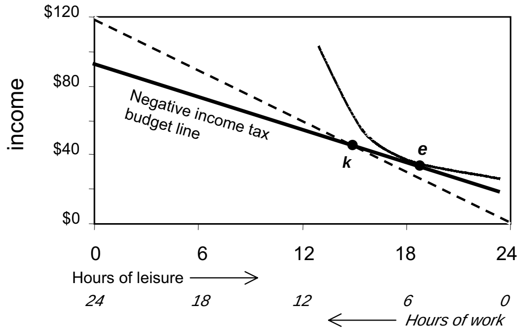


Fig. 4.21. Negative income tax
Here are the choices available to a worker earning \$5 per hour and receiving a negative income tax payment of \$20 per day, less 2/5ths of earned income. If the worker does not work, he receives \$20. If he works 8 hours he earns \$40 and receives a negative income tax check equal to $\$20 - 2/5 \times \$40 = \$4$ or a disposable income of \$44. If he works 12 hours he receives \$60 in wages plus the \$20 welfare payment but pays \$24 in taxes, netting \$56. If he works 10 hours he will earn \$50 and receive a welfare check of $\$20 - 2/5 \times \$50 = 0$. A 10 hour worker breaks even at intersection point k where his disposable income of \$50 is exactly what it would be if there were no negative income tax.

The "negative income tax" is an alternative form of low-income support summarized by the tax equation:

$$T = -k + tY\,. \tag{56}$$

This tax is equivalent to a positive lump-sum transfer of k dollars (i.e., a tax of $-k$) coupled with a tax at rate t on income. After-tax income is $Y_t = Y - T = k + (1 - t)Y$.

For example, if k is \$20 and $t = 2/5$ths, we have $T = -20 + 0.4Y$ and the consumer's after tax income is $Y_t = 20 + (1 - 0.4)Y$.

Examples:

1. With these negative income tax parameters, a worker who earned income of \$40 at McDonalds would have a tax of $T = -20 + 0.4 \times \$40 = -\$4$ or an after-tax income of \$44.
2. A welfare recipient who did not take a job would receive \$20.
3. A welfare recipient who earned $k/t = \$50$ would break even for $T = 0$ and $Y_t = Y$.

An advantage of the negative income tax is that it does not impose such an abrupt penalty for working. Economists, both liberal and conservative, have long been united in regarding the negative income tax as a great improvement over the traditional welfare system.[19]

4.4.5 *Saving and the rate of interest*

The indifference curve apparatus can be used to examine the response of saving to changes in the rate of interest. In order to simplify the analysis so that the alternatives confronting an individual can be displayed on a two dimensional graph, let us make the morbid assumption that our utility maximizer has only two years to live. With only a two-year life span, this year's income and consumption can be plotted on the abscissa and next year's on the ordinate as on Figure 4.22. Suppose that our utility maximizer will receive income of $M_1 = \$100$ this year and $M_2 = \$75$ next year. Then this individual can obviously consume at no-save point a on the graph by spending all of this year's income

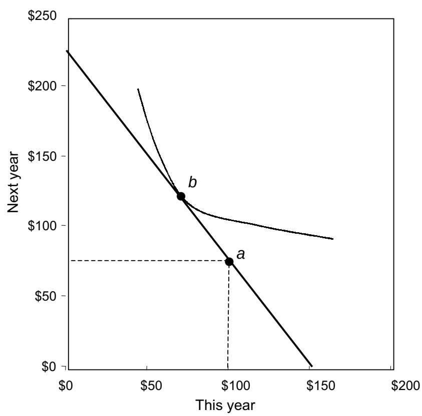


Fig. 4.22. Savings and the rate of interest
This period's income and spending are plotted on the abscissa; next year's are plotted on the ordinate. Our worker, earning $M_1 = \$100$ this year and anticipated earning $M_2 = \$75$ next year, could consume at no-saving point a. But with an interest rate of 50% this worker maximizes utility at point b, this year consuming $C_1 = \$75$ and next year consuming $C_2 = \$75 + (1 + 50\%) \times \$25 = \$112.50$. This year our worker is saving $M_1 - C_1 = \$100 - \$75 = \$25$ in order to consume more than earned income next year.

[19]The "Earned Income Tax Credit," (EITC), introduced into the Internal Revenue Code by President Nixon, is a weak form of the negative income tax that will be discussed in Chapter 7.4.3.

this year and all of next year's income next year. But if our consumer can borrow or lend at interest rate $r = 50\%$, a whole new range of additional possibilities is available. Our worker could consume precisely what is earned each year, saving nothing, which yields no-saving point a on the budget line with $C_1 = M_1$ and $C_2 = M_2$. Or our worker could consume nothing this year, earn \$50 in interest on the \$100 saved for the future, and consume \$225 next year. The full range of savings opportunities confronting our thrifty consumer is indicated by the budget line on the graph. In general, the equation for the budget line is

$$C_2 = (1 + r)(M_1 - C_1) + M_2\,, \tag{57}$$

where r represents the rate of interest, C_1 and C_2 this year's and next year's consumption, M_1 and M_2 are this year's and next year's income, and $M_1 - C_1$ is saving. For our specific example, we are given that $r = 50\%$, $M_1 = \$100$ and $M_2 = \$75$; therefore, the budget line plotted on the graph is

$$C_2 = (1 + 0.5)(100 - C_1) + 75\,. \tag{58}$$

The indifference curve on the graph reveals that this thrifty utility maximizer, encouraged by the high interest rate to forgo immediate gratification, might spend only \$75 this year in order to place \$25 in the bank, earn $\$12.50 = 50\%$ of \$25 in interest, and consume $\$75 + \$25 + \$12.50 = \$112.50$ next year. *Interest is the reward for waiting!*

How our utility maximizer responds to a change in the rate of interest is shown on Figure 4.23, which elaborates on the preceding graph. When the interest rate drops to zero, but with $M_1 = 100$ and $M_2 = \$75$, just as before, the equation for the new budget line is

$$C_2 = 75 + (1 + 0\%)(100 - C_1)\,. \tag{59}$$

Although the interest rate is now zero, the new budget line, like the old, goes through no-save point $(M_1, M_2) = (100, 75)$ because if the individual consumes all of his income in the first year he can only consume M_2 in the second year regardless of the rate of interest.

This utility maximizing consumer selects point c because the indifference curve through that point is tangent to the new budget line. The consumer has decided to respond to the drop in the interest rate by cutting C_1 from \$75 to \$50. That is to say, saving is now $M_1 - C_1 = \$50$, larger than when the interest rate was high. While this example is counter-intuitive,

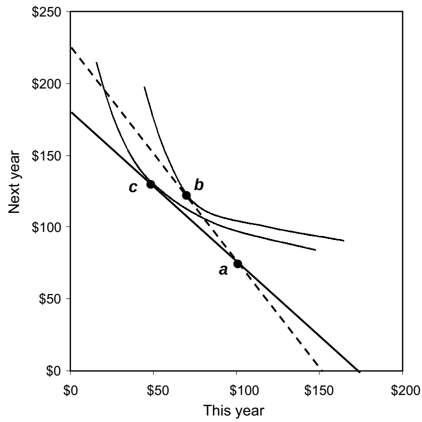


Fig. 4.23. A fall in the interest rate
When the interest rate drops to zero, the budget line rotates through no-saving point a. Now our utility maximizing worker consumes at point c. Consumption is now $C_1 = \$50$ and $C_2 = \$75 + (1 + 0)(\$100 - \$50) = \$125$. Savings of $M_1 - C_1 = \$100 - \$50 = \$50$ will augment consumption next year.

The reduction in the interest rate caused this consumer to increase saving from \$25 to \$50.

it is not implausible. Our utility maximizer could be saving funds to meet an important future need, such as paying the children's college tuition or providing for retirement. A consumer with different tastes might consume more this period. This would be the case for a consumer with the tangency point on the new budget-line located to the right of b.

The natural rate of interest

When we think of interest rates we think of bankers, but even in the absence of banks and money and uncertainty there would be an interest rate. To illustrate with absolutely the simplest possible example, consider Daniel Defoe's *Robinson Crusoe* (1719). The sole survivor of a shipwreck, Crusoe finds himself stranded alone on a deserted isle. Suppose that Crusoe has salvaged 120 pounds of corn from the shipwreck. He must choose how much of the corn to eat and how much to set aside as seed to plant for next year.

The choices available to Crusoe are displayed on Figure 4.24. The axes are labeled as on Figure 4.22 — this year or next year, the present or the future. Point a denotes the 120 pounds Crusoe could consume now if he put nothing aside for the future. If he were to store some of the grain for next year, but not plant it, he would have the choice of the opportunities on the dotted line with slope -1 emanating from point a. Suppose, instead, that he considers planting some of his corn. The curve labeled $T(C_1)$ shows the production possibilities that are available to Crusoe. Specifically, if C_1 pounds of corn are consumed this year, $T(C_1)$ will be available next year. Because nature is bountiful and farming is a productive enterprise, there is a reward for waiting: every additional pound of corn used as seed yields more than a pound of corn at harvest time; i.e., $dT/dC_1 < -1$. That is why at every point this production possibility curve is steeper than -1. There is a reward for waiting in that Crusoe's total consumption of corn will be larger the more corn he puts aside for the future.

Now Crusoe will select tangency point e, judging by the indifference curve on the graph. At this tangency point the slope of the indifference curve and the slope of the production possibility frontier must be equal. That slope must be greater than one because farming is productive, the crop harvested in the fall being larger than the pounds of seeds that were planted in the spring. This slope is said to be equal to $-$ (1 + the *natural rate of interest*), just as the slope of the budget line on Figure 4.22 is $1 + r$. But the natural rate of interest implicit in Crusoe's choice between

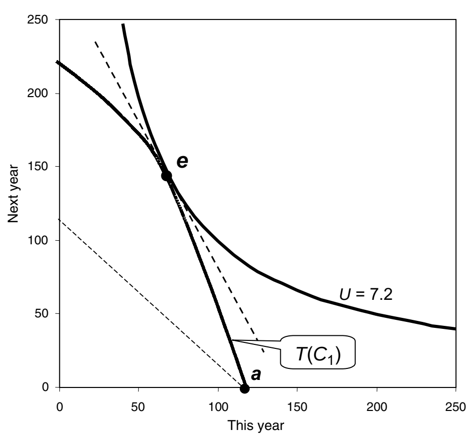


Fig. 4.24. The natural rate of interest
The production possibilities curve, $T(C_1)$, shows the technological options available in choosing whether to consume this year or next – the more we consume this year the less we will have available tomorrow.

The indifference curve reflects the consumer's choices between today and tomorrow.

The slope of the dotted tangency line through equilibrium point e depends on both the technological options and the tastes of the consumer. The natural rate of interest plus 1 equals minus the slope of the tangency line.

the present and the future is not determined by banks because there are no banks on that lonely isle. This interest rate is determined at tangency point e where the preferences of Crusoe between present and future consumption are balanced with the technological tradeoff between the present and the future made possible by investing part of his corn as seed.

What are we to conclude from this parable? First of all, our economy's natural rate of interest is that rate that balances the desire of consumers to save, as determined by their preferences between current and future consumption, with the technological tradeoff between the present and the future provided by opportunities for investment in productive capital. Second, we should observe that the natural rate of interest is not the same thing as the rate of interest you could earn by placing your funds in a savings account at your local bank; nor is it the rate of interest you must pay if you have credit card or mortgage debt. Uncertainty causes the rate of interest charged on a credit card or a mortgage on a home to depart from the natural rate of interest because a risk premium must be added to compensate the lender for the possibility that the borrower may default on the loan. In addition, governments borrow and central banks intervene in financial markets, which may cause a further deviation of money market rates from the natural rate of interest. Although the natural rate of interest may not be directly observable, it is a concept of fundamental importance. It explains why interest rates tend to be higher in countries that have many investment opportunities ripe for development. And the concept helps us to understand that interest rates are influenced by both the availability of profitable investment opportunities and the public's preferences between current and future consumption.

4.5 The box diagram — Two goods and two consumers

Considering the problem of allocating two goods fairly between two consumers will clarify the distinctions between the concepts of equity and efficiency. To analyze this problem we will make use of a clever graphical device known as the box diagram. It is sometimes referred to as the "Edgeworth Box," after its developer, Francis Edgeworth [1845–1926]. Of Irish birth, Edgeworth taught at the University of London and at Oxford. He analyzed utilitarianism in his *Mathematical Psychics*, 1881.

Suppose that we have 10 units of good X and 20 units of good Y to allocate between Dick and Jane. If we let X_J and X_D indicate the quantities of commodity X we allocate to our two consumers, we must obviously have

$X_J + X_D = 10$; similarly, $Y_J + Y_D = 20$. The possible allocations must all lie in the allocation box displayed on Figure 4.25. The important thing to note about the box is that we can measure quantities from either Dick's or Jane's perspective. If Dick has $X_D = 2$ and $Y_D = 15$, he is located at point a, where we measure Dick's consumption from the origin at the lower left (south west) corner of the box. But the same point represents Jane's consumption if we choose to measure distance from the upper right hand corner of the box, as indicted by the second set of scales on the upper and right edges of the allocation box: Jane has $X_J = 10 - 2 = 8$ and $Y_J = 20 - 15 = 5$. The task of allocating the fixed quantity of X and Y between our two individuals is equivalent to finding the appropriate point in the allocation box.

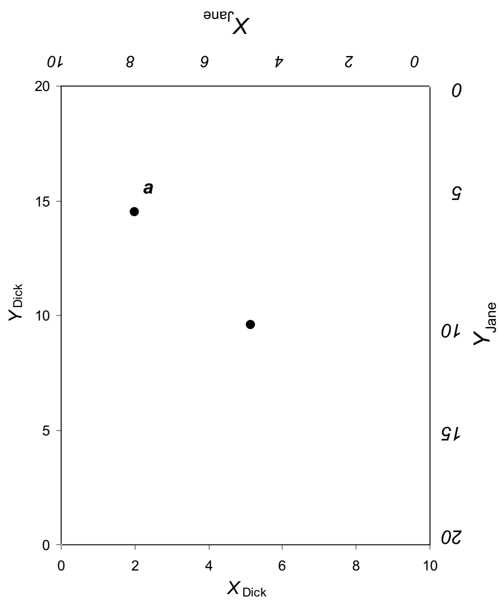


Fig. 4.25. The allocation box
Because there are 10 units of good X and 20 units of good Y to divide between two consumers, Dick and Jane, the allocation box showing all conceivable allocations of the two goods between our two consumers is 10 units wide and 20 units high. At point a Dick is receiving 2 units of X and 15 units of Y, which means that there are $8 = 10 - 2$ units of X and $5 = 20 - 15$ units of Y left for Jane. Jane's allocation can be read directly off the graph at point a with the alternative scales emanating from the upper right hand corner of the graph, as may be readily seen by turning your book upside down.

4.5.1 *Equity versus efficiency*

One possible allocation is to divide the available supply of the two goods equally, as indicated by the point $(5, 10)$ on the graph. While this may seem like an eminently reasonable solution, we shall find that this might be "inefficient" in the sense that there may exist another allocation that would make at least one person better off without making anyone worse off. Examination of Figure 4.26 will reveal why this might happen. This graph will also help to clarify what is meant by an efficient allocation of resources; and it will allow us to specify necessary conditions that must be satisfied in order to achieve an efficient allocation of resources among consumers. The *box diagram* is constructed by carefully superimposing the indifference maps

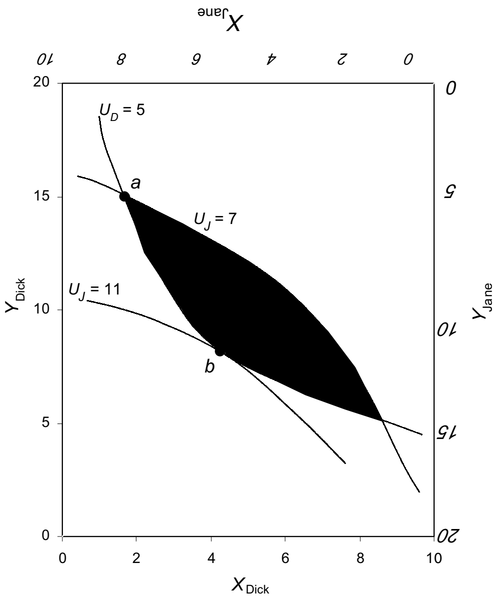


Fig. 4.26. The box diagram
The box diagram is obtained by adding Dick and Jane's indifference curves to the allocation box. While Dick's indifference curves are conventional, Jane's are inverted in orientation because the goods that she receives are measured with the alternative scale emanating from the upper right hand corner of the graph.

Observe that point a is not efficient. Both Dick and Jane like all the points inside the shaded banana better than point *a*. Point *b*, on the other hand, is *efficient* because there is no alternative allocation that is preferred by one of the individuals that doesn't make the other worse off.

of Dick and Jane onto the allocation box.[20] Although Dick's indifference curves are similar in shape to those that we considered earlier, Jane's had to be inverted in order to line up with her scale on the allocation box. Remember, the origin on the allocation box, from Jane's viewpoint, is at the upper right-hand corner. It was necessary to plot Jane's indifference map upside down because any movement downward and to the left is an improvement in her position. That explains why her indifference curves bend in the opposite direction from Dick's — turning the graph upside down will confirm that Jane's indifference curves bend conventionally when viewed from her perspective.

- Point a on the graph is *not* efficient: At this point Dick is on indifference curve $U_D = 5$ and Jane is on indifference curve $U_J = 7$. Point a is not efficient because any of the points between point a and b on Dick's indifference curve $U_D = 5$ would leave Dick as well off as before but make Jane better off in that she would be southwest of indifference curve $U_J = 7$.

Indeed, Dick and Jane would both find any of the points inside the entire shaded banana shaped area between indifference curves $U_D = 5$ and $U_J = 7$ are better than point a.

Key Concept

Economists call adjustments that make at least one individual better off without making anyone worse off *Pareto improving changes*, after famous Italian economist Vilfredo Pareto [1848–1923].

- Point b on the graph is efficient: It is impossible to move from this tangency point without making at least one individual worse off, which means that no Pareto improving change is possible.

If we move down and to the left we can place Jane on a higher indifference curve but only by making Dick worse off. If we move up to the right we make Dick better off but Jane worst off. If we move upward toward the left or downward to the right we might make both worse off. We can move from point b in the allocation box to any other point only by making at least one individual worse off.

[20] This maneuver implicitly assumes that the utility functions are "individualistic"; i.e., it is assumed that the utility each market participant enjoys is not affected by knowledge of what others are receiving. If there is jealousy, we might write $U\ (X\ \ ,X\ \ ,Y\ \ ,Y\ )$, which complicates the analysis tremendously.

Note that the indifference curves are tangent at efficient point b. If they crossed there would be a new banana shaped region in which mutual improvement would be feasible. At any point where the indifference curves of the two individuals are tangent the indifference curves have the same slopes; i.e., the marginal rates of substitution of good X for good Y is the same for both individuals. Thus we can say that efficiency requires that resources be allocated so that

$$MRS_{xy}^{D} = MRS_{xy}^{J}\,.^{21} \tag{60}$$

4.5.2 *The gain from trade*

Now suppose that initially the goods were allocated between Dick and Jane so that they were at inefficient point a. While this initial allocation is not efficient, the market participants can engage in mutually beneficial trade. Suppose Dick offers to trade one unit of his Y for a unit of Jane's X. Since the trade is beneficial to Jane as well as Dick, she will accept the offer. This means that a mutually beneficial Pareto improving change will take place. However, they will not stop trading after exchanging one unit of X for one unit of Y because the resulting allocation may still not be efficient; they can both do even better through further trading. If trading were to stop at an inefficient point, Dick and Jane would not be taking advantage of a mutually beneficial exchange. Is it not reasonable to expect trading to continue until an efficient point is reached where both individuals cannot simultaneously be made better off through further trading?

Economists agree that it is usually a mistake for the government to prohibit mutually beneficial trade. Exceptions are based on two arguments:

1. Both parties must be capable of making an informed judgment about what is in their own best interest. For example, the government may prohibit the sale of cigarettes to minors on the grounds that they are not old enough to make informed judgments.
2. Further, others must not be negatively affected by the exchange. An *external diseconomy* is said to arise when others are affected by the consumption decision, as when the explosion of firecrackers disturbs the neighborhood or smoking in a restaurant creates discomfort.

[21] A corner tangency, similar to that considered on Figure 4.10, is the only exception. Condition (60) must hold for efficiency if both consumers are consuming positive amounts of both goods.

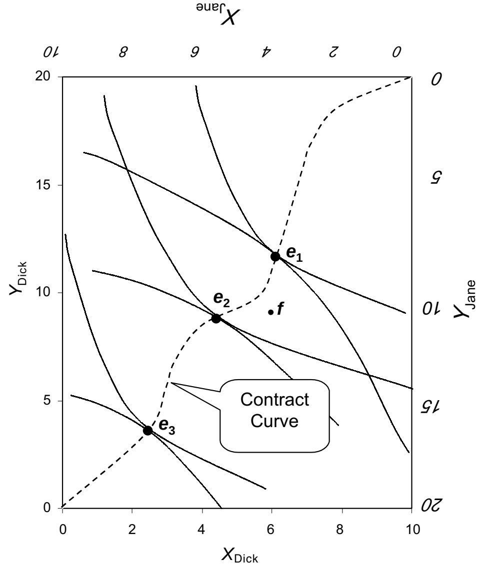


Fig. 4.27. The contract curve
An allocation of resources is efficient if no one can be made better off without making at least one other person worse off. Each point where Jane's indifference curve is tangent to one of Dick's is efficient. The contract curve is the set of all efficient points. Point f is inefficient, but some (including Dick) would argue that it is fairer than efficient point e_3.

4.5.3 *The contract curve*

The next box diagram, Figure 4.27, shows a whole set of efficient points. Each efficient point is characterized by the tangency of the two individuals' indifference curves. The curve constituting the locus of all efficient points is called the *contract curve.* At any point off the contract curve the indifference curves of our two consumers cross, and it is possible to make at least one person better off without making anyone worse off. As the concept of the contract curve makes clear, there is a whole range of efficient points. Obviously, Dick prefers efficient points on the upper-right end of the contract curve while Jane prefers those on the lower left.

The contract curve illustrates the following points:

- It is always possible to improve on an *in*efficient resource allocation by making a Pareto improving exchange.
- Some efficient resource allocations seem manifestly unfair, such as those at the extreme ends of the contract curve.

Some efficient points may be worse than some inefficient points. For example, point e^3 gives such small shares to Dick that many would judge it to be worse than point f, even though the latter is inefficient. Equity matters!

4.5.4 *Competitive equilibrium*

The box diagram can be used to illustrate what is known as the First Fundamental Law of Welfare Economics. This is the proposition that a competitive market generates an efficient allocation of resources.

Initial situation: A competitive market is characterized by a large number of participants, so suppose there are 100 participants with indifference maps like Dick's and 100 with indifference map like Jane's. Also, let us suppose that initially each of the 100 Dick's has, say, 2 units of X and 15 units of Y while each of the 100 Jane's has 8 units of X and 5 units of Y. Then point a on Figure 4.28 denotes this initial allocation.

Competitive equilibrium: Suppose that all 200 market participants are allowed to trade at the price ratio specified by the slope of the straight line through a. Then each Dick will want to move to point e and each Jane will want to move to the same point. The market clears, for the amount of X that each Jane is willing to give up at this price ratio is precisely equal to the quantity that each Dick wants to purchase, so they all move to the competitive equilibrium at point e.

This is a *competitive equilibrium* because all three of the following conditions are satisfied:

1. All parties are price takers.
2. No trader would want to trade away from point e, given the prevailing price ratio.
3. Supply is equal to demand for all commodities.

Efficiency of competitive equilibrium

Now condition (2) for competitive equilibrium requires that the indifference curves be tangent to the budget line; hence they are tangent to each other; therefore, the competitive equilibrium is efficient, as asserted by the First Fundamental Law of Welfare Economics.

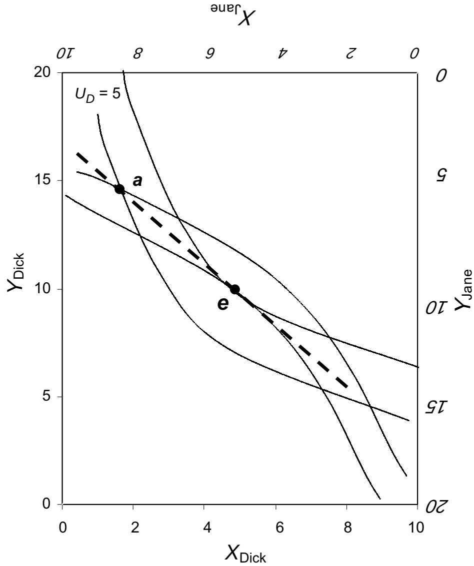


Fig. 4.28. Competitive equilibrium
Dick has an initial endowment at point a of X = 2 and Y = 15; Jane has X = 8 and Y = 5. With market prices of p = \$3 and p = \$5, Dick will want to slide down the budget line to tangency point e, giving up 5 of Y in exchange for 3 of X. And when Jane considers trading at these market prices she will maximize her utility by exchanging 3 of X for 5 of Y, also moving to point e. With these prices the supply by Dick of 5 units of Y precisely equals Jane's demand; and the supply by Jane of 3 units of X precisely equals the quantity of X that Dick wants to buy. Thus the market clears in happy competitive equilibrium.

Equity versus efficiency

All the points on the contract curve are efficient, but some seem manifestly unfair. In particular, giving everything to Dick on Figure 4.27, the point $(10, 20)$ is efficient, for we cannot make Jane better off without making Dick worse off nor can we make Dick better off without hurting Jane. But no one (perhaps not even Dick) would say that such an allocation is fair or equitable. Clearly, just showing that competitive markets are efficient does not mean that they necessarily result in an appropriate allocation of resources. This suggests that it may be necessary for government to intervene in the operation of markets in order to insure that the outcome is fair or equitable. But how big a role should this be?

Help in considering this issue is provided by the Second Fundamental Law of Welfare Economics, which states that any Pareto efficient allocation of resources can be achieved as a competitive market equilibrium. In terms of Figure 4.27, this is the proposition that any desired point on the contract curve (i.e., any efficient allocation of resources) can be obtained from an appropriate set of initial allocations. For the simple case of two consumers this is easily visualized by considering any particular point e on the contract curve. Draw a tangent line to Dick's and Jane's indifference curve at this point (their indifference curves must be tangent because they are on the contract curve). From any initial point on that tangent line competitive forces will lead to the desired efficient point. This Second Fundamental Law of Welfare Economics implies that an appropriate allocation of resources could be obtained in two steps. For the first step the government can establish a fair initial distribution of resources. Second, the market mechanism can be relied upon to guide the economy to the corresponding efficient distribution without further government intervention. The hope might be that if the initial allocation were fair and equitable than the resulting competitive equilibrium would also be fair.

In Chapter 7 we shall be looking at how unevenly income is in fact distributed and how governments can affect the distribution through tax policy, subsidies, and welfare programs.

Summary

1. Positive economics has to do with explaining the world as it is. Economists explain consumer behavior by invoking the assumption that consumers are motivated by the desire to maximize happiness or "utility." Normative economics is concerned with the world as it ought to be. Economic thought has been shaped by utilitarian philosophers, who argued that society should strive to maximize the sum total of happiness. In contrast, John Rawls argued that the objective of society should not be to maximize the sum total of happiness but rather to maximize the position of the least advantaged.
2. Because it is not possible to compare one person's utility with another, even if we agreed that the objective of society should be to maximize total satisfaction, the principle of utilitarianism does not offer decisive guidance on equity issues, such as when trying to decide how income ought to be distributed. Limited guidance on policy issues is offered by the Pareto Principle, the proposition that a change is clearly worth while if it will make at least one person better off without making anyone else worse off.

3. Where do demand curves come from? We learned how to derive a demand function explaining consumption of a commodity in terms of its price, the price of other consumption goods, and income. The strategy is to maximize the consumer's utility function subject to a budget constraint.
4. Only certain types of behavior are compatible with the assumption that individuals are rational utility maximizers. In particular, we found that the demand curve of a rational consumer must be homogeneous of degree zero in income and prices. This means that if a period of inflation leads to a proportionate rise in all prices and wages, consumers will end up consuming the same quantities as before the inflation.
5. The effect of inflation is more complicated when, as is usually the case, some prices rise more than others. We derived an expenditure function showing the income required to attain a specified level of utility, given prices. We showed that except when all prices rise in the same proportion, consumers are overcompensated for inflation if their wages (or Social Security benefits) increase by enough to allow them to purchase the same bundle of goods as before the inflation.
6. We used a graphical technique — indifference curves — to analyze the response of a utility maximizing consumer to a change in income and price. We found that consumption of some goods may decline when income increases — such goods are called inferior goods. We found it useful to partition the effect of a price change on consumer behavior into two parts, the income effect and the substitution effect. The substitution effect asks how consumption would be affected by a price change *if*, hypothetically, income were adjusted to as to allow the consumer to remain on the initial indifference curve (enjoy the pre-price-change level of utility) — a price increase always leads to a negative substitution effect, consumption of the good being reduced. The income effect is the change in the quantity consumed that would occur *if* relative prices had not changed but the consumer's money income changed by just enough to allow the enjoyment of the same level of utility (i.e., move to the same indifference curve) as is realized after the price change.
7. We noted that if a commodity's income effect is sufficiently negative, a price increase may lead to an increase in the consumption of the good, which means that the assumption of maximizing utility does not imply the Law of Demand. Such "Giffen goods" are thought to be extremely rare.

8. Indifference curves help analyze how changes in wages motivate workers and changes in interest rates affect savings behavior. We found that a fall in wages might conceivably lead a utility maximizing consumer to work longer hours! We also saw how poorly designed welfare programs can discourage people from working. We noted that a rise in interest rates might conceivably lead to a reduction in saving.
9. The box diagram is a graphical technique used in analyzing the distribution of fixed available quantities of two goods between two consumers. An efficient allocation is defined as one in which no individual could be made better off without making someone else worse off. The set of all the efficient points on the box diagram is called the contract curve.
10. We observed that some efficient allocation of resources are manifestly unfair — it is necessary to worry about questions of equity as well as efficiency issues in formulating economic policy.
11. The box diagram also revealed that under competitive conditions any efficient allocation of resources can be obtained by the price mechanism.

Appendix 4.1. Lagrangian Multipliers

Economists frequently encounter constrained maximization problems. In Chapter 3 we worried about maximizing tax revenue and in Chapter 4 we have analyzed the behavior of consumers who are attempting to maximize their utility subject to a budget constraint. In this appendix we compare an application of the Lagrangian multiplier technique with the substitution technique used in the chapter.

In Section 4.3.2 we derived a consumer's demand function by solving the problem of maximizing utility function

$$U(X,Y) = X^{\frac{3}{4}}Y^{\frac{1}{4}} \tag{24}$$

subject to the budget constraint

$$p_xX + p_yY = M\,. \tag{28}$$

Our strategy was to convert the original problem of maximizing a function of two unknowns subject to one constraint into an equivalent problem involving only one unknown and no constraints. The trick was to use the budget constraint to eliminate the unknown Y from the utility function, which left us with the standard problem of maximizing a new function:

$$U^*(X) = U\left(X, \frac{M}{p_y} - \frac{p_x X}{p_y}\right) = X^{\frac{3}{4}}\left[\frac{M}{p_y} - \left(\frac{p_x}{p_y}\right)X\right]^{\frac{1}{4}}. \qquad (31)$$

The Lagrangian multiplier technique is an alternative procedure leading to the same solution.

Step #1

The first step is to form the Lagrangian expression

$$L(X, Y, \lambda) = X^{\frac{3}{4}}Y^{\frac{1}{4}} - \lambda(p_x X + p_y Y - M)\,. \qquad (61)$$

This Lagrangian function involves three unknowns, both X and Y, and a new parameter λ, which is known as the Lagrangian multiplier.

Step #2

The second step is to solve for the maximum of $L(X, Y, \lambda)$ by setting its derivative with respect to each of its three arguments equal to zero:

$$\frac{\partial L(X, Y, \lambda)}{\partial X} = \frac{3}{4}X^{-\frac{1}{4}}Y^{\frac{1}{4}} - \lambda p_x = 0\,, \qquad (62)$$

$$\frac{\partial L(X, Y, \lambda)}{\partial Y} = \frac{1}{4}X^{\frac{3}{4}}Y^{-\frac{3}{4}} - \lambda p_y = 0\,, \qquad (63)$$

$$\frac{\partial L(X, Y, \lambda)}{\partial \lambda} = p_x X + p_y Y - M = 0\,. \qquad (64)$$

Step #3

The final step is to solve this system of three simultaneous equations. A convenient approach for this example is to write the first two equations as

$$\frac{3}{4}X^{-\frac{1}{4}}Y^{\frac{1}{4}} = \lambda p_x\,, \qquad (65)$$

$$\frac{1}{4}X^{\frac{3}{4}}Y^{-\frac{3}{4}} = \lambda p_y\,, \qquad (66)$$

Dividing (65) by (66) yields

$$\frac{3Y}{X} = \frac{p_x}{p_y}\,, \qquad (67)$$

or

$$p_y Y = \frac{p_x X}{3}\,. \tag{68}$$

Thus we spend 1/3 as much on Y as on X. This result, coupled with the 3rd condition for a maximum of the Lagrangian function, which in this application just reproduces the budget constraint, yields the demand equation $X = (3/4)M/p_x$.

Interestingly, Lagrangian multiplier λ is the derivative of the indirect utility function (38) with respect to M, which is sometimes called the "marginal utility of income." Thus λ tells us how much utility would increase by as a result of a small increase in money income.

The Lagrangian approach yields the same answer as the method of substitution, as explained in calculus textbooks. The Lagrangian technique seems more complex than the method of substitution in this example, but it comes into its own when there is more than one constraint.

Key Concepts

Exercises

1. Cindy has utility function $U_C(X_C) = X_C^{1/2}$. Robert has utility function $U_R(X_R) = X_R$. You are to decide how to allocate 4 units of X between these two citizens.

   a. Sketch the utility possibility frontier.
   b. Determine the allocation of the 4 units of X that will maximize the sum total of happiness; i.e., maximize $U_\Sigma = U_C(X_C) + U_R(X_R)$, subject to $X_C + X_R = 4$. Then calculate U_C, U_R, and U_Σ. What U_C and U_R will maximize U_Σ if the supply of X increases to 6?
   c. Determine the allocation of X that will maximize the position of the least advantaged; i.e., $\max[\min(U_C(X_C), U_R(X_R)]$ given that $X_C + X_R = 6$.

2. Julie has utility function $U = 2X + \ln Y$. She maximizes utility subject to the income constraint $M = p_x X + p_y Y$.

   a. If her income is $M = 100$ and $p_x = \$1.00$ and $p_y = \$2.00$ how much X and how much Y will she consume?
   b. Determine Julie's demand function for X; i.e., find the function $X(p_x, p_y, M)$.

3. Draw an indifference map for the case of an inferior good that is not a Giffen good; i.e., consumption of good X decreases when income increases or the price of X increases. Compare with Figure 4.15. Indicate both the substitution effect and the income effect of the price increase.
4. Tom's utility function is $U(M, L) = 10 + 3L^{1/2} + 12M$, where L is leisure and M is income to spend on consumption goodies.

   a. What is the marginal utility of leisure (L)?
   b. What is the marginal utility of income (M)?

      Note: This is an example of a "constant marginal utility of income" utility function.
   c. Determine the labor supply function showing the number of hours our utility maximizer will work as a function of the wage rate w, where $M = w(24 - L)$, $0 \le L \le 24$.
   d. If Tom receives a \$1,000 inheritance, will he work less and enjoy more leisure?
   e. How will the supply function of labor be affected if he receives a welfare benefit $B = \max(\$20 - M, 0)$?

5. Consider Mary, a utility maximizer consuming goods X_1 and X_2 at prices p_1 and p_2. Her income is M dollars. Suppose her utility function is

$$U = 2X_1^{1/2} + X_2 \,.$$

   a. Solve for the demand function for X_1 showing how the quantity consumed of this commodity depends on the price of the two goods and income. Note that you are to maximize U subject to both the income constraint $M = p_1X_1 + p_2X_2$ and the sign constraints $X_1 \geq 0$ and $X_2 \geq 0$.
   Caution: Why does it make a difference if $(p_2/p_1)^2 \geq M/p_1$?
   b. Derive the *indirect utility function* showing how Mary's utility depends upon her money income and the prices of the two commodities; i.e., find a function of the form $U = g(M, p_1, p_2)$.
   c. Find from the indirect utility function the *expenditure function* showing how much income Mary must have to achieve a specified level of utility, given prices; i.e., find a function $M = f(U, p_1, p_2)$.
6. Determine the demand functions for X and for Y implied by the utility function $U = X^3Y$. Does this utility function satisfy the law of diminishing marginal utility?
7. Carefully draw the supply of labor curve generated by the indifference map plotted on Figure 4.19. Contrast with Figure 4.18.
8. It is said that a cut in personal income taxes might provide such a stimulus to work effort as to cause the income tax to yield more revenue to the government. Others assert that a cut in the personal income tax might lead to a decrease in work effort. Is either (or both) of these assertions compatible with the assumption that individuals maximize utility? Explain, drawing appropriate indifference maps.
9. Sorry, John will only live for two years! (This morbid two-period example keeps the algebra simple.) His income is Y_1 in year 1 and Y_2 in year 2. He can borrow or save at interest rate $r = 20\%$. If he saves $Y_1 - C_1$ in year 1 he can spend $C_2 = Y_2 + (1 + r)(Y_1 - C_1)$ in period 2. He will not leave a bequest to his estranged children, spending everything!

   a. Can we write John's budget constraint as

$$Y_1 + Y_2/(1 + r) = C_1 + C_2/(1 + r)\,;$$

i.e., is the "Present Discounted Value" (PDV) of the income stream equal to the present discounted value of consumption?

b. Suppose that $Y_1 = Y_2 = \$100$ and that $r = 20\%$, and $C_1 = \$80$; what is C_2? Plot the initial income point and the consumption point and the appropriate budget line on an indifference map with period 1 consumption and income on the abscissa and period 2 consumption and income on the ordinate.

c. Suppose John's utility function is $U = \ln(C_1) + \ln(C_2)$. Determine how much he will consume if his income is \$100 each period and the interest rate is zero. How much will he consume in period 1 if the interest rate is 25%? Will he save more or less as a result of the increase in the rate of interest?

10. Definition: A function $Y = f(X_1, X_2)$ is homogeneous of degree k if and only if for any coefficient ρ,

$$\rho^k f(X_1, X_2) = f(\rho X_1, \rho X_2,) .$$

Example: if $k = 0$, $\rho^k = 1$ and $f(X_1, X_2) = f(\rho X_1, \rho X_2,)$.
Determine which of the following demand functions is homogeneous of degree 0 in income and prices:

a. $X(p_x, p_y, M) = 5 + 3p_y - 4p_x + 3M$.
b. $Y(p_x, p_y, M) = (p_x + M)/p_y$.
c. $\ln Y(p_x, p_y, M) = 10 + 3 \ln M - \ln p_y$.

Query: If a demand function is not homogeneous of degree zero in money income and prices, can it describe the behavior of a utility maximizing individual? Explain.

11. Draw indifference maps illustrating each of the following possible outcomes:

a. An increase in the price of a commodity leads to an *increase* in the quantity consumed, given money income and all other prices.
b. An increase in the tax on wage income leads to an increase in hours worked.
c. An increase in interest rates leads to a reduction in saving.

12. In England during World War II, the slogan "fair shares for all" was popularized in order to encourage patriotic citizens to abide by the strict rationing of basic foodstuffs that was imposed in an effort to

cope with severe wartime shortages. Thus if there are 10 units of X and 20 of Y to be allocated between Dick and Jane we would have $X_D = X_J = 5$ and $Y_D = Y_J = 10$. Show on a graph that while the equal division might be considered fair, it does not necessarily yield an efficient allocation of resources.
Hint: You are to construct a box diagram with an appropriate banana shaped region.

13.# The Stone-Geary utility function is of the form

$$U(X_1, X_2) = (X_1 - \beta_1)^{\alpha_1}(X_2 - \beta_2)^{\alpha_2} ,$$

where $\beta_1 \geq 0$ and $\beta_2 \geq 0$ are interpreted as subsistence or "wolf-point" levels of consumption.

a. For what values of the parameters α_1, α_2, β_1, and β_2 does the Stone-Geary utility function reduce to equation 24 of Ch 4?
b. Derive the demand function for X_1 implied by the Stone-Geary utility function.
   Hint: You could simplify the problem by working with the transformed variables $X_1^* = X_1 - \beta_1$, and $X_2^* = X_2 - \beta_2$.
c. Demonstrate that the demand for X_1 is homogeneous of degree 0 in income and prices.
   Note: The Stone-Geary utility function yields a demand function that is linear in relative prices and real income, which facilitates its estimation from empirical data by econometric techniques.

14.* Consider the following policy issue:
Suppose that the government wants to undertake a Reform to improve the productivity of the economy. As a result, everyone will be better off, but the improvement in life will not affect people equally. A million people (people who respond energetically to the incentives in the plan and people with certain skills) will see their incomes triple while everyone else will see only a tiny income increase, about 1%.

a. Suppose that the only option is either the status quo or the Reform. Would you support the change? Why?
b. Do you think that either Jeremy Bentham or John Rawles or both would support the change? Explain why.

When in May of 1990 random samples of residents in Moscow and New York were asked precisely this question by telephone interviewers, 55% of Moscow residents and 38% of New York city residents said they would support the change.[22]

c. Why do you think the proposal did not receive stronger support?

[22] See Robert Shiller, Maxim Boycko and Vladimir Korobov, "Popular Attitudes Toward Free Markets: The Soviet Union and the United States Compared," *American Economic Review*, June 1991, pp. 385–400.

5

The Business Enterprise: Theory of the Firm

5.1 Introduction

In 1975 William H. Gates III and Paul G. Allen founded Microsoft as a partnership. That same year they shipped, as their first product, an implementation of the BASIC language interpreter — this was the first computer language program to run on a personal computer. In 1981 they reorganized as a privately held corporation, Microsoft, Inc. In 1986 Microsoft went public with an initial public offering that raised $61 million. In May 1998, the Antitrust Division of the U.S. Department of Justice and a group of 20 state Attorneys General filed two antitrust cases against Microsoft. In 1999 Microsoft realized profits of $11.891 billion on sales of $19.747 billion, employing more than 34,000 people and incurring taxes of $4.106 billion. That, in a nutshell, is the twenty-five year history of what is arguably the most successful business enterprise in the history of capitalism.

This chapter starts with an explanation of the various forms of business organization, including a discussion of the differences between partnerships and corporations. You will also learn basic accounting principles that business enterprises use to report their financial condition to their owners and the tax authorities.

Later in this chapter we shall develop the theory of the firm. This basic engine of analysis examines the implications of the assumption that business firms maximize profits. We will ask how a profit maximizing firm will respond to an increase in wages or interest rates and how its pricing decisions may be influenced by a change in taxes.

The concluding section discusses Management Science, a discipline concerned with the application of mathematical techniques to the practical problems of business. This section considers the problem of efficient inventory management and explains the Just-in Time style of production management pioneered by Toyota.

Left for Chapter 6 is a discussion of alternative types of market organization and a report on the procedures used by governments to limit and control monopolies in an attempt to establish an environment for business in which profit maximization will best contribute to the benefit of society.

5.2 Organization of the firm

5.2.1 *Types of organizations*

Individual proprietorship

A student is recruited by one of her instructors to develop a web page. They agree that the student will receive $500 when the project is satisfactorily completed. With this arrangement, the student is not an employee. She is "self employed" — she has become a *sole proprietor*. This is the simplest form of business structure. The instructor who recruits her to undertake the project will not have to worry about income tax withholding or social security taxes.[1] The student is legally required by the Internal Revenue Service to report her income on "*Schedule C: Profit or Loss from Business* (sole proprietorship)," which is a bit of a pain unless she uses *Turbo Tax* or some other income tax software to expedite the April 15th process. But using Schedule C may make it possible for her to deduct at least a part of the cost of her computer and the software used on the project when the time comes to calculate her profits from the project and the taxes that she owes.

Partnership

If our student's web page business prospers, she may find it advisable to recruit friends to help her with her projects. She might hire them as employees, she might have them work for her under contract (in which case her friends would also become sole proprietors[2]), or she might form a *partnership.* Any two or more people can agree to work together in a partnership. Physicians, lawyers, accountants and other professionals often adopt this form of organization.

When forming even the simplest of partnerships, it is essential that the participants draw up a written agreement carefully specifying the fraction of the work to be performed by each partner, the fraction of the capital that each shall provide, and how they will share profits, losses and debts. Even though the partners may have carefully specified their mutual obligations when they initially established the business relationship, it all too often happens that, after the initial enthusiasm for the business has worn thin, there will be a falling out, each partner thinking he is doing two thirds of the work and providing three quarters of the thinking.

A major limitation of the partnership mode of organization is that each partner assumes unlimited liability. Suppose your partnership fails. If you owned half the partnership, you will be assigned one half of the outstanding debts of the firm to be paid out of your own pocket. Worse than this, if

[1] The professor may be required to file a Form 1099 with the IRS notifying them of the amount that the student is receiving. The student will have to pay the equivalent of the Social Security and Medicare taxes when she files her income tax return.

[2] The Internal Revenue Service has rather strict requirements that must be met before one can claim that one is a proprietor rather than an employee.

your partner lacks the financial resources to pay his half of the partnership's debt, you will be legally liable to shoulder the entire burden![3]

Corporations

Corporations are the dominant form of business organization. A corporation is an artificial "person."[4] At one time, anyone seeking to form a corporation had to obtain a charter through special legislative act. Thus, the British East India Company was established by a special act of Parliament in 1600. Today it is relatively easy in the United States to incorporate in any of the fifty states. Indeed, some law firms advertise on the back pages of the *New York Times* that you can incorporate over the phone for a fee of $99.95.

- One advantage of incorporation is that, unlike a partnership, the organization continues even if one of the investors dies — corporations attain immortality.
- Another advantage is that the owners of the corporation enjoy the right of *limited liability.* Limited liability means that if the corporation goes bankrupt, perhaps because the courts rule against it in a major suit, the investors only lose what they have already invested in the firm. While limited liability has a long history as a special privilege granted to those entrepreneurs enjoying political connections, the first general limited-liability law for manufacturing companies was passed by New York State in 1811. Other states quickly followed in order to avoid disadvantaging their own corporations, and Britain adopted limited liability in 1854. By reducing the risk of owning stock, establishing limited liability encouraged even passive investors to shift their funds from bonds to stocks. The

[3]Young associate lawyers at major law firms aspire to make partner, but they are well advised to check out the liabilities of the firm and to be aware that they might be required to help pay any losses arising from pending lawsuits.

[4]The Fourteenth Amendment to the United States Constitution, ratified shortly after the Civil War, not only provided that all persons born in the United States, regardless of race, are citizens of the United States; it also provided that no state shall "deprive any person of life liberty or property without due process of law." The Courts subsequently interpreted the word "person" in the *due process clause* to include corporations and ruled that the clause prevented the states from regulating public utilities to the point where they receive anything less than a "fair rate of return." Economists serving as expert witnesses at public utility commission rate hearings are amply compensated for testifying about what constitutes a "fair rate of return."

establishment of limited liability made it much easier for corporations to raise the funds required for financing large-scale industrial capitalism.

- The *Corporate Profit Tax* is the primary disadvantage of incorporation. The federal tax is 15% on the first $50,000 of net income, but it rises in steps to a maximum rate of 35% on taxable net income over $10 million — such is the price for immorality and limited liability.

As can be seen from Figure 5.1, there are many more sole proprietary businesses than there are corporations or partnerships. But corporations are much larger on average, so if we look at the size of firms, as measured by their sales receipts, the corporations clearly dominate. Corporations today are the dominant form of economic organization in all modern industrial societies.

In addition to sole proprietorships, partnerships, and corporations, there are several other types of business organizations, some of which are more interesting than others. An *S Corporation* is a cross between a partnership and a corporation. If the participants in a business enterprise decide to form an S Corporation rather than a partnership they will receive the

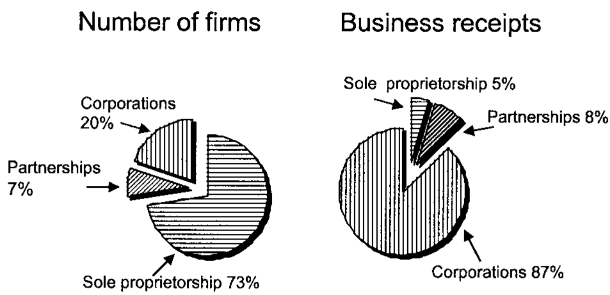


Fig. 5.1. Forms of business organization
The pie-chart on the left shows that sole proprietorships outnumber all other forms of business organization. The pie-chart on the right shows that the vast bulk of the nation's business is done by corporations.

Table 5.1. Relative importance of different forms of business organization.

	Number (1,000)	Receipts (Billions $)	Receipts ($)/Firm
Proprietorship (nonfarm)	17,409	918	52,731
Partnership	1,855	1,534	826,954
Corporation	4,849	16,543	3,411,631

Source: Data for 1998 from *Statistical abstract of the United States*, 2001, Table 711.

benefit of limited liability, but they will not have to worry about the federal corporate profits tax. Instead, all the profits of the corporation will be "passed through" to count as current income of the shareholders. A major limitation of S Corporations is that there can be no more than 75 stockholders. In many states Limited Liability Companies (LLC) provide still another form of organization. In some states Chapter S corporations are subject to the state corporate profit tax but an LLC is exempt.

5.2.2 *How the modern corporation works*

The owners of a corporation are its shareholders, but what are shares? A corporation can raise funds by issuing shares of stock representing an ownership interest. For example, in year 2000 Microsoft had 5,250 million shares of stock outstanding. You could have purchased a share of Microsoft in 1990 for $21. In 1999 they had sold for as much as $119.93 per share, but after the District Court Judge ruled that Microsoft was guilty of serious violations of the antitrust laws, the stock declined dramatically in value to a low of $60.37.

You may have purchased shares in a corporation as an investment, but the act of purchasing also means that you are a part owner of the company; you have become a corporate *shareholder*. As part owner, you are entitled to vote for the corporation's *board of directors*. The Board in turn appoints the managers of the firm, oversees operations, and tries to insure that the company is run for the benefit of the stockholders. In corporate democracy it is not "one person one vote." The number of votes each shareholder may cast equals the number of shares that she owns. When the ballot arrives in the mail you may find that there is only one set of names on the ballot, those recommended by the incumbent directors. Critics complain that corporate democracy is like having only one political party — it is hard for dissatisfied stockholders to organize in opposition to the incumbent board of directors.

Microsoft is a publicly held corporation. Its shares are tradable on the NASDAQ stock exchange — that is, any individual with the money and the courage to invest may buy shares in the corporation. You can contact a stockbroker or purchase shares on the Net at the going market price. Small corporations are privately held. For example, if a construction firm partnership decides to incorporate in order to obtain the protection of limited liability, the owners may divide the shares up among themselves, but they will not be sold on a stock exchange.

Corporations raise funds to finance their operations by borrowing as well as by issuing stock. The corporation may take out a loan in order to meet short term financial needs or it may seek a mortgage to pay for a major investment project, just as a home buyer will seek a mortgage to finance the purchase of that dream house. Corporations also borrow funds by issuing bonds. A *bond* is a legally enforceable obligation to pay back the loan on maturity and to make interest payments periodically in the interim. For example, a $10,000 bond maturing in 2015 paying 6% is a promise to pay the owner of the bond $10,000 when the bond matures in 2015 plus $600 each year until then. Bonds of major corporations are actively traded in the marketplace and their current prices, determined by supply and demand, are quoted daily in the *Wall Street Journal* and other financial publications. This means that if you inherit a $10,000 bond from your late uncle you will not have to wait until it matures before you can get your money; you can sell it to someone else who is willing to hold the bond until maturity. A broker will be happy to handle the transaction for you for a fee. You will find that the going price of the bond today may be substantially more or substantially less than the $10,000 face value that it will pay on maturity, depending on market conditions.

Owners versus creditors

There is a fundamental distinction between the firm's owners and its creditors. *Debtors* are private individuals or organizations that have borrowed money. A student using a credit card and a homeowner paying off a mortgage on their home are debtors. Corporations are debtors because of the funds they owe to bankers and bondholders. The *creditors* are those who have loaned the money. The banker, the insurance company or a private individual buying the corporation's bonds or lending it funds receives a commitment from the borrower to pay off the loan at a specific date. The borrower also has a contractual obligation to make regular interest payments on the loan. The firm's owners — its stockholders — have no such guarantee even though they also have advanced funds to help finance the firm's investments and operations. If the firm prospers, the owners will be rewarded with a share of the firm's profits, but if the firm loses money the owners are the first to shoulder the loss. The lenders must be paid on time before the owners get anything. Creditors lose only if business is so bad that there is not enough to fully pay off all the outstanding loans.

When a firm cannot meet its commitments to its creditors, *bankruptcy* may follow. Bankruptcy may be declared by the firm itself or initiated by disgruntled creditors. Because the United States Constitution grants Congress the authority to "establish ... uniform laws on the subject of Bankruptcy throughout the United States," bankruptcy proceedings are supervised by and litigated in the United States Bankruptcy Courts rather than in state courts. There are two types of bankruptcy: If the business enterprise is liquidated, a court appointed trustee will sell off the assets of the firm and distribute the proceeds to the creditors in proportion to the amount they are owned — this might be 90¢ or 10¢ for each dollar of debt, depending on how much is realized from the sale. Alternatively, the courts may allow the debtor firm to continue in business in an effort to rehabilitate itself and, it is hoped, eventually be able to pay off its creditors.

Here are two bankruptcies that undermined investor confidence and traumatized the stock market in year 2002:

1. Enron Corporation: Late in 2001, shortly after admitting major accounting errors that inflated earnings by almost $600 million since 1994, this giant Texas-based energy trading company filed for bankruptcy. With $62.8 billion in assets, it became the largest bankruptcy case in U.S. history, up to that time. The price of Enron stock tanked, many employees lost their pensions and life savings. Arthur Andersen, one of the very largest accounting firms in the United States, was convicted of obstructing justice by shredding Enron documents and forced out of business.
2. WorldCom, the telecom giant, filed for bankruptcy in July of 2002 after reporting that its profits had been overstated by $3.8 billion in bogus accounting, and assumed from Enron the dubious distinction of suffering the largest bankruptcy in the nation's history. Subsequent investigation revealed that there had been more than $7 billion in bogus accounting. The company's former CEO and Chairman Bernard Ebbers received a low interest rate loan of $408 million dollars and $1.5 million-a-year for life in severance pay. The bankrupt firm took legal proceedings to rescind Ebbers' severance package, arguing that the severance agreement was negotiated under fraudulent terms. The company had laid off 17,000 workers earlier in the year.

Corporate Governance

The collapse of stock markets in 2001–2002 plus revelations about gross accounting improprieties at a sizable number of major corporations raised serious concern about the structure by which corporations run their affairs and the proper role for government in their regulation. Do the shareholders who own the firm have effective control? Does the company's Chief Executive Officer (CEO) run the company in the interest of the owners or to maximize personal gain? These were not new questions. In 1932, in the aftermath of the 1929 stock market crash, Adolf Berle and Gardiner Means focused in *The Modern Corporation and Private Property* on the "separation of ownership from control." Seventy years later, Alan Greenspan, the powerful Chairman of the Federal Reserve Board (the central bank of the United States), who had himself served on the boards of 15 major corporations, commented as follows on the problem:[5]

> ... as our economy has grown, and our business units have become ever larger, de facto shareholder control has diminished: Ownership has become more dispersed and few shareholders have sufficient stakes to individually influence the choice of boards of directors or chief executive officers. The vast majority of corporate share ownership is for investment, not to achieve operating control of a company. Thus, it has increasingly fallen to corporate officers, especially the chief executive officer, to guide the business, hopefully in what he or she perceives to be in the best interests of shareholders. Indeed, the boards of directors appointed by shareholders are in the overwhelming majority of cases chosen from the slate proposed by the CEO. The CEO sets the business strategy of the organization and strongly influences the choice of the accounting practices that measure the ongoing degree of success or failure of that strategy. Outside auditors are generally chosen by the CEO or by an audit committee of CEO-chosen directors. Shareholders usually perfunctorily affirm such choices.
>
> To be sure, a CEO can maintain control over corporate governance only so long as companies are not demonstrably in difficulty. When companies do run into trouble, the carte blanche granted CEOs by shareholders is withdrawn. Existing shareholders, or successful hostile

[5]Remarks by Chairman Alan Greenspan delivered at the Stern School of Business, New York University, New York, New York, March 26, 2002: http://www.stern.nyu.edu/greenspan/alan_greenspan_remarks_0326.pdf.

bidders for the corporation, usually then displace the board of directors and the CEO. Such changes in corporate leadership have been relatively rare but, more often than not, have contributed to a more-effective allocation of corporate capital.

Reform

Public outrage at the growing list of corporate scandals coupled with the plunge of the stock market led to the passage of the Sarbanes-Oxley Act in July of 2002. The reform legislation provides for the establishment of the Public Company Accounting Oversight Board to watch over the auditing of corporations. It requires that each corporation's Chief Executive Officer (CEO) and Chief Financial Officer (CFO) certify to the validity of their firm's financial statements. Each corporation must have an audit committee composed entirely of directors of the firm who are independent in the sense that they do not accept consulting advisory or other compensatory fees from the company apart from their capacity as members of the firm's board of directors. The act also increased the penalties for violating the security laws and prohibited the destruction or falsifying of records. It banned corporations from making personal loans to their executive officers and directors.

While many critics objected that the law did not go far enough, some complained that it might inhibit entrepreneurial innovation and place American enterprises at a disadvantage in competing with foreign firms. The continuing decline in the stock market in the months immediately following the passage of the act implied that the reform legislation failed to restore investor confidence in the integrity of the system.

5.3 Profits and accounting

Accounting records report the financial position of the business enterprise, providing management with information about the business that is essential for making wise decisions. Accountants also generate reports for use by those outside of the firm, such as the financial reports for shareholders and potential investors, including the tables that appear in the corporation's annual report. Before banks will make a loan, they appraise the creditworthiness of the borrower and carefully scrutinize the accounting records of the business in order to learn how much profit the firm has been making and how much the business is worth today. And of course, the Internal

Revenue Service requires corporations to file tax returns reporting their profits and related information.

The Securities and Exchange Commission was created in 1934 to establish financial accounting and reporting standards for publicly held corporations. Since 1973 the responsibility for establishing financial accounting standards has been delegated to the Financial Accounting Standards Board, a private organization. This organization promulgates the *Generally Accepted Accounting Principles* (GAAP),which are officially recognized as authoritative by both the Securities and Exchange Commission and the American Institute of Certified Public Accountants. Nonetheless, GAAP leave considerable room for the firm's accountants to exercise discretion in preparing the books. In deciding such arcane issues as how to measure depreciation and what inventory accounting procedures to employ, the accountant has a certain amount of leeway to manage the profit picture that will be reported on the firm's books. Accountants may be encouraged by management to adopt strategies that puff reported profits in order that the firm's stock will catch the eyes of investors, push up the value of the stock, and possibly lead to a bonus for the corporation's president. For tax returns the objective is just the opposite. The accountant may adopt tax-reduction strategies that will minimize profits that have to be reported for tax purposes.

5.3.1 *The balance sheet*

A firm's *balance sheet* provides a picture of the financial posture of the enterprise at a particular date. There are three major items on the balance sheet:

- *Assets*: An asset is something of value owned by the firm, such as its holdings of land, buildings and machinery. Inventories of raw material waiting to be processed and finished goods that are unsold or awaiting shipment to customers are also counted among the assets of the firm. Cash in the safe and funds on deposit in the bank are also assets. So are patents and copyrights. If the firm has accounts receivable from its customers, they are listed among the firm's assets.
- *Liabilities*: In accounting, liabilities are what the enterprise owes to its creditors. Included are funds owed to the banks, the value of bonds outstanding, accounts payable to suppliers, and taxes owed to the Internal Revenue Service.

- *New Worth*: The net worth of an enterprise is the accountant's estimate of the value of the firm to its owners. It is the excess of the assets of the firm over its liabilities.

The fundamental equation of accounting states the relationship between these three concepts:

$$\text{Net Worth} = \text{Assets} - \text{Liabilities} \tag{1}$$

As this equation reveals, Net Worth is a residual. It is what is left over for the owners of the firm after the claims of all the creditors are met. Note that the firm's creditors are those who have lent funds (i.e., granted credit) to the firm. Stockholders are not creditors; they are the owners of the corporation, sharing in the profits or losses and, at least in principle, exercising control over the enterprise.

Now let us consider the simplified balance sheet for a hypothetical firm displayed on Table 5.2. First of all, note that the Assets of the firm are listed on the left side of the balance sheet and its Liabilities and Net Worth are on the right. This is in conformity with the following modification of equation (1):

$$\text{Assets} = \text{Liabilities} + \text{Net Worth} \tag{2}$$

Thus the sums of the two sides of the table must add to the same number, and indeed they do, $1,550,000!

The accountant's treatment of plant and equipment requires explanation. The $800,000 listing for plant and equipment is the amount that the firm paid for the various items on the dates when it purchased them; it is their *historic cost* rather than replacement cost. Because the plant and

Table 5.2. Balance sheet of the Fly-by-Nite Aircraft Company.

December 31, 2000 (All figures in $1,000)					
Assets			*Liabilities*		
Cash		150	Accounts payable	50	
Inventory		500	Bonds	350	
Accounts receivable		300	Total liabilities		400
Plant and equipment	800		Net worth		
less depreciation	(200)		Common stock	1,000	
		600	Retained earnings	150	
			Total net worth		1,150
Total Assets		$1,550	*Liabilities + net worth*		$1,550

equipment are far from new; they are not as valuable as they were when purchased. The *depreciation* item of \$200,000 is the accountant's estimate of an appropriate allowance for the wear and tear on the plant and equipment.[6] The number is placed in parentheses, which is the accountant's way of indicating that it is to be subtracted. Subtracting the depreciation from the original purchase price yields the accountant's estimate of the current value of the plant and equipment. The final Plant and Equipment figure of \$600,000 is the value of factory buildings and equipment at historic cost adjusted for wear and tear. But the figures have not been adjusted for inflation, which means that they may grossly understate the cost of replacement.

Liabilities appear as the first heading on the right hand side of the balance sheet. As of December 31, 2000, our firm owed \$400,000 to its suppliers and bond holders.

The Net Worth item on the right side of the balance sheet requires further explanation. Net Worth is a *residual*. It is not apparent from the balance sheet, but as equation (1) made clear, net worth is obtained by subtracting Liabilities from Assets — this gives us the Total Net Worth entry of \$1,150,000. The *retained earnings* entry is the sum of profits that, over the years, have been plowed back into the firm rather than distributed to the owners as dividends. The entry for Common Stock is the total net worth entry minus the retained earnings.

The balance sheet is a statement about the firm's financial position at a particular point of time, in this case December 31, 2000. The balance sheet does not tell us how much in the way of profit the firm has made during the year or how much it has been paying out in dividends. Profits and dividend payments are examples of flows that take place over a period of time. Just as a snapshot cannot accurately reveal the speed of a cross-country runner, so too, a balance sheet cannot reveal profits. A second type of accounting record, the income statement (aka a profit and loss statement) reports on what happened to the firm over a period of time, just as a video camera may vividly reveal the speed of our cross-country runner.

[6] How depreciation is calculated is a complicated topic that will deserve further discussion a bit later in this chapter.

Table 5.3. Income statement of the Fly-by-Nite Aircraft Company.

January 1, 2001 to December 31, 2001 (all figures in $1,000)							
Net sales				800			
Less Cost of manufacturing							
Materials	200						
Labor	250						
Depreciation expense	100						
		550					
Less inventory increase		(100)					
Cost of goods sold			(450)				
Gross margin				350			
Less selling cost			30				
Interest expense			20				
				(50)			
Profits					300		
Less corporate profit tax					(105)		
Profits after taxes						195	
Less dividends						(100)	
Addition to retained earnings							95

5.3.2 *The income statement*

Suppose that during year 2001 the following transactions are recorded on the books of the Fly-by-Nite Aircraft Company (in $1,000): Materials Purchased $200; Labor costs $250; Sales $800; Selling Cost $30; Dividends $100; Depreciation Expense, $100; Bond interest, $20; Corporate Profits Tax, $105. Further, as of December 31, 2001, Accounts Payable were $100, Accounts Receivable $200, Inventories $600. How can we determine the firm's profits from this information? Table 5.3, the income statement, lays out the information in a way that makes it easy to see what has been going on. Here are some key points to note about the income statement:

- The income statement summarizes flows that took place during year 2001. We shall find that these flows will explain why entries in the end of year 2001 balance sheet differs from the year 2000 balance sheet.
- Observe that inventories have increased during the year. That is to say, part of the expenditures of the firm on purchased materials and labor costs were incurred in accumulating stocks of purchased materials and/or finished product that will be available for future use. These expenses should not be counted as part of cost of goods sold during the current year.

- Interest costs (which are income to the firm's creditors) are an expense that can be subtracted before computing the firm's profits, which are subject to the corporate profit tax.
- The corporation pays dividends from after-tax profits, but only part of after-tax profits are distributed as dividends to the owners of the firm.
- The addition to retained earnings of $95 constitutes funds kept back by the firm for future expansion rather than distributed to the stockholders. Stockholders may expect eventual rewards whenever they sell their stock because it should sell for more, thanks to the investment in plant and equipment financed from retained earnings.

5.3.3 *Taxing complications*

Capital gains

Suppose you had purchased 100 shares of Fly-by-Nite Aircraft for $10 a share on March 10, 1996. By December 31, 2000, Fly-by-Nite stock was selling for $15 a share. Thus the value of your holdings had increased from $1,000 to $1,500. This increase in the value of the shares is known as a capital gain, but because you have not sold the stock it is an *unrealized capital gain*. Unrealized capital gains are not taxable. If you sold the stock on July 1 for, say $14 a share you will have *realized* a capital gain. The excess of what you sell your stock for over what you paid for it, $1,400 − $1000 = $400 is your *realized capital gain*, which is taxable. You will be pleased to find when you complete your income tax return the following April that yours is a long-term capital gain because you had owned the asset for more than a year. This means that your capital gain will be taxed at a maximum rate of 15% (or 5% if your income is so low that you are in the 15% income tax bracket). If you had sold the asset for less than you paid for it, you would have realized a capital loss, which would lead to a reduction in your taxes for that year.

Why should investors pay a lower tax rate on realized capital gains than a worker pays on earned income? This hardly seems fair, but there are two arguments for the favorable treatment of capital gains. First of all, the realized capital gain may be partially offset by inflation. Suppose that during the period you owned the stock the consumer price index had increased by 10%. Then the $1,400 you receive when you sell your asset is worth only as much purchasing power as $1,400/1.1 = $1,272.72 when you purchased the asset, which means that your capital gain net of inflation, i.e., your *real capital gain*, is only \$272, but you will be taxed on the \$400 nominal capital gain. The favorable treatment of capital gains provides an imprecise adjustment for inflation. The second argument for favorable taxation of capital gains is that the taxation of realized capital gains has a *lock-in effect*. Faced with the prospect of a substantial capital gains tax if they sell their assets, investors may decide to hang on to stocks that they would otherwise sell. Taxing capital gains at a reduced rate is one way of limiting the strength of the lock-in effect.

Step-up-of-basis-at-death

Those who invoke the "lock in effect" to justify the favorable tax treatment of capital gains do not always champion the elimination of another type of "lock in effect." If you hold your stock until you die, both you and your heirs will avoid the capital gains tax! The "step-up-of-basis on death" provision of the internal revenue code allows an heir selling an inherited asset to compute the realized capital gain on the basis of the increase in the value of the asset from when it was inherited rather than its price when initially purchased by the deceased. This means that no tax is ever paid on the gain in value that took place during the period in which the asset was owned by the deceased.

Double taxation

Stock market investors often complain that they are unfairly subjected to double taxation. First, the corporation in which they have purchased stock has to pay the corporate profits tax. Second, they have to pay income tax on the dividends they receive from the corporation. Thus, their investment earnings are taxed twice. Workers may sympathize with the stockholders' lament because they have to pay both the income tax and the Social Security tax on their wage income.

The actual burden of *double taxation* of corporate profits is not quite as severe as it appears for two reasons. First of all, only that part of corporate profits that is paid out as dividends is taxable income for the investor. Corporate retained earnings are not immediately taxable. The hope is that the shares will increase in value because the funds are being put to good use by the corporation. Thus, the personal income tax on the retained portion of corporate profits is postponed until the capital gains are realized when the stock is sold. And when the return is finally taxed it is at a much lower rate than wage income. Second, endowment investments of universities, hospitals and other tax exempt institutions escape double taxation entirely.

5.3.4 *Creative accounting*

There is a fundamental difference between tax avoidance and tax evasion. Tax evasion is fraudulent. For example, counting the dog along with the children as an exemption in computing the Personal Income Tax is tax evasion, even if the dog is treated like family. A waiter or waitress who neglects to report tips is also guilty of tax evasion. A plumber who gives customers a discount for paying by cash rather than by check is probably not going to report the payment on his income tax return. Stu Leonard, the founder of an extraordinarily successful supermarket in Stamford, Connecticut, served time in jail for fraudulently underreporting the revenues of his enterprise in order to understate his true profits to the tax collectors!

Tax avoidance is entirely different from tax evasion. Tax avoidance is legal! Tax avoidance is organizing your activities so as to maximize your after tax income. Avoid paying taxes where you can; and if you cannot avoid paying a tax, at least postpone the payment into the future. Postponing a tax payment lets you put the funds to use in the meantime at zero cost — it's like an interest free loan from the government.

To the extent that corporations retain earnings rather than pay out profits as dividends, they help their stockholders avoid or postpone paying income taxes. Some corporations pay zero dividends, which provides a tax benefit to shareholders as well as making capital available to the firm. Microsoft did not start paying dividends until 2003.

Accelerated depreciation

The firm has been using straight-line depreciation in evaluating its durable capital assets, such as machinery and equipment, trucks and buildings and other items with a useful life of more than a year. Our firm had paid \$90,000 for a machine that has a useful life of seven years; at the end of that time it will have a scrap value of \$20,000. On average then, the annual cost of using the machine is $\$10{,}000 = (\$90{,}000 - \$20{,}000)/7$. With *straight-line depreciation*, precisely this amount is deducted as depreciation expense each year over the machine's life-span. This means that the net value of the machine reported on the balance sheet will decline in a succession

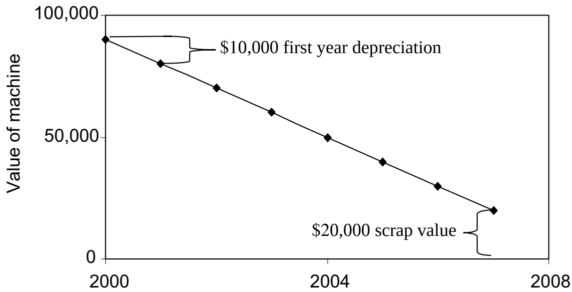


Fig. 5.2. Straight-line depreciation
A new machine costing \$90,000 will have a scrap value of \$20,000 after seven years. If the accountant uses "straight-line depreciation," the value of the machine will be depreciated (written down) by \$10,000 = (\$90,000 − \$20,000)/7 each year. The graph shows the value of the machine on the firm's books over each year of its remaining lifespan. Depreciation expense is the change in the book value of the machine during the year. For this machine, \$10,000 of depreciation expense is charged each year in computing the firm's profit.

of \$10,000 depreciation steps until the \$20,000 scrap value is reached at the end of the seven years, as indicated by the solid line on Figure 5.2.[7]

Over the years, the United States tax code has been gradually rewritten to allow more and more flexibility to accountants. The life of capital equipment for tax purpose has been cut repeatedly, reaching the point where the accounting lifespan is substantially shorter then the useful economic life of the equipment — this is *accelerated depreciation.* For example, the United States tax code allows trucks and cars to be depreciated over a five year period, books and furniture over seven years and residential housing over 27.5 years.

To see how accelerated depreciation affects our firm's income statement, balance sheet, and tax liability, suppose the tax authorities allow our firm to depreciate the machine over only five years. With the life of the machine shortened for tax purposes our firm will have $\$14,000 =$ (\$90,000 – \$20,000)/5 of depreciation expense to deduct each year, which means that accounting profits will be \$4,000 lower than was reported on

[7] Straight-line depreciation may be too conservative in that it implies that the machine is equally useful to the firm each year when in fact an aging machine may be more prone to breakdowns and suffer from obsolescence. Various alternatives to straight line depreciation, such as "double declining balance," allow accountants to take a larger depreciation deduction in the early years of asset ownership. While this must be offset by smaller deductions in later years, it is advantageous because it postpones the tax burden into the future, which is like an interest free loan from the IRS.

Table 5.3. Thus our corporation's profit will be artificially reduced by \$4,000, and if the corporate tax rate is 35%, after tax profit will be down by \$1,400. Furthermore, the firm's balance sheet will now show lower total assets because a lower depreciated value is assigned to plant and equipment. But our firm has an extra \$1,400 of cash because of the tax saving. The firm looks poorer, thanks to our creative accounting; but this is deceptive. Because the profits reported to the tax authorities are reduced, the corporate profit tax has been cut and the firm is actually better off!

The balance sheet will show a lower book value for the machine because the depreciation will be twice as high. After five years have passed, the machine will be "fully depreciated"; i.e., the value of the machine for tax purposes (the "basis") will be zero. From then on out, no more depreciation expense can be taken; therefore, before-tax profits will be \$10,000 higher, and taxes will be \$3,500 higher each year. This means that over the entire useful life of the machine, tax payments have been the same as before. But the tax payments have been delayed, which provides the firm with an opportunity to put the funds to work; if it cannot think of a better use it can earn interest by putting the funds in the bank.

Conversely, management may decide to puff profits by lengthening the depreciation period, perhaps in order to justify a bonus for key executives or to impress the stock market. It is said that in the 1990s major airlines gave multimillion-dollar lifts to their profits by claiming a longer useful life for their aircraft. There are legal limits to such manipulations. In 2002 key executives of WorldCom Corporation took the 5th Amendment when called before the House Financial Services Committee of the House of Representatives to respond to allegations that they had recorded as capital expenditures some \$3.8 billion dollars of current expenses, such as maintenance, salary, and interest payments. This meant that only the depreciation on the \$3.8 billion of bogus capital expenses had to be subtracted in calculating profits during the current year, which inflated profits and deceived the stock market.

5.3.5 *Accounting versus economic profit*

The profit figure appearing on the firm's income statement is calculated in accordance with Generally Accepted Accounting Principles (GAAP). But that does not mean that the accountants do not have a fair amount of discretion in determining what that profit number will be.

Economic profit is the excess of the firm's revenue over the costs of *all* the resources used by the firm in producing those goods. The resources involve more than the cost of materials, salary and wage payments, and the interest paid on borrowed funds. The economic costs also include the opportunity cost of using two other types of resources.

1. The owners of a small enterprise may pay themselves a salary or wage that is far less than the value of the countless hours they devote in an effort to make the firm thrive. They choose to take less than the value of their services out of the firm as salary in order to plow funds back into their enterprise.
2. The funds invested by the owners would earn interest if deposited in the bank or used to purchase government bonds. This foregone interest is the opportunity cost of the funds invested by the owners in the business and should be regarded as a cost of doing business rather than true economic profit. Profit is a reward for taking risk rather than compensation for providing funds to the firm.

Subtracting these two types of costs from accounting profits gives a good approximation of economic profit. As a further refinement, the economist would use an unbiased estimate of depreciation rather than the figure reported by accountants to the Internal Revenue Service.

5.4 The technological constraint

Accounting records report on the firm's current financial position. They provide information about what the firm has done. They do not fully explain what options or choices were available to the firm, why the firm made the decisions it did, or how it would respond in the future to a change in circumstances, such as an increase in demand or a decrease in tax rates. Our next task is to explain what determines the behavior of a representative business firm. We shall be building a simple model relying on the fundamental assumption that the primary motivating force is the desire to maximize profit. The model will provide a powerful engine for analyzing such questions as how taxes affect the behavior of the firm and for studying, in the next chapter, different types of market structure. The first step in constructing the model will be to specify the technological options available to the representative business enterprise.

5.4.1 *Total product curve: One variable input*

The first step in analyzing the behavior of the firm is to look at the available technological options. The simplest possible case involves a single variable input, say hours of labor, L, used to produce one type of output, q. For example, we might have $q(L) = \beta L$ or $q(L) = L^{\lambda}$, $\lambda > 0$. Or we might have the slightly more complicated total product curve plotted on Figure 5.3. Here are two key concepts:

- *Average product of labor* $\sim q(L)/L$. The output to labor ratio is sometimes referred to as "labor productivity" or as just plain "productivity." The remarkable increases over the last century in the average product of labor were discussed in Chapter 1.
- *Marginal product of labor* $\sim dq/dL$. The marginal product of labor is the extra output that will be produced when an additional unit of labor is employed, all other inputs remaining unchanged.

Example #1

If $q(L) = \beta L$, then the average product of labor, $q(L)/L$, equals the marginal product of labor; i.e., $dq/dL = q/L = \beta$.

Example #2

If $q(L) = L^{1/2}$, then the average product is $q/L = L^{-1/2}$ and the marginal product is $dq/dL = (1/2)L^{-1/2} = (1/2)q/L$.

In the second example, the marginal product of labor gradually declines as more and more workers are employed. This may well make sense because when more and more workers are hired to work in a factory of a given size, they will have to queue up to share equipment and they may gradually get more and more in each other's way — the reduction in efficiency means that the average product of labor gradually declines as L increases, given that there are no changes in the other inputs. As another application, if more and more farm laborers work the soil of a farm of given size, total output may increase in response to more intensive farming, but at a slower and slower rate because of crowding. That is why output per farm worker may be expected to decline.

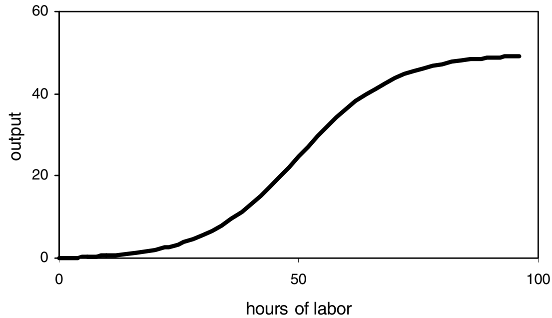


Fig. 5.3. The total product curve
The total product curve shows how the total output of the firm depends on the number of hours of labor applied to the production process, holding fixed the number of machines and other equipment.

This total product curve displays first increasing and then diminishing marginal productivity: An additional worker adds substantially to total output when there are few workers relative to the number of machines; but beyond a certain level, hiring additional workers does not yield much of an increase in output because they may have to share the fixed number of machines.

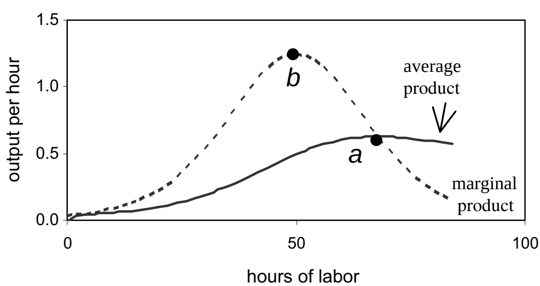


Fig. 5.4. Marginal versus average product
The marginal product of labor is $dq(L)/dL$; the average product is $q(L)/L$.

Example #3

The average and marginal products of labor corresponding to the total product curve that was plotted on Figure 5.3 are re-plotted on Figure 5.4. Observe that average and marginal product both increase initially but then they taper off and eventually decline. Note too that the marginal product curve crosses the average product curve at the point where average product is at a maximum.

The relationship between the labor input and output on Figure 5.4 deserves careful explanation. When there are only a few workers in the factory, output per worker hour may be low because each worker has to do a large number of different tasks and as a result does few of them well. Or certain machinery requires many workers to make it operate properly. When more workers are employed at the factory, each can specialize on one or two tasks. Average product increases because the workers gain more and more skill when they are not trying to function as jack-of-all-trades. But after a point the advantages of specialization are outweighed by the crowding effect and first the marginal and then the average product decline. Economists customarily regard the more complex total product curve generated by this type of behavior as a better representation of reality than either Example #1 or #2.

Geometry of marginal and average product

The relationship between the marginal and average product concepts is clarified by looking at Figure 5.5, which reproduces the total product curve from Figure 5.3. Now consider the ray connecting the origin with point m. The slope of this line is $q(50)/50$ or the average product of labor when $L = 50$. Again, the slope of the radius vector through point a is $q(68)/68$, which is the average product of labor when $L = 68$. Observe that the line linking the origin with point a is the steepest ray emanating from the origin that just touches the total product curve. That is why the average product attains a maximum at this output level, as illustrated on Figure 5.4. Now the marginal product is the derivative of total product; i.e., the marginal product at point m is the slope of the tangent line going through point m and the marginal product at point a is the slope of the tangent line at that point. Since m is also the point of inflection, the marginal product is at a maximum at that level of output, as was indicated on Figure 5.4. Also, the line segment from the origin through point a of Figure 5.5 is tangent to the total product curve, which means that marginal and average product of labor are equal at this level of employment.

The following proposition concerns a fundamental link between marginal and average product for any production function $q(L)$.

Fundamental proposition

The marginal product must equal average product at that level of output at which average product is at its maximum.

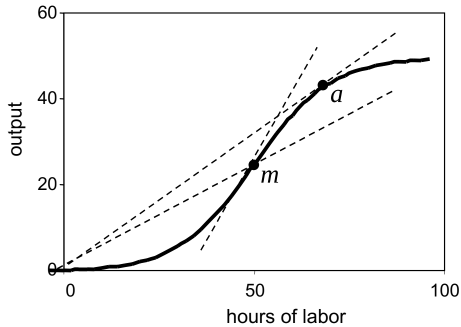


Fig. 5.5. Total versus marginal and average product of labor
The slope (rise over run) of the line segment connecting point m with the origin is output divided by hours of labor, which is the *average product* of labor. The steeper line through m is tangent to the total product curve — its slope is the *marginal product* of labor at point m. Since point m is at the inflection point, the marginal product of labor attains its maximum value at this point.

The line through point a is tangent to the total product curve and it also goes through the origin. Therefore, its slope represents both the marginal product and the average product of labor at point a. Also, this line is the steepest line segment emanating from the origin that touches the total product curve, which means that its slope is the maximum attainable value of the average product of labor. This suggests that marginal product must equal average product at that level of output where average product is at a maximum.

To prove this proposition, note that if average product is maximized we must have

$$\frac{d(q/L)}{dL} = \frac{d(q)/dL}{L} - \frac{q}{L^2} = 0\,. \tag{3}$$

On multiplying through by L we have

$$L\frac{d(q/L)}{dL} = \frac{dq}{dL} - \frac{q}{L} = 0\,,$$

or

$$\frac{dq}{dL} = \frac{q}{L}\,; \tag{4}$$

i.e., the marginal product equals average product at that level of output at which average product attains its maximum value.

5.4.2 *The law of diminishing returns*

The Law of Diminishing Returns is the proposition that the additional output from successive increases of one input will eventually diminish, provided all other inputs are held constant. This "law," stressed as early as 1817 by David Ricardo, has been regarded as of fundamental importance by generations of economists. Point a on Figure 5.4 and Figure 5.5 marks the point where additional labor starts to reduce the average product of labor. Point m marks the *point of diminishing marginal productivity.*

Two exceptions to the law of diminishing returns

While the proposition may initially appear obvious, here are two exceptions that deserve consideration.

1. *Pipeline*: When building an irrigation pipeline, a major cost factor is the steel for the pipe. The number of square feet of steel required per running foot of pipeline obviously depends on the circumference of the pipe: $C = 2\pi r$, where r is the radius of the pipe. The capacity of the pipe in gallons per hour, given the pumping pressure, will be proportional to the cross-section area of the pipe, which is a circle with area $A = \pi r^2$. Thus the cost/capacity ratio will be proportional to $A/C = 2/r$.[8] This ratio, and hence the steel required per unit of output declines with output, violating the Law of Diminishing Returns.
2. *Networks*: Telephone, faxes and the Internet are all examples of *networks*, and they all violate the law of diminishing returns. Their common feature is that the value of the network to any user depends positively on the number of users. If you were the only subscriber to the local telephone company, the service would be of no value because there would be no one to talk to. With two telephone subscribers you can call one person, with three two, and so forth. If x is the average value to you of being able to contact someone, then the value to you of the telephone is $(n-1)x$, where n is the number of subscribers (including you). The value of a network to the n users is $v(n) = n(n-1)x = (n^2 - n)x$. This demonstrates "Metcalfe's law," the proposition that the value of a network is proportional to the square of the number of connected clients.[9]

[8] For a more elaborate discussion, see Hollis B. Chenery, "Engineering Production Functions," *Quarterly Journal of Economics*, 1949.

[9] Robert Metcalfe founded 3Com Corporation and designed the Ethernet protocol for computer networks.

> So too with Email, the fax machine and any other innovation where the value to each user of the device is proportional to the number of other users one can contact.

The concept of networking explains why many innovations are slow to catch on. It does not pay to join the network when there are few users. But once the number of users reaches a critical mass, the rewards clearly outweigh the cost, and more and more users rush to take advantage of the innovation.

5.4.3 *Multiple inputs and the production function*

A natural extension of the total product curve is to include more than one input, which brings us to the concept of the production function. The *production function* is like a recipe book in that it shows the level of output that can be obtained by applying various combinations of labor, machinery and equipment, raw materials and other inputs.

While in reality the typical firm uses a wide range of resources in producing a number of different products, it is appropriate to narrow our attention to the essential features of the problem by supposing that our hypothetical firm produces one type of product utilizing only two inputs, labor and capital.[10] Here we mean by *capital* the machinery, buildings, equipment and other durable items that are available for use in the production process.[11] In reality there are many kinds of labor and types of capital employed by a business enterprise, but we shall neglect these complications and assume that we can talk meaningfully about the quantity of labor and capital services that are employed by the firm. Thus we shall write our production function as

$$q = q(L, K)\,. \tag{5}$$

The production function is analogous to the utility function we used in building the theory of the consumer in Chapter 4. But a major difference is that we can measure the output the firm produces with a given set of resources while in Chapter 4 it was necessary to concede that we cannot directly measure the utility realized by the consumer.

---

[10] Students who have studied multivariate calculus will find that the analysis that follows may be easily extended from 2 to any finite number of inputs.

[11] The term "capital" is used in two senses. Here we are measuring the stock of buildings and durable physical equipment. But the term capital is used in both accounting and finance to refer to the total funds that have been contributed by the owners of the firm to finance their business enterprise.

A simple example is provided by the Cobb-Douglas production function:

$$q = q(L,K) = \alpha L^{\lambda} K^{\lambda'}\,, \quad L \geq 0\,,\; k \geq 0\,,\; \lambda \geq 0\,, \lambda' \geq 0\,. \tag{6}$$

Note that in the Cobb-Douglas case $q(0,K) = q(L,0) = 0$; i.e., both labor and capital are required if anything is to be produced.

This production function is named after economist Paul H. Douglas [1892–1976] and C. W. Cobb, his math department colleague at Amherst College.[12] Utilizing annual data on industrial production in the United States covering the years 1899–1922, they estimated:[13]

$$q(L,K) = 0.0156 L^{0.807} K^{0.232}\,; \tag{7}$$

i.e., $\alpha = 0.0156$, $\lambda = 0.807$ and $\lambda' = 0.232$. This empirical investigation established Douglas's reputation as one of the leading economists of the day.[14]

The iso-product curves on Figure 5.6 provide an informative contour plot of the production function, which is analogous to the indifference maps of Chapter 4. The quantity of labor is plotted on the abscissa and the quantity of capital on the ordinate. The parameter values are $\alpha = 2$, $\lambda = 2/3$ and $\lambda' = 1/3$. Each *iso-product curve* (aka "isoquant") reveals the various combinations of the two inputs that would just suffice to produce the specified level of output. Points a, b, c and d on the light horizontal line at $K = 10$ indicate how the level of output would respond to increasing quantities of labor, keeping the stock of capital fixed at $K = 10$. These same three points are also plotted on the total product curve for $K = 10$ on Figure 5.7, where the equation for the total product curve is $q(L,10) = 2L^{2/3}10^{1/3}$.

For the Cobb-Douglas production function we have as the average product of labor

$$\frac{q}{L} = \alpha L^{\lambda-1} K^{\lambda'}\,. \tag{8}$$

[12]As often happens in the history of science, the Cobb-Douglas function is misnamed. Although the theoretical properties of the production function they used had already been studied by Philip Wicksteed [1844–1927], this function is universally known as the Cobb-Douglas.

[13]Taking logs of both sides of equation (6), they calculated without the aid of computer the values of the unknown parameters λ and λ' and α that would minimize $\sum e^2$ in the equation $\log q \;= \alpha + \lambda \log L \;+ \lambda' \log K \;+ e$. This simple estimation procedure is known as the "method of least squares."

[14]Douglas taught at the University of Chicago as well as Amherst. After service in the Marines during World War II, he was elected to three terms in the United States Senate, where he chaired the Joint Economic Committee.

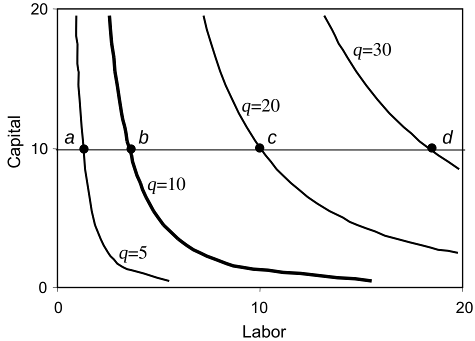


Fig. 5.6. Iso product curves, Cobb-Douglas production function
The iso-product curves show the combinations of capital and labor that will yield the specified level of output. Points a, b, c and d show the quantity of labor required (read off the abscissa) to obtain the specified level of output, given that $K = 10$. The production function is Cobb-Douglas: $q = 2L^{2\ 3}K^{1\ 3}$.

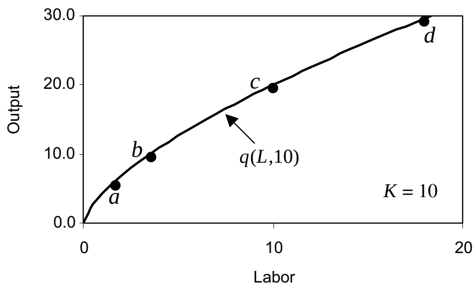


Fig. 5.7. Total product curve ($K = 10$)
The total product curve, derived from the iso-product graph, plots output as a function of labor, given the specified quantity of capital. Points a, b, c and d are carried over from Figure 5.6.

Unlike Figure 5.3, this total product curve does not display an area of increasing returns because the Cobb-Douglas functional form is so simple.

The marginal product of labor is

$$\frac{\partial q}{\partial L} = \lambda\alpha L^{\lambda-1}K^{\lambda'}\,, \tag{9}$$

which simplifies to

$$\frac{\partial q}{\partial L} = \frac{\lambda\alpha L^{\lambda}K^{\lambda'}}{L} = \frac{\lambda[\alpha L^{\lambda-1}K^{\lambda}]}{L} = \frac{\lambda q}{L}\,. \tag{10}$$

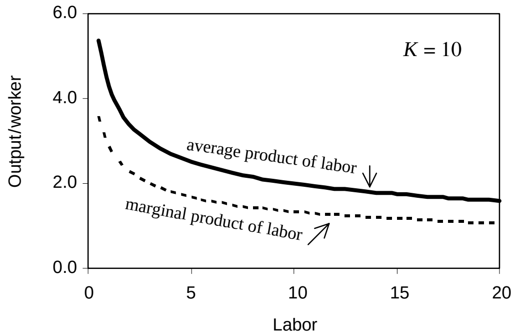


Fig. 5.8. Average and marginal product of labor (Cobb-Douglas case)
For the special case of the Cobb-Douglas production function, marginal product ($\partial q/\partial L$) is proportional to the average product of labor (Q/L); specifically, $\partial q/\partial L = \lambda Q/L$.

The average and marginal product of labor are plotted on Figure 5.8 for the case of 10 machines ($K = 10$), $\lambda = 2/3$ and $\lambda' = 1/3$. These are simpler than those on Figure 5.4 because the Cobb-Douglas production function is subject everywhere to diminishing returns.

The production function provides a simple but useful summary of the available technology. Our next task will be to put it to work in generating cost functions.

5.5 Maximizing profit

The profit maximizing level of output for the representative firm is determined by the appropriate balancing of revenue and costs. We will show how the firm's cost function may be derived from its production function. Then we will seek the profit maximizing level of output.

5.5.1 *The cost function*

The firm's *cost function* explains how its cost of production depends on the level of output. The *short-run cost function* shows how production costs depend on output, given a fixed number of machines, $\overline{K}$, where the bar over the K indicates that it is to be regarded as fixed. Figure 5.8 showed how output depended on the number of workers, given $\overline{K} = 10$. Suppose we are told that it costs $r = \$5$ per hour to use a machine and the workers get a wage $w = \$10$ per hour. Then total costs will be

$$C = r\overline{K} + wL = 5\overline{K} + 10L\,. \tag{11}$$

But this accounting relationship is not the total cost function because it does not relate costs to output. To find the short-run total cost function $C(q, \overline{K})$ showing the minimum cost of producing the specified output, given $\overline{K}$, we need to harness the technological information provided by the production function. We must obtain from the production function the inverse function showing the minimum quantity of labor required to produce q units of output, given $\overline{K}$. This required quantity of labor function, call it $L^*(q, \overline{K})$, is in the case of the Cobb-Douglas production function (6)

$$L^*(q, \overline{K}) = \alpha^{-1/\lambda} q^{1/\lambda} \overline{K}^{-\lambda'/\lambda}\,. \tag{12}$$

This is not as complicated as it seems, for $L^*(q, \overline{K})$ is just the inverse of the total product of labor curve plotted on Figure 5.7 for $\overline{K} = 10$. Substitution of $L^*(q, \overline{K})$ into (11) yields the cost of producing q units of output, given $\overline{K}$:

$$\begin{aligned} C(q, \overline{K}) &= wL^*(q, \overline{K}) + r\overline{K} \\ &= w\alpha^{-1/\lambda} q^{1/\lambda} \overline{K}^{-\lambda'/\lambda} + r\overline{K} \end{aligned} \tag{13}$$

From this short-run total cost function we find that in the short run (given $\overline{K}$) the average total cost is

$$\frac{C}{q} = w\alpha^{-1/\lambda} q^{(1-\lambda)/\lambda} \overline{K}^{-\lambda'/\lambda} + \frac{r\overline{K}}{q}\,. \tag{14}$$

Differentiating (13) yields marginal cost:

$$\frac{dC(q, \overline{K})}{dq} = \frac{w\alpha^{-1/\lambda} q^{(1-\lambda)/\lambda} \overline{K}^{-\lambda/\lambda}}{\lambda} = \frac{[C(q, \overline{K}) - rK]}{\lambda q}\,. \tag{15}$$

Average and marginal costs are plotted on Figure 5.9. The total costs of producing various levels of output per hour are summarized on Table 5.4 for parameter values $\lambda = 2/3$ and $\lambda' = 1/3$; the wage is $w = 20$ and the cost of using a machine is $= \$5$.

Because the Cobb Douglas is a particularly simple form of the production function it generates quite simple cost functions. While the Cobb

Douglas case will suffice for many purposes, in Chapter 6 we will find it necessary to consider a somewhat more complicated shape for the cost function. In general, given any production function $q(L, K)$, the short-run total cost function is

$$C(q, \overline{K}) = r\overline{K} + wL^*(q, \overline{K}), \tag{16}$$

where $L^*(q, \overline{K})$ is the required quantity of labor function derived from $q(L, K)$.

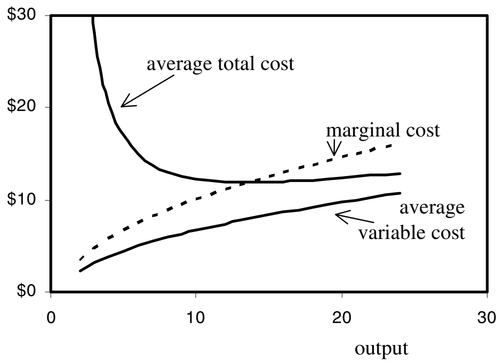


Fig. 5.9. Short-run cost functions
These costs functions were derived from our Cobb-Douglas production function under the assumption that $r = 5$, $w = 20$ and $\overline{K} = 8$.

Note that at the level of output where average total cost is at its minimum, marginal cost equals average total cost.

Table 5.4. Shortrun total cost function.

Output	Labor required	Variable cost	Total cost	Average cost	Average variable cost	Marginal cost
q	$L^*(q)$	$C(q) - r\overline{K}$	$C(q)$	$C(q)/q$	$(C(q) - rK)/q$	$dC(q)/dq$
0	0.0	0.0	40.0			
5	1.4	28.0	68.0	13.6	5.6	8.4
10	3.9	79.1	119.1	11.9	7.9	11.9
15	7.3	145.2	185.2	12.3	9.7	14.5
20	11.2	223.6	263.6	13.2	11.2	16.8
25	15.6	312.5	352.5	14.1	12.5	18.8

5.5.2 *Short-run profit maximization*

The total profit of our firm equals total revenue minus total costs.

$$\boldsymbol{\pi}(q) = R(q) - C(q)\,, \tag{17}$$

where $R(q)$ is total revenue, price times quantity. (We may simplify notation by omitting the fixed capital stock $\overline{K}$ from the total cost function, writing $C(q)$ for total costs.) If our firm sells its product on a competitive market, it will be a price taker, the price it receives for its product being independent of how much it desires to sell. If our firm is a monopoly, however, the price it will receive for its product will depend on the quantity it decides to sell. Quite generally, we may specify total revenue to be

$$R(q) = p(q)q\,, \tag{18}$$

where $p(q) = q^{-1}(p)$ is the inverse demand function discussed in Chapter 3. In the competitive case, $dp/dq = 0$ and so the demand curve is horizontal. Under monopoly, $dp/dq < 0$, the demand curve having a negative slope. To find the q that maximizes profits we set the first derivative of profit function (17) equal to zero:

$$\frac{d\boldsymbol{\pi}(q)}{dq} = \frac{dR(q)}{dq} - \frac{dC(q)}{dq} = 0\,. \tag{19}$$

Therefore, profit maximizing requires

$$\frac{dR(q)}{dq} = \frac{dC(q)}{dq}\,. \tag{20}$$

This is the fundamental proposition that in order to maximize profits it is necessary to adjust output so that *marginal revenue*, (dR/dq), equals *marginal cost*, (dC/dq).

Since marginal revenue is a concept of fundamental importance, it deserves closer inspection. Differentiation of (18) reveals that marginal revenue is

$$\frac{dR}{dq} = p + \frac{dp}{dq}q\,. \tag{21}$$

In the case of competition, $dp/dq = 0$ and marginal revenue equals price. The monopoly case is clarified on Figure 5.10. From equation (21) it is clear that at $q = 0$, $dR/dq = p$, as indicated on the graph. Also, if as on the graph the demand curve is linear, the slope of the marginal revenue curve will be $d^2R/dq^2 - 2\,dp/dq$; i.e., the marginal revenue curve is twice

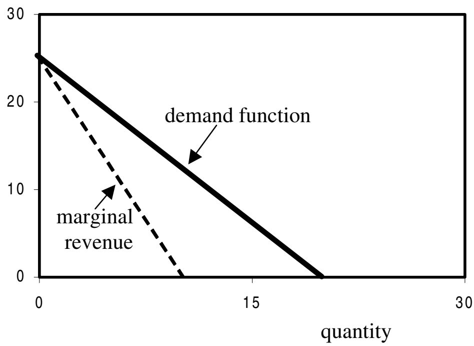


Fig. 5.10. Demand function $q = 20 - (4/5)p$
Since the inverse demand function is $p(q) = 25 - (5/4)q$, revenue is $R(q) = qp(q) = 25q - (5/4)q^2$ and marginal revenue is $dR(q)/dq = 25 - (5/2)q$.

as steep as the demand curve. All this means that whendemand is linear the marginal revenue line will cross the abscissa at $q(0)/2$.

There is an interesting relationship between marginal revenue and elasticity:

$$\frac{dR}{dq} = p + \frac{dp}{dq}q = p\left(1 + \frac{dp}{dq}\frac{q}{p}\right) = p\left(1 - \frac{1}{\eta}\right), \tag{22}$$

where $\eta = -(dq/dp)(p/q)$ is the price elasticity of demand discussed in Chapter 3.

Note the following two fundamental points:

#1. Because a firm in a competitive industry is a price taker facing a horizontal demand curve, $dp/dq = 0$ and so marginal revenue equals price.
#2. If demand is inelastic, $(\eta < 1)$, then $dR/dq < 0$.

The second proposition implies that no profit maximizing firm would sell at a point where its demand elasticity is less than one. Why? Because by selling less it could increase revenue without increasing total cost.

The profit maximization story for our firm is summarized graphically on Figure 5.11. The average and marginal revenue curves have been reproduced from the preceding graph and the average and marginal cost curves from Figure 5.8. Profits are maximized at the output level q for this yields marginal revenue equal to marginal cost at point e. The highest price that can be obtained at this level of output is p. Average cost at output q, call it c_a, may be read off the average total cost curve. Profit per unit of

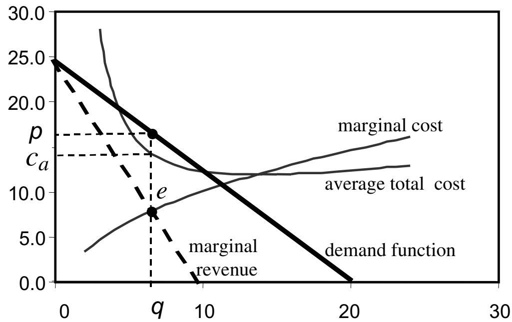


Fig. 5.11. Profit maximization ($\overline{K} = 10$)
The profit maximizing level of output can be determined once the demand and marginal revenue curves of Figure 5.10 are superimposed on top of the marginal and average total cost functions of Figure 5.9.
Profits are maximized by setting output at q because at point e marginal revenue equals marginal cost. Price p is the maximum price at which q can be sold.

output is the excess of price over average cost, $p - C_a$, and total profits of $\pi = (p - C_a)q$ are indicted by the upper rectangle on the graph.

5.5.3 *Least-cost input mix — Costs in the long run*

The preceding analysis relied on the assumption that the timeframe for the analysis was so short that our firm could only adjust the labor input; there was insufficient time to adjust the quantity of physical capital (e.g., machinery and factory space) — that is what is meant by the *short run*. In the *long run*, the firm has enough time to adjust the number of machines as well as the number of workers. Our firm must adjust its capacity on the basis of its projections of future demand for its product.

As a first step toward deriving the firm's long-run cost function, suppose our firm expects to sell 10 units of output. What would be the best input mix? That is to say, what combination of capital and labor will minimize the total cost of producing this level of output? Our firm's task is to minimize

$$C(L, K) = wL + rK \tag{23}$$

subject to the constraint

$$q(L, K) = 10\,. \tag{24}$$

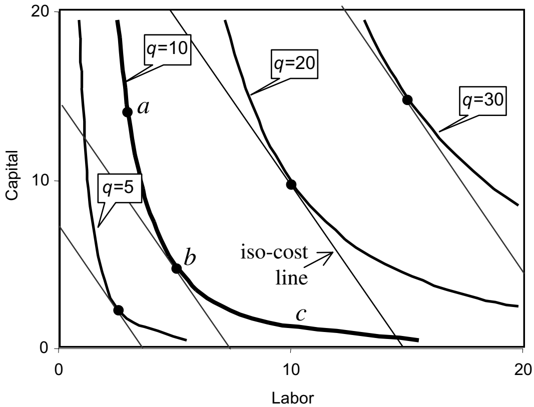


Fig. 5.12. Least cost input mix
Each iso-cost line identifies all input combinations involving a particular level of expenditure, given that the wage is \$20 per hour and the rental cost of a machine is \$10 per hour. An iso-cost line is tangent to the iso-product curve at the input mix that produces the specified level of output at least cost.

The problem is clarified by Figure 5.12, which reproduces the iso-product curves from Figure 5.6, but with the addition of some iso-cost lines. Each *iso-cost line* plots the combinations of labor and capital incurring a given cost; i.e., from equation (23) we have for a typical iso-cost line

$$K = \frac{\overline{C}}{r} - \frac{w}{r}L\,, \tag{25}$$

where $\overline{C}$ is the specified cost, say \$150. Note that all iso-cost lines have slope $dK/dL = -w/r$, which explains why they are parallel on Figure 5.12. Now ten units of output could be produced at any point on the $q = 10$ iso-product line, including points a, b and c. But tangency point b marks the least costly method of producing 10 units because the $q = 10$ iso-product curve cannot be reached from any lower iso-cost line. The tangency condition may be written

$$-\left.\frac{dK}{dL}\right|_q = \frac{w}{r}\,, \tag{26}$$

where $dK/dL\big|_q$ is the slope of the iso-product line. Now $-dK/dL\big|_q$ is called the *marginal rate of substitution in production* and so our efficiency condition is that profit maximization requires that the mix of inputs be such that the marginal rate of substitution equals the price ratio, w/r.[15]

There is a useful link between the marginal rate of substitution in production and the marginal products of capital and labor that can be explained intuitively. Suppose we increase by a small amount dL the quantity of labor employed in the production process and simultaneously reduce the amount of capital by an amount dK that is just sufficient to leave total output dq unchanged; then by the chain rule of calculus we must have:

$$dq = \frac{\partial q}{\partial L} dL + \frac{\partial q}{\partial K} dK = 0\,. \tag{27}$$

which yields

$$-\left.\frac{dK}{dL}\right|_q = \frac{\partial q/\partial L}{\partial q/\partial K}\,; \tag{28}$$

i.e., the marginal rate of substitution in production is the ratio of the marginal productivity of labor to the marginal productivity of capital. This result, with (26), yields an alternative statement of the efficiency condition:

$$\frac{\partial q/\partial L}{w} = \frac{\partial q/\partial K}{r}\,. \tag{29}$$

This means that the profit-maximizing firm will employ the mix of inputs that equates the ratios of the marginal productivity of each input to its price.

The task of deriving the optimal input mix in the theory of the firm parallels the consumer's problem, central to Chapter 4.3.3, of choosing what commodities to purchase, subject to a budget constraint. We shall exploit this parallelism in deriving efficiency condition (29) with the aid of the calculus by noting that our firm obviously is not operating efficiently if it could produce more output without any increase in cost. We will seek to maximize

$$q(L, K) \tag{30}$$

subject to the constraint

$$\overline{C} = wL + rK\,,$$

[15] The "monopsonistic" complications that arise when the price of the inputs depends on the quantity the firm purchases are discussed in Chapter 7.

where $\overline{C}$ is the current level of expenditure.[16] The first point to note is that if costs are to remain fixed at $\overline{C}$, the quantity of capital we can employ must satisfy equation (25). Substituting for K in production function (30) converts the problem of maximizing a function of two variables subject to a budget constraint into the task of maximizing output as a function of L alone, given $\overline{C}$, w, and r:

$$q^*(L) = q\left(L, \frac{\overline{C}}{r} - \frac{w}{r}L\right).$$

As a necessary condition for a maximum,

$$\frac{dq^*(L)}{dL} = \frac{\partial q}{\partial L} + \frac{\partial q}{\partial K}\frac{dK}{dL} = \frac{\partial q}{\partial L} - \frac{\partial q}{\partial K}\frac{w}{r} = 0\,, \tag{31}$$

The last equality yields efficiency condition (29).

Now let us look at the Cobb-Douglas production function, equation (6). The marginal product of labor was given by equation (10): $\partial q/\partial L = \lambda q/L$; similarly, the marginal product of capital is $\partial q/\partial K = \lambda' q/K$. Substituting into (29) we have

$$\frac{K}{L} = \left(\frac{\lambda'}{\lambda}\right)\frac{w}{r},. \tag{32}$$

This means that the capital/labor ratio for the Cobb-Douglas production function depends only on the ratio of the wage rate to the hourly cost of using the machine. Since K/L does not depend on output, the efficient points identified on the successive iso-product curves on Figure 5.12 all lie on a straight line emanating from the origin. The Cobb-Douglas production function is only one member of a larger class of "homothetic" production functions with this interesting property.

5.5.4 *Constant returns to scale (homogeneity of degree 1)*

Suppose that a successful wheat farmer decides to double profits by doubling the acreage and spending twice as much on hired labor, seed, fertilizer and other supplies. Is it reasonable for the farmer to presume that with twice the inputs twice as much will be produced? This is the assumption of *constant returns to scale.* It is the assumption that the production func-

[16] The problem of maximizing output for given costs is sometimes called the "dual" of the corresponding problem of minimizing the cost of producing a given output.

tion is homogeneous of degree 1. A function $q(L,K)$ is said to be *positive homogeneous of degree* κ if for every $\rho > 0$

$$q(\rho L, \rho K) = \rho^{\kappa} q(L,K)\,. \tag{33}$$

The Cobb-Douglas production function is homogeneous of degree $\kappa = \lambda + \lambda'$. To see why, note that if $q(L,K) = \alpha L^{\lambda} K^{\lambda'}$, then $q(\rho L, \rho K) = \alpha(\rho L)^{\lambda}(\rho K)^{\lambda'} = \rho^{(\lambda+\lambda')}\ \alpha L^{\lambda} K^{\lambda'} = \rho^{(\lambda+\lambda')} q(L,K)$. As mentioned in connection with equation (7), Cobb and Douglas found in their pioneering empirical study that $\lambda = 0.807$ and $\lambda' = 0.232$, which implies slightly increasing rather than constant returns to scale. But $0.807 + 0.232 = 1.039$ is so close to unity that it is reasonable to conclude that the empirical evidence is compatible with the hypothesis of constant returns to scale. After all, the sample of historical data used in their study was not all that large and the presence of random measurement error might well suffice to explain such a small discrepancy.

Constant returns to scale has an important implication for the shape of the total cost function. Let L^* and K^* be the least-cost combination for producing q^* units of output and C_a^* the average cost of producing that output. Then if we change both inputs by the proportionate factor ρ, output will go up by that same proportion and so will the total cost of producing it. But this means that average cost will remain unchanged (assuming the cost of inputs do not change). This demonstrates that average costs are constant in the long run if the production function is homogeneous of degree one; i.e., $C = cq$, where c is average cost; obviously $dC/dq = c$ as well. This is an exceedingly simple form of the cost function. We consider a slightly more complicated decreasing average cost function in the next section.

5.5.5 *Pricing to maximize profit*

Research and development of a new product can be a very expensive undertaking. The patent system rewards the inventor by prohibiting imitators from marketing a similar product for a period of time, usually 20 years in the United States. The rational behind the patent system is that it provides a temporary monopoly in order that the developer of the new product will be able to recover its development expenses and be compensated for risk by charging a higher price for its new product. How would inventors recover development costs in the absence of the temporary monopoly provided by the patent system? They might try to keep their invention secret, but that is difficult to do. Defenders of the patent system point out that the disclosure of information required when one applies for a patent contributes to the dissemination of information that can contribute to technological progress once the patent has expired.

Consider, for example, the development of pharmaceuticals. It is estimated that developing a new pharmaceutical product through final testing may cost half a billion dollars, but many products fail in the development stage and never make it to market. Once developed, there are costs of distribution and marketing, which often exceed development costs. But the marginal cost of producing an additional dosage of the medicine may be relatively cheap. For example, Fluconazole, an anti-fungal agent patented by Pfizer, is the treatment of choice for cryptococcal meningitis, a disease that attacks more than 20 percent of AIDS victims in some countries. If untreated this disease all too frequently kills its victims within a month. The wholesale price of the drug was \$9.78 per pill in the United States in year 2000, which means it may retail for as much as \$40. In India, which did not honor American patents, a major drug company profitably marketed a generic version of the drug at 64 cents a pill, which suggests that actual production costs are minimal.[17]

Figure 5.13 considers a patented product with total costs of $C(q) = \$1{,}000{,}000 + 5q$. The plot of average costs $C_a = \$1{,}000{,}000/q + \$5$ and marginal cost $dC/dq = \$5$ reveals that marginal cost is constant but average total cost is decreasing. The hypothetical linear demand function is $q = 200{,}000 - 5p$. Profits are maximized at a price of \$22.50 a dose, but not quite half of the infected receive the drug. Under the patent system, monopoly pricing means that many do not receive treatment. There is no easy answer to the question of how the development of wonder drugs should be financed and how they should be distributed, but we shall return to the topic in the next chapter when we discuss alternative pricing procedures, the patent system, and monopoly.

5.6 Management science illustrated: OLS & JIT

Economists customarily assume that business enterprises succeed in maximizing profit. In contrast, faculty at major business schools teaching courses in Production Scheduling, Operations Research, or *Management*

[17]The information on Fluconazole was reported in the *New York Times*, July 9, 2000, page 8 ne.

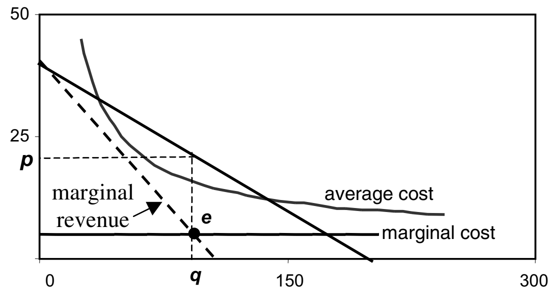


Fig. 5.13. Pharmaceutical pricing
With demand function $q = 200 - 5p$ and cost function $C = 1{,}000{,}000 + 5q$ (where output is measured in thousands), the profit maximizing level of output (determined by the intersection at e of marginal revenue with marginal cost) is $q = 87.5$ thousand. This output can be sold at price $p = \$22.50$. It will be shown in Chapter 6 that this high a price is inefficient and causes a substantial loss in welfare. But if the price is too low, who will pay for the next generation of miracle drugs?

Science make substantial sums consulting for business firms. They are hired on a part-time basis to tell the business firms how they can operate more profitably — their efforts bring the traditional maximization assumption of economists closer to reality.

A simple example will illustrate how the calculus can be applied to help a business firm reduce costs. Our firm produces over 600 different products on a complex machine. Because substantial cost is involved in setting up the machine to produce a particular item, the firm produces each product in batches large enough to meet two months of sales. After a batch is produced, it is stocked in the warehouse. The inventory is gradually drawn down as orders for the good are received from the firm's customers. For example, Product A sells 100 units a month on average, so the two-months supply rule dictates the production of 200 units in each batch. If a batch of 200 is produced on January 1, the stock in the warehouse will initially be 200, but it will gradually be drawn down as illustrated on Figure 5.14. Around March 1, when the inventory is exhausted, the machinery will be set up again and another batch will be produced. Thus, the time plot of the inventory stock pictured on the graph looks like an upside down saw blade. Similarly, Product B sells 150 units a month so it is produced in batches of 300 while Product C sells only 50 units and is produced in batches of 100. Note that on average the ratio of inventories to monthly sales is 1 for each product.

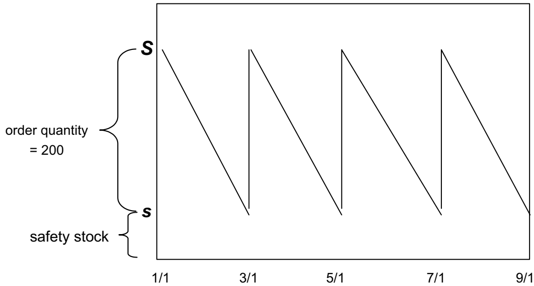


Fig. 5.14. Inventory of good A
The level of inventories, measured on the ordinate, is initially set at level S. Sales gradually draw down the stock of inventories until they reach the lower s trigger point after two months. Then a new lot of size S-s is produced, which rebuilds the stock up to level S; then the cycle repeats. The base stock level of s, the "safety stock," is maintained in order to avoid stocking out as a result of unexpected sales or supply disruptions.

The top half of Table 5.5 summarizes the experience of the firm for three of its many products under the assumption that each setup costs \$150 and that it costs \$10 to carry an item in inventory for one year. A consultant claims that the firm could do better using an "optimal-lot size" procedure. The bottom half of the table shows the implications of using the procedure recommended by the consultant, which involves a smaller than the two months batch size for heavier selling items and a larger then two months supply for the slow selling item. The bottom half of the table shows that an annual cost savings of \$137 will be realized on these three items using optimal lot sizes. This is on only three items. Since the firm produces more than 600 different items, the total saving from adopting the rule will be substantial.

5.6.1 *Optimal lot size formula (OLS)*

The optimal lot size formula recommended by the consultant (first derived in the 1920s) minimizes the sum of setup plus inventory carrying costs. Let q equal annual sales of a product and D the size of a batch. Then there will be q/D setups a year. If c_s is the cost of a setup, then setup costs for the year will total $c_s q/D$. Assuming sales take place at a roughly uniform rate throughout the year, inventory will range from 0 to D with an average level of $D/2$. If the cost of carrying a unit of output in inventory for a year

Table 5.5. Inventory carrying plus setup costs.

	Annual sales	Lot size	Setups/ year	Average inventory	Annual costs Setup cost	Inv costs	Total costs	Months Supply	Inventory/ Sales ratio
Current practice, two months' supply									
Good A	1,200	200	6	100	900	1,000	1,900	2.0	1.0
Good B	1,800	300	6	150	900	1,500	2,400	2.0	1.0
Good C	600	100	6	50	900	500	1,400	2.0	1.0
Totals	3,600		18	300	2,700	3,000	5,700		1.0
Optimal lot size (square root rule)									
Good A	1,200	189.7	6.3	94.9	949	949	1,897	1.9	0.9
Good B	1,800	232.4	7.7	116.2	1,162	1,162	2,324	1.5	0.8
Good C	600	134.2	4.5	67.1	671	671	1,342	2.7	1.3
Totals	3,600		18.5	278.1	2,781	2,781	5,563		0.9
Saving on three items							137		

Note: Setup costs: \$150; inventory carrying cost: \$10 per year.

is c_i, then yearly inventory cost for this item will be $c_iD/2$. We must pick the batch size D that minimizes total cost:

$$C = \frac{c_iD}{2} + \frac{c_sq}{D}\,. \tag{34}$$

The inventory carrying costs and setup costs for the year plus their annual sum are plotted on Figure 5.15. We must find the lot size D that gets us to the minimum annual total cost point indicated on the graph. To derive the optimal lot size D which minimizes the annual costs of meeting sales demand q, we differentiate (34), obtaining as a necessary condition for a maximum $\partial C/\partial D = c_i/2 - c_sq/D^2 = 0$, or

$$D = \sqrt{\frac{2c_sq}{c_i}}\,. \tag{35}$$

For example, in the case of Good A with $q = 1,200$, $c_s = \$150$ and $c_i = \$10$ per year, we have

$$D = \sqrt{\frac{2 \times 150 \times 1200}{10}} \doteq 190$$

Equation (35) is the famous square-root rule for the optimal lot size. As expected, the optimal lot size increases if the setup cost increases; it decreases if inventory carrying costs go up. The surprising result is that if sales go up four fold, the optimal lot size will only double — average costs fall when output increases.

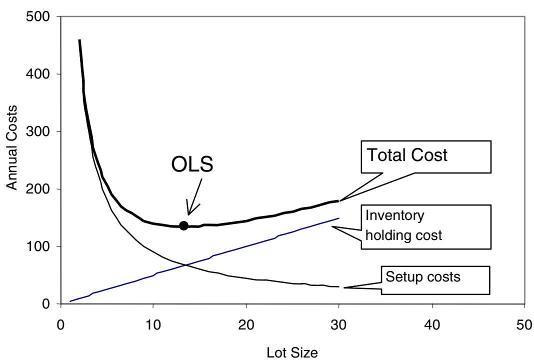


Fig. 5.15. Optimal lot size
The optimal lot size (OLS) is the lot size that minimizes the sum of annual setup costs plus inventory holding costs.

Our derivation of the Optimal Lot Size formula invoked the simplifying assumption that sales took place at a constant rate, but Professor Herbert Scarf of Yale University showed that (35) is the optimal rule under quite general conditions.[18]

A closely related concept, the "efficient order quantity," (EOQ) arises in procurement. If a retailer orders items in larger quantities from the wholesaler, it will incur lower annual ordering costs, but its inventories will on average be larger. Precisely the same square root relationship arises from this argument; equation (35) holds with c_s now representing ordering costs.

These examples are simplified. If a time-lag is involved in producing the product, it is necessary to add a "safety stock." Otherwise, the firm will lose sales because it will stock out of the item if it waits until stocks are zero before producing another batch. This complication leads to what is known as an S,s inventory policy. Our firm sets up the machine and starts to make a run as soon as the stock hits the lower trigger level s. It produces enough to rebuild the stock to level S. If it decides to set s = 50, the firm will start producing the item as soon as the stock reaches that level. It will produce the quantity given by the optimal lot size formula plus enough to replenish the safety stock.

5.6.2 *Just in time*

Just in Time (JIT), a managerial concept originating in Japan, is an alternative to batch processing.[19] The JIT strategy is to re-engineer the production process, and perhaps also modify the design of the product, so as to eliminate or to reduce drastically the set-up-cost. The development of JIT in the 1970s gave a tremendous cost and quality advantage to Japanese auto manufacturers, but the technique has now been adopted by many manufacturing firms around the world.

Here is an application illustrating the basic principles involved. Toyota designed an automated factory for making the wide variety of gauges required for its autos. With batch processing substantial setup costs were incurred because of the idled labor and the loss of production during down time while the robotic machines were adjusted to switch to the produc-

[18]Scarf, H., "The Optimality of (S,s) Policies in the Dynamic Inventory Problem," in *Mathematical methods in the Social Sciences* (1959).

[19]See the online OR notes of J. E. Beasley, http://www.ms.ic.ac.uk/jeb/or/jit.html.

tion of a different type or model of gauge. In order to implement JIT, the robots were reengineered to automatically switch from producing one type of gauge to another. The switch was accomplished by sending a dummy *signal gauge* down the assembly line, which was sensed by each robot. Instead of doing its thing to the dummy gauge as it progressed to its station on the assembly line, each robot in turn would read the dummy signal gauge's instructions to self-adjust its settings to reflect the next type of gauge to be produced. Thus, the setup cost (c_s) is reduced to the opportunity cost of not producing one part. With negligible setup cost, the optimal inventory is cut to essentially zero, as can be seen from equation (35).

A variety of benefits accrue to the firm that switches to JIT. Advocates of JIT argue that in addition to reducing inventories and their carrying cost, JIT also contributes to tighter quality control. Inventories hide defects! With batch processing, maladjustment of a machine may lead to the production of an entire lot of defective parts before the problem is recognized. The problem is likely to be compounded because an assembly-line worker encountering a defective part will be tempted to throw it back into the inventory pile and take another part rather than addressing the problem. Not only does this adverse selection mean that the next shift may inherit an inventory with a high number of defects; it also means that a substantial number of defective parts will have been produced before the defect is properly recognized and corrective actions taken. That is why reducing inventory with JIT can make a key contribution to improved quality control.

Summary

This chapter looked at the business enterprise — the firm.

1. There are many forms of business organization, of which corporations, partnerships and single proprietorships are the most important.
2. A corporation's shareholders have invested in the firm by purchasing shares of its stock, which represent the proportion of the corporation that they own. The shareholders will be rewarded for their investments with dividends and capital gains, provided the firm prospers. The shareholders, because they are owners, participate in the governance of the firm by voting for the directors of the corporation, getting to cast one vote for each share of stock that they own. Corporations benefit from limited liability, which means that if things turn out badly

investors can lose only as much as they have invested in the business. However, corporations have to pay the corporate profit tax.

3. The firm's balance sheet is a statement of the firm's financial position at a particular point of time. The typical balance sheet will list the assets of the firm, which are the things it owns, such as building and equipment and inventories of purchased materials and finished product. It also lists what the firm owes to its creditors, such as loans to the bank or bonds outstanding. The firm's net worth is the excess of assets over liabilities.
4. By appropriately selecting depreciation procedures, inventory valuation techniques and other arcane procedures, accountants may strive to reduce the corporation's accounting profits for tax purposes in order to reduce the firm's corporate tax liability. In preparing data for the annual report to the shareholders, accountants may strive to enhance reported profit in the hope that it will make the firm appear more profitable to potential investors.
5. Economic profit differs from accounting profit in that it includes the opportunity cost of the time and capital that the owners invest in the firm. Further, economic profit is calculated with depreciation representing an accurate picture of the wear and tear on capital equipment rather than the accelerated depreciation that the firm's accountants may report for tax purposes.
6. The total product curve shows how the output of the firm changes as a function of the quantity of labor (or some other input), given the quantity of all the other inputs (e.g. capital).
7. The production function shows how the level of output depends on the various inputs used by the firm, such as labor and capital.
8. The firm's short-run cost function is derived from the production function, given the quantity of physical capital, wage rates and other cost factors. The long-run cost function reports total cost as a function of output, given that the least cost mix of capital, labor and other inputs is utilized.
9. Given a firm's total cost function and its demand curve we derived the optimal level of output. We proved that at the profit maximizing level of output marginal cost must equal marginal revenue.
10. Management Science and Operations Research are sister disciplines concerned with the application of quantitative techniques to improve business efficiency. The optimal lot size formula shows that optimal

inventory stocks should be proportional to the square root of sales volume.

11. Just in Time (JIT), a revolutionary technique developed by Toyota and imitated by manufacturing firms around the world, redesigns the production process in order to drastically reduce setup costs. The result is much smaller lot size, drastically reduced setup and inventory carrying costs, and better quality control.

Key Concepts

Exercises

1. On the basis of the Fly-by-Nite Aircraft Company's December 31, 2000 balance sheet (Table 5.2) and the information in Section 5.3.1 that was used in developing the year 2001 income statement (Table 5.3), draw up the firm's balance sheet for December 31, 2001.
2. A new accountant hired by Fly-by-Nite Aircraft recommends two changes in the firm's accounting practice: First, she argues that the firm should employ accelerated depreciation, which will increase its depreciation charge for year 2001 by $100,000. Second, she offers to reclassify $20,000 of labor and $30,000 of materials as capital improvements to be depreciated over a ten-year period.
   a. How would Table 5.3: Income statement of the Fly-by-Nite aircraft company have been affected if these three procedures had been adopted?
   b. Would it be smart for the firm to reclassify $20,000 of labor and materials from a current cost to a capital expense? Explain carefully how this would affect the income statement. What are the pros and cons of this accounting shift?
3. A firm uses two inputs, physical capital K (machinery, buildings and other capital equipment) and L (hours of labor) to produce output q per hour. The Cobb-Douglas production function is $q = K^{0.5}L^{0.5}$.
   a. Suppose our firm has 1 machine and employs 16 workers. What is output per hour, the average product of labor, the marginal product of labor, and the marginal product of capital? In general, find equations for each of these concepts as a function of L and K.
   b. Is this production function homogeneous of degree 1? Show why or why not.

   Suppose workers earn $10 per hour and the firm has only one machine, which costs $20 per hour to operate.

c. How much labor will our firm have to hire to produce 10 units of output per hour? What will be the total cost to our enterprise of producing 10 units of output per hour? What is average (total) cost? What is average variable cost?

d. Now derive the firm's total cost function, $C(q)$, showing how the total costs depend on the level of output per hour, given that the firm has only one machine.
Note: this is referred to as the "short-run" total cost function because the firm does not have enough time to adjust the number of machines. Hint: First find the function showing how much labor is required as a function of q, given that $K = 1$.

e. Determine the (short run) marginal cost, the average cost, and the average variable cost functions, given that $K = 1$.

4. In the *long run* the firm of question 3 has sufficient time to adjust the number of machines.

   a. What is the least cost technique for producing 10 units of output, given that workers earn $10.00 per hour, that machines cost $20 per hour to operate, and that the production function is $q = K^{0.5}L^{0.5}$? That is to say, find the quantities of labor and capital we should employ if we are to produce 10 units of output at minimum cost. What will be total cost and cost per unit (average cost) when the firm produces output of 10 at minimum cost?

   b. What is the minimum average cost if the long run level of output is 20? Hint: Doubling inputs will double output and double total costs.

   c. Determine the firm's long-run total cost function, given that it can have as many machines as it wants at a cost of $20 per machine.

5. Monopoly: A firm facing inverse demand function

$$p(q) = 20 - 2q$$

has total cost function

$$C(q) = 4q + q^2 .$$

   a. Determine the total revenue function $R(q) = q \cdot p(q)$ and the marginal revenue function $dR(q)/dq$ for our firm.

b. Solve for the level of output and price yielding maximum profits $\pi = R(q) - C(q)$.

What is total revenue, total costs, total profits, profits per unit of output, marginal cost, average total cost, and average variable cost at this level of output?

c. Plot on a graph the demand function, the marginal revenue function, the marginal cost curve, and average total cost. Then indicate on the graph the information you reported in b above.
d. Calculate the amount of consumer surplus that is enjoyed under the monopoly.

6. The Brand X firm has cost function $C(q) = 24 + 4q + q^2$.

a. Determine average total cost, average variable cost and marginal cost when output is 20. Determine average total cost, average variable cost and marginal costs as functions of output. Label these three curves on a graph similar to Figure 5.9.
b. If the firm's product sells in a competitive market at price $p = \$20$, what quantity would the firm sell in order to maximize profits?
c. If a recession meant that the competitive market price for the firm's product fell to \$10, how much would our firm market in the short run?

7. The Brand Y firm has the same cost function as the Brand X firm. It faces the demand function $q = 10 - 1/3\,p$.

a. Plot the demand curve on a neat graph.
b. What is our monopolist's inverse demand function?
c. Derive the equation for marginal revenue. Plot the marginal revenue curve on the graph.
d. Show on the graph what quantity the firm will sell in order to maximize profits. What price will it charge? Indicate total revenue, total cost and profit rectangles on the graph.
e. Is it true for this demand curve that at $q = 0$, marginal revenue equals price? Is it true that $dMR/dq = d^2R/dq^2 = 2 \cdot dp(q)/dq$ for this demand curve? Verify that these two properties hold for *any* linear demand curve $p = a - bq$, with $a > 0$ and $b > 0$.
f. What is the elasticity of demand at the profit maximizing level of output? Recall: $\eta_p = -dq/dp \times (p/q)$.
g. Show that Marginal Revenue $= p(1 - 1/\eta_p)$ at the profit maximizing level of output. Does this condition hold at any level of output? Explain.

8. Management Science
   a. Use the Optimal Lot Size equation (square root rule of section 5.6.1) to check whether the consultant has indeed calculated the appropriate lot sizes for Table 5.5.
   b. Determine how the optimal lot sizes recorded on the table would change if the inventory carrying cost for Good A were $15, unchanged for Good B, and only $8 for Good C?
   c. Observe on that table that the setup size recommended by the consultant yields annual setup costs that are precisely equal to annual inventory carrying cost. Is this a coincidence, or can you show it must always be the case when the optimal lot size is being produced?

9.* An auto dealer leases you a new car. It finances such sales with a loan from the local bank. It depreciates the car over a five-year period. Does the IRS give the car dealer a tax advantage over that which you would receive if you were to buy the car outright and finance it from the bank? Or to put it another way, does car leasing rather than car ownership cost the government tax revenue? Explain.

10.* On the Road:
Use the information on the following table about the cost of driving a car to construct the best estimate you can of the cost function for driving a car, assuming it is the simple linear form

$$\text{Total Cost} = \alpha + \beta \text{ mileage}.$$

   a. Find the marginal cost of driving the car an extra mile.

Table 5.6. Cost of owning and driving a car, 1997 (assuming the car is driven 10,000/year)

Total cost per mile	Cents	53.08
Variable cost	Cents/mile	10.80
Gas and Oil	Cents/mile	6.60
Maintenance	Cents/mile	2.80
Tires	Cents/mile	1.40
Fixed Cost	Dollars	4,228
Insurance	Dollars	809
License & registration	Dollars	220
Depreciation	Dollars	3,268
Finance Charges	Dollars	793

Source: *Statistical abstract of the United States*, 2000.

b. Joe Carman offers to drive a college classmate to the airport (15 miles each way) in his 1987 Honda for \$10. Is he making money? If you drive me to the airport, should you charge me marginal cost, average cost, or the taxi fare? Explain.

11.# The CES (Constant Elasticity of Substitution) production function has the form

$$q(K, L) = [\beta K^\rho + (1-\beta)L^\rho]^{\varepsilon/\rho}\,.$$

Here β is the distribution parameter determining the relative importance of capital and labor in the production process, ε is known as the elasticity of substitution, and ρ is called the scale parameter.[20]

a. What is the marginal product of labor?
b. For what values of β, ε, and ρ does this function exhibit constant returns to scale?

[20] This production function was introduced into the literature by K. J. Arrow, H. B. Chenery, B. S. Minhas and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics, 1961.*

6

Market Structure

6.1 Overview

A *market* is an institutional arrangement facilitating the interaction of buyers and sellers in a process that determines price and quantity sold. Some markets, such as the American Stock Exchange or the town produce market where local farmers congregate to sell their crops, conduct their business at a particular location. Other virtual markets are conducted by telephone or over the Internet. Some markets, such as the Middletown real estate market, are focused on a particular geographic region while some financial markets are international in scope. The different types of markets to be considered in this chapter can be classified along two dimensions:

1. *Markets can be classified in terms of the number of firms in the industry:* At one extreme, a large number of firms may constitute a competitive market and at the other extreme, a single producer enjoys being a monopoly. In between there are oligopolistic industries, which are characterized by a relatively small number of enterprises, as in automobiles, airline travel, and household appliances.
2. *Markets can also be distinguished by the degree of product differentiation:* In some industries, different firms produce essentially the same product, as is the case with heating oil and sunflower seed. In other industries, such as automobiles and computer software, products are differentiated, no two firms producing precisely the same product.

The types of market structure to be considered in this chapter are indicated on the following table.

Table 6.1. Markets classified.

	Identical products	Differentiated products
One firm	Monopoly	Discriminating monopoly
Few firms	Oligopoly	Differentiated oligopoly
Many firms	Competition	Monopolistic competition

In this chapter we will skip from one pigeonhole in this two dimensional array to the next. It is best to start by focusing on the case of competition, which involves a large number of firms producing the same product. Then we shall contrast this competitive form of economic organization with monopoly before moving on to the hybrid blend of monopolistic competition and to oligopoly.

Throughout the chapter we will be assuming that business enterprises are motivated by the desire to maximize profit. We will be putting to work the theory of the profit maximizing firm developed in Chapter 5 in seeking answers to the following questions: What forms of market structure contribute to the efficient allocation of resources and what forms contribute to inefficiency? When is it necessary for the government to intervene in the market place in order to provide a level playing field in which the efforts of firms to maximize profits will lead to the efficient allocation of resources?

6.2 Competitive markets

6.2.1 *Key features of competitive markets*

A *competitive market*, as defined by economists, is characterized by a large number of small firms all selling the same identical product, such as Red Winter Wheat #2 or 3/4 inch fir C-D plywood. The presence of a large number of identical sellers has powerful implications:

1. *No firm can charge more than the price prevailing in the marketplace:* Why, after all, would any customer pay a premium if the identical item can be purchased from another supplier for less?
2. *And no firm would sell for less than the prevailing price:* Selling at a price below that charged by other firms in the competitive market will attract many more customers than the enterprise can possibly serve.

Thus, firms selling in a competitive market are *price takers*. They lack pricing discretion, charging whatever price is prevailing in the market. The demand curve facing the individual firm (as opposed to the *market demand curve* showing the total quantity that customers will purchase of the good as a function of price) is horizontal. If dq/dp_i is extremely large than dp/dq_i must be very small — indeed, it is negligible, which means that in a truly competitive market (as economists define the term) the decision of the individual producer about how much to market has a negligible effect on market price.

3. The market price is that price which equates supply and demand. If the price were so high that sellers could not sell all they wanted to at that price, there would be excess supply; sellers would shave the price rather than be stuck with unsold goods. If the price were so low that buyers could not purchase all they wanted, some would offer to pay more rather than do without, which would push the market price upward. The only price that can prevail in a competitive market — the *equilibrium price* — is that price which equates the quantity that suppliers want to sell at that price with the quantity that customers wish to buy.

These propositions mean that the theory of competition must determine price and industry output from information about the industry demand function and the typical firm's cost function. Solving for this competitive equilibrium is a two step process: We first solve for the firm's supply function showing the profit maximizing level of output for the representative firm, given price. Once this firm supply function is determined, we will find it possible to solve simultaneously for the level of industry output and price.

6.2.2 *The behavior of the competitive firm*

If a firm sells q units of its product in a competitive market at price p, it will realize total revenue

$$R(q) = pq\,. \tag{1}$$

How much profits it will make will depend upon its total cost function, $C(q)$, a concept introduced in Chapter 5.5. Here we shall find it useful to employ for illustrative purposes the function[1]

$$C(q) = 20 + 10q - 3.5q^2 + 0.5q^3\,. \tag{2}$$

In general, a firm's profits, the excess of total revenue over total cost, will be

$$\pi(q) = pq - C(q)\,, \tag{3}$$

or for our numerical example

$$\pi(q) = pq - (20 + 10q - 3.5q^2 + 0.5q^3)\,. \tag{4}$$

Function $R(q)$ is plotted on Figure 6.1 for price $p = \$20$. The functions $C(q)$ and the difference $\pi(q)$ are plotted on the same graph. Profits are maximized at the peak of the $\pi(q)$ profit mountain.

To solve for the profit maximizing level of output, we should set the derivative of $\pi(q)$ with respect to q equal to zero:

$$\frac{d\pi}{dq} = \frac{dR}{dq} - \frac{dC}{dq} = 0\,, \tag{5}$$

with second-order condition for a maximum of $d^2\pi/dq^2 = d^2R/dq^2 - d^2C/dq^2 < 0$. Equation (5) yields the fundamental proposition, already considered in Chapter 5:

> *A profit maximizing firm adjusts output to the point where marginal revenue, dR/dq, equals marginal cost, dC/dq.*

[1] The cost function is slightly more complicated than those of Chapter 5.5 because it includes $-3.5q^3$ in order to allow for a zone of increasing returns.

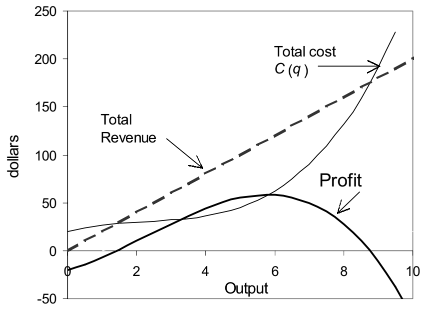


Fig. 6.1. Total revenue, costs and profit in a competitive market
With $p = \$20$, total revenue is $R(q) = pq = 20q$. Total cost is $C(q) = 20 + 10q - 3.5q^2 + 0.5q^3$. Profits are $\pi(q) = R(q) - C(q)$.
The firm sets output q at the level that maximizes profits.

For a firm operating in a competitive market, price p is determined by the market and so $dp/dq = 0$; the firm's total *revenue* is $R = pq$ and the firm's marginal revenue $dR/dq = p$, which yields a second fundamental proposition:

> *A profit maximizing firm producing in a competitive market adjusts its output to the point where marginal cost equals price*

$$\frac{dC}{dq} = \frac{dR}{dq} = p\,. \tag{6}$$

This leaves open the question, to be considered shortly, of whether the price is so low that the firm should not produce anything at all.

To illustrate, consider a firm with cost function (2). Profit maximization requires

$$\frac{d\pi}{dq} = p - 10 + 7q - 1.5q^2 = 0\,, \tag{7}$$

an equation with solution

$$q = \frac{7}{3} + 2.45\frac{(p - 1.83)^{0.5}}{3}$$
$$\text{for } p \geq 1.83;\ \text{otherwise } q = 0\,. \tag{8}$$

This is the *supply function* showing how much our firm will market as a function of market price, except for complications that arise if the suggested level of output should yield negative profits.

The nature of the supply response is clarified by considering Figure 6.2, which displays the marginal cost curve dC/dq and the average total cost curve $C(q)/q$. If the price is $p = \$20$, for example, how do we find the profit maximizing level of output? We must find the level of output equating marginal cost with price. We draw the competitive firm's horizontal demand curve at market price $p = \$20$ and set output where this line crosses the marginal cost curve.

Price is revenue per unit, average cost is $C(q)/q$, and profit per unit realized by the firm $\pi/q = p(q) - C(q)/q$.

Total profit is just profit per unit times output or $\pi(q) = [p(q) - C(q)/q]q$. Thus profit can be read off Figure 6.2 by looking at the rectangle with height $[p(q) - C(q)/q]$ and width q.

Breaking even and the long-run supply function

On Figure 6.2 the level of output has been determined for a particular price. To map out the supply curve showing the quantity that the firm will sell as a function of price it is essential to note that the quantity supplied at each point can be read off the marginal cost curve, provided that profits are positive. If price is so low that profits are negative at the point where price equals marginal cost, the firm may close its doors. The firm will never

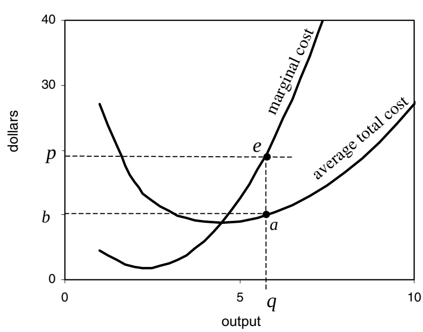


Fig. 6.2. Profit maximization in a competitive market
The curves on this graph may be derived from those on Figure 6.1:
- Average total cost is $C(q)/q$ and marginal cost is dC/dq.
- Marginal revenue equals the price of 20 because the market is competitive.
- Profits are maximized at that output level q where marginal cost equals marginal revenue.

Profits equal the excess of price over average cost times quantity sold — this is represented on the graph by profit rectangle p, e, a, b.

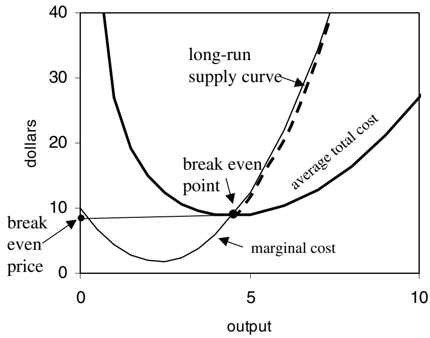


Fig. 6.3. The break-even point and long-run supply curve
The breakeven point is the minimum point on the average total cost curve. The corresponding price is the breakeven price.

At any price below the break-even price the firm will go out of business in the long run, once it has worn out or sold off its plant and equipment.

The long-run supply is the part of the marginal cost curve that is above the break-even point.

produce at a loss if the alternative is to go out of business, perhaps retire. By the long-run supply function we mean the quantity that the competitive firm will choose to sell at a given price when it has sufficient adjustment time to consider the alternative of going out of business.

On Figure 6.3 the minimum point on the average total cost curve has been labeled the break-even point and the corresponding price is referred to as the *break-even price.* For cost function (2) the break-even price is 8.82 at output level $q = 4.5$. If the market price falls below the break-even price the firm will lose money at any positive level of output. If this unhappy situation is expected to prevail, the firm has no alternative but to go out of business. Workers will be laid off, if lucky with severance packages, and the managers may take early retirement, perhaps with a generous "golden parachute" of retirement benefits. At any price above the break-even price, the firm will produce where price equals marginal cost. This argument is summarized by the concept of the long-run supply curve.[2]

$$S_{Lr}(p) = \left[\frac{dC(q)}{dq}\right]^{-1} \quad \text{if} \quad p > p_{be}\,; \quad \text{otherwise,} \quad S_{Lr}(p) = 0\,; \qquad (9)$$

[2] As will be recalled from Chapter 2 and Chapter 5, economists customary plot quantity on the abscissa even though it is a function of price.

here $[dC(q)/dq]^{-1}$ denotes the inverse of the marginal cost function. To recapitulate:

> In the long-run our firm will never produce at a loss but will go out of business instead. If it does produce, the firm will maximize profit by setting output so as to equate marginal revenue with marginal cost; but since competitive firms are price takers, this means that marginal cost is equal to price. Therefore, the firm's long run supply curve is the segment of the marginal cost curve that is above the average total cost curve

The shut-down point and the short-run supply curve

When the economy moves into recession, producers may find that the demand function for their product has shifted downwards to the point where they can no longer sell enough to cover their costs no matter what quantity they produce. Nevertheless, firms that are losing money may decide not to shutdown for good. This is most likely to be a wise decision if prices may reasonably be expected to recover shortly. But even if today's painfully low-price is expected to prevail indefinitely, it may be wise to continue producing for a while because it will take time to liquidate the enterprise, particularly if the firm uses specialized equipment that is difficult to sell off except at distress prices. In such a situation the enterprise may find that a positive level of output with negative profits is the least painful way to make the best of an unfortunate situation. To sum up, the firm may find that it can minimize its losses by postponing going out of business until its productive equipment has worn out.

To see the validity of this counter-intuitive proposition, note that if the firm shuts down it will reap zero revenue but incur fixed costs $C(0)$; thus $\pi(0) = -C(0)$. Our unhappy firm may lose less by producing at the point q^* where $p = dC/dq$. This will be the case if and only if

$$\pi(q^*) = R(q^*) - C(q^*) > \pi(0) = -C(0)\,. \tag{10}$$

Now this inequality condition can be manipulated to obtain

$$R(q^*) > C(q^*) - C(0)\,, \tag{11}$$

where $C(q^*) - C(0)$ is variable cost. This means that it is better to produce at a positive level of output if total revenue will exceed variable costs, thus making a partial contribution toward covering fixed costs! The firm may be losing money, but it would lose even more if it shuts down and incurs costs $C(0)$!

Dividing both sides of inequality (11) by q^* reveals,

$$\frac{R(q^*)}{q^*} = p > \frac{[C(q^*) - C(0)]}{q^*}. \tag{12}$$

On the right of this inequality, we have average variable cost. Thus the inequality specifies that losses will be minimized at output q^* rather than at $q = 0$ if the price is greater than average variable cost at this level of output. This is obviously an unhappy situation, but continuing production will help to recover at least a part of fixed costs.

If price is below the minimum point on the average variable cost function, there is no positive level of output that will satisfy condition (12), and the firm should shutdown; with prices so low, profits will be less than $\pi(0) = -C(0)$ at any positive level of output. That is why the minimum point on the average variable cost curve is labeled the *shut-down point* on Figure 6.4. For cost function (2) the shutdown point is at $p_{sd} = 3.975$ with $q_{sd} = 3.5$. At any price above this minimum *shut-down price* of $p_{sd} = 3.975$, the firm will, in the short run, produce where marginal cost equals price.

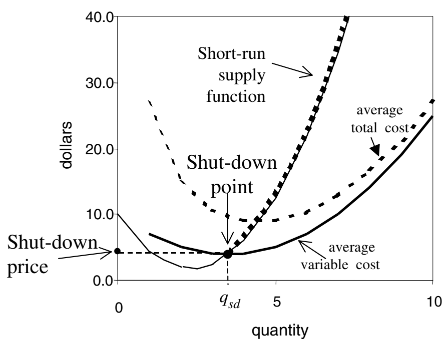


Fig. 6.4. The shutdown point and the short-run supply function
In the short-run the firm that shuts down will have zero revenue but will be stuck paying fixed costs $C(0)$; therefore its profits will be $\pi = -C(0)$.

The shutdown point is the minimum point on the average variable cost curve. The corresponding price is the shut-down price. The shutdown price is the minimum price at which the profit maximizing firm will find it better to stay in production rather than shut down and incur a loss equal to fixed costs. At any price below the shutdown price, the firm will shut down and suffer losses equal to fixed cost $C(0)$.

The short-run supply curve is the segment of the marginal cost curve that is above the shutdown price.

This establishes a fundamental proposition:

> The firm's short-run supply curve is the segment of the marginal cost curve that is above the shut-down price.[3]

6.2.3 *Industry supply*

The assumption of profit maximization allowed us to determine how much a firm selling in a competitive market will decide to market, given the market price and its total cost function. In moving from the firm's supply curve to the industry supply curve, we will find it useful, once again, to make a distinction between the short and the long run.

Short-run supply and demand interaction

In analyzing the competitive market it is useful to define the *short-run* as a period so short that the number of firms in the industry is regarded as historically determined or fixed. Thus, if the function $s_i(p)$ describes the quantity that will be supplied by the ith firm in the industry, and if there are n such firms, industry supply $S(p)$ will be

$$S(p) = s_1(p) + s_2(p) + \ldots s_n(p) = \sum s_i(p)\,. \tag{13}$$

Therefore, $dS(p)/dp = \sum ds_i(p)/dp$. The fact that $ds_i/dp \geq 0$ for all p means that $dS(p)/dp \geq 0$; i.e., the industry supply curve has the expected positive slope.

Now suppose that $Q(p)$ is the market demand function for our commodity. Then the market clearing condition $S(p^e) = Q(p^e)$ yields the equilibrium price p^e and market supply Q^e, as illustrated on Figure 6.5. This explains only the short-run competitive market outcome, given the historically determined number of firms.

The long-run industry supply curve

Only in the short-run may the number of firms in the industry be regarded as fixed. In the longer run firms may enter or leave the industry. To solve for the *long-run competitive equilibrium* requires that we determine the number of firms that will survive in the industry, the output of each firm, the price that will prevail, and total industry output. Here are the

---

[3]The marginal cost curve must go through the minimum point on the average variable cost curve — see exercise 6.2.

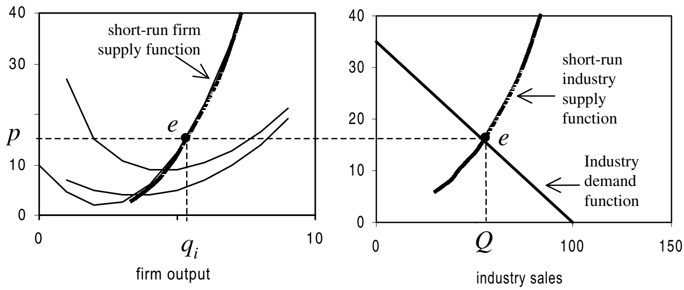


Fig. 6.5. Short-run competitive equilibrium
In short-run competitive equilibrium, the number of firms is historically determined — there is insufficient time in the short run for firms to enter or leave the industry.

The industry supply function on the right-hand panel is determined by adding up the quantities that each individual firm is willing to supply at each specified price. Price is determined on the right by the intersection of the demand and industry supply curves at point e. The representative firm, plotted on the left, is supplying the profit maximizing quantity at the market determined price.

three basic assumptions:

1. Every firm strives for maximum profits.
2. Market demand equals market supply in equilibrium.
3. There is free entry and exit; i.e., new firms can enter the industry and old firms may leave.

Because finding the solution is a tedious task, it is advisable to focus initially on the easy case by invoking the simplifying assumption that all firms, including potential entrants into the industry as well as old and established enterprises, have the same cost function.

Step #1

The first step in deriving the long-run competitive solution is to note that with free entry and exit from the industry, economic profit will be driven to zero in the long run.

- If profits are initially positive, new firms will continually be attracted into the market, which will lead to larger industry supply. This will generate downward pressure on prices and profits. This process will continue until economic profits are driven to zero.
- Conversely, if firms are initially losing money, their numbers will gradually shrink to the point where the surviving firms have just enough business to make a non-negative profit.

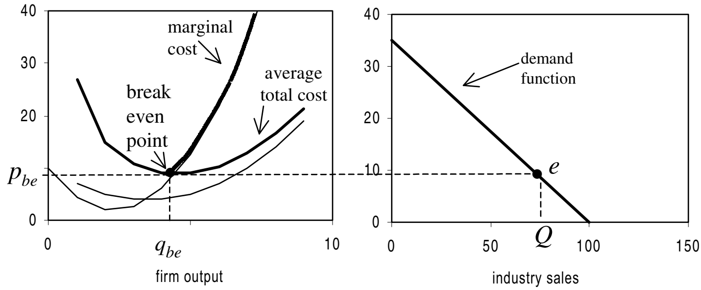


Fig. 6.6. Long-run competitive equilibrium
In the long run there is time for new firms to enter and/or old firms to leave the industry. If new firms are free to enter with the same cost structure as existing firms, the price cannot prevail above the breakeven point on the typical firm's average total cost function. And if the price fell below the breakeven point, firms would exit the industry. Thus the only price that can prevail in the long run is the break-even price.

Needless to say, every firm is eagerly trying to maximize profit; but the number of firms in the industry adjusts until economic profits are driven to zero.[4]

Step #2

The second step in deriving the long-run competitive solution is to note that zero profits imply that $p = C(q)/q$; i.e., price is equal to average total cost. Further, profit maximization implies that firms are producing that quantity at which price is equal to marginal cost, $p = dC/dq$. These two conditions are satisfied only when the representative firm is producing at the minimum point on its average total cost curve, the break-even point on Figure 6.3. In the long-run each firm will produce at the break-even point, producing output $q^e = q_{be}$ and charging price p_{be}. The zero profit condition means that the long-run, equilibrium price $p^e = p_{be}$.

Step #3

The third step in solving for long-run competitive equilibrium is to find total industry sales by substituting equilibrium price p^e into the market demand equation:

[4]This does not mean that accounting profits are zero. As was pointed out in Chapter 5, accounting profits tend to be higher than economic profits because accountants do not subtract all the costs of owner supplied labor and capital in computing profit.

$$Q^e = Q(p^e)\,. \tag{14}$$

Step #4

The fourth and final step is to note that the number of firms in the industry must be

$$n = \frac{Q^e}{q_{be}}\,. \tag{15}$$

How will the competitive industry respond to a change in demand? To find the answer, suppose the market demand curve shifts to the left, perhaps because of a hike in taxes, a recession, or an increase in foreign competition. Initially firms will experience losses, causing an exodus of firms from the industry. This process will continue until the industry has adjusted to a new long-run competitive equilibrium with zero profits. As illustrated on the right-hand panel of Figure 6.7, the adjustment requires a decline in the number of firms in the industry.

How long is the long run?

How quickly and easily firms can enter or leave the industry will be influenced in part by the nature of the product and the production process. Establishing an apple orchard takes both capital and time. Many years will elapse after the land is cleared and the trees are planted before the new orchard will start producing a commercial crop. Once the trees have

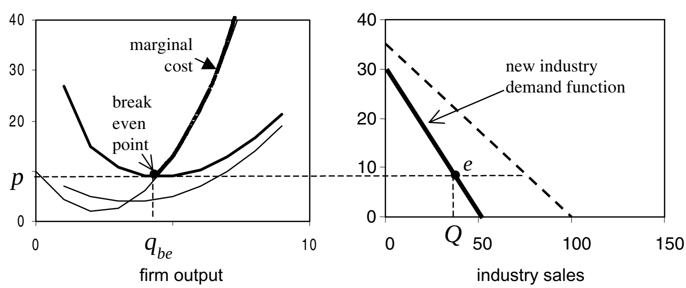


Fig. 6.7. Long-run competitive supply response
When the long run competitive equilibrium is disturbed by a shift to the left of the industry demand curve, firms in the industry suffer a profit squeeze. In the short-run the price will fall and firms lose money. In the long run firms will leave the industry and the price recovers. The industry eventually converges to a new long-run competitive equilibrium with fewer firms, each once more producing at its break-even point. Thus the long-run industry supply curve is horizontal.

reached maturity it pays to continue to harvest the apples off the trees even if the price drops substantially and farmers have difficulty meeting their mortgage payments. Even though price may fall far short of what is required to cover operating costs plus the fixed payments on the mortgage used to finance the purchase of the land and the investment in the trees, it may be better to continue to grow apples in the orchard until the trees die off in old age. Provided that minimum average variable cost (primarily harvesting expense) is less than the price of apples, the farmer will lose less by harvesting the apples than by cutting down the orchard. This means that the adjustment of the apple industry to a downward shift in demand may require twenty or more years. In contrast, restaurants can be opened and closed out fairly rapidly. Thus one can expect that the restaurant industry will approach long-run equilibrium much more rapidly than the apple industry.

Size Indeterminacy and returns to scale — an esoteric complication

When deriving total cost function $C(q)$ from the Cobb-Douglas production function in Chapter 5.5.1, we found that the assumption that the production function was homogeneous of degree 1 (constant returns to scale) implied that the firm's average total costs were constant; i.e., the total cost function was of the form $C(q) = kq$. Hence, $dC(q)/dq = C(q)/q = k$. This creates a problem when considering a firm operating in a competitive environment. If the price is above average cost k, the firm will always find it can increase profits by increasing output — equilibrium output is infinite! If the price is below k, the firm should go out of business because it will lose money on every unit no mater how much it sells. And if $p = k$, the size of the firm is indeterminate because economic profits will be precisely zero at any level of output! If there are constant returns to scale, the price will be k under competition, and industry sales will be $Q(k)$. While we can determine industry output under competition with constant returns to scale, we cannot determine either the size of the average firm or the total number of firms in the industry!

We could leave open the question of firm size in a competitive industry, but it may be more reasonable to avoid indeterminacy by abandoning the assumption of constant returns to scale in order to allow for a more complicated cost function.[5] The assumption of constant returns to scale

[5] Alternatively, we could abandon the assumption that the firm can purchases its inputs at a constant price.

does seem plausible: if we double all the ingredients in a cake recipe plus the cook's time, won't we double output? But every college dining hall chef knows that this argument at best goes only so far. Operating on a larger scale permits the use of larger pots and more efficient machinery and the hiring of specialized chefs, all of which contributes to declining average cost. On the other hand, the larger the kitchen the more difficult is the task of coordinating the activity of the multitude of cooks. Thus it may be reasonable to assume that at low levels of output there are increasing returns to scale and declining long-run average costs while at high levels of output production is subject to diminishing returns and increasing long-run average cost. These two tendencies were captured by total cost function (2), as can be seen by looking back at the cost function plots on Figure 6.1.

6.2.4 *Rent and the value of land*

The long-run industry supply curve derived on Figure 6.7 is horizontal, implying that a shift in the industry demand curve would not cause a permanent change in price. This argument relied on the assumption that all firms in the industry, potential entrants as well as existing firms, face exactly the same cost conditions. But almost two centuries ago British economist David Ricardo relaxed this assumption in analyzing the implications of two obvious facts: farmland is in fixed supply and it varies in fertility. As a result, one farm may yield much larger crops than another when the same amount of labor is applied. Since the supply of the best land is limited, poorer and poorer land may be called into production when demand for the crop increases, and this will cause production costs and prices to rise.

This process is explained on Figure 6.8, which for simplicity assumes there are only three grades of farm land. The cost functions on the left are for the most fertile farms. The cost functions in the next panel are higher because they are located on mediocre land. And the cost functions on the third panel are still higher because their soil is even poorer. The poorest farmland will be bought into production only if the price rises above its \$20 breakeven price. The market supply curve is obtained by summing at each price the quantity of corn that will be supplied at that price by all the farms of each type. Because the poorer farms will be brought into production only when the price reaches the point where it profitable, the market supply curve has a positive slope.

The equilibrium price on the graph is equal or above the breakeven price of the fertile and mediocre farms but below the break-even price for

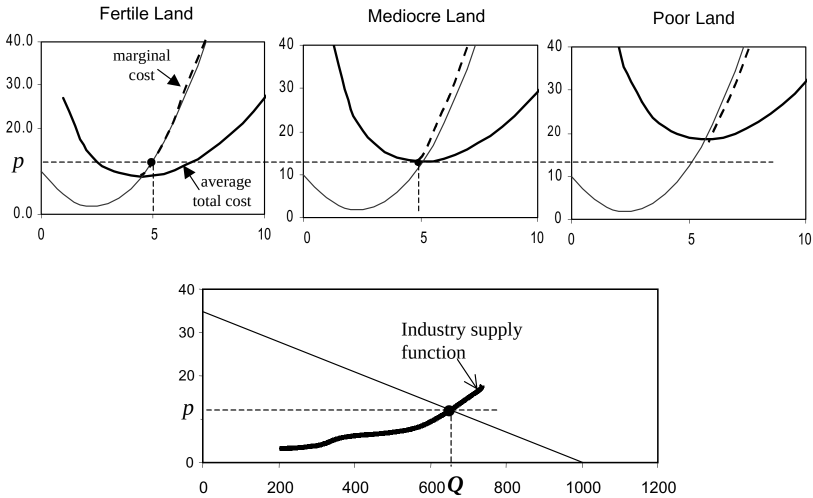


Fig. 6.8. Farmland of differing fertility and the theory of Rent
Each farm produces where price equals marginal cost, provided that price is above its shut-down price. When farmland differs in fertility, more fertile farms with lower than average cost functions will obviously do better than others. Farms with especially good land will yield a positive rental income to any owner who decides to lease out the farm and go to work in the factory or retire.

the poorest farms. The farms with the most fertile land enjoy an excess of market price over average cost, which means that the owner of such land enjoy a surplus over the costs of production equal to the product of quantity produced times the excess of price over average cost. Ricardo chose to call this surplus *rent* rather than profit. This makes a certain amount of sense, for the owner of a more fertile farm who does not wish to continue farming can rent out his farm for more than owners of mediocre or poor farm land would be able to charge. Clearly, this surplus is indeed the maximum amount that a potential tenant would be willing to pay for the privilege of farming the land for one year. Assuming there are a number of potential tenants, they will bid the rent up to this value in their competition for the privilege of farming the land. Note from the graph that the mediocre farmland is cultivated but yields no rent because the price of corn equals minimum average total cost. The poorest farmland is left fallow — no one will till the land because the price of corn is below the breakeven point.

If the market demand curve for wheat were to shift upward, the price of wheat would rise and the rents would increase. Specifically, an expansionary shift in the demand function causes rents to rise on the fertile land; furthermore, the mediocre land will start to yield a rent and even the poor farmland may be brought into production.

Suppose that the annual rent being earned on fertile farms is \$10,000 but that of a mediocre farm is only \$3,000. How will the rent affect the value of the land if a tenant farmer wishes to buy the farm? Suppose that buying a farm will yield in the years to come a stream of future returns, R_t, R_{t+1}, $R_{t+2}, \ldots$ But the return in the future is of less value that this year's return because the value of each future return must be discounted to reflect the time cost of money.

Example

\$100 at the end of one year when the interest rate is $i = 10\%$ is worth only \$100/1.1 = \$90.90 today. Why? Because \$90.90 in the bank at $i = 10\%$ interest would grow to \$100 in one year.

More generally, the value today of a sum R_1 to be received at the end of one year is $V = R_1/(1+i)$. This holds because, if we were given the sum V today we could place it in the bank earning interest rate i and have $(1+i)V = R_1$ at the end of the year. By the same logic, a payment R_2 to be received two years away is equivalent to $R_2/(1+i)^2$ today and the present value of a payment R_t to be received t years into the future is $R_t/(1+i)^t$.

Definition

The value today of a payment to be received in the future is its present value.

Because the farm will be yielding a return into the indefinite future, it is necessary to consider the present value of the entire stream of future payments in deciding how much we should pay to buy the land today. The value of the farm is the sum of all these discounted future rents:

$$V = \frac{R_1}{(1+i)} + \frac{R_2}{(1+i)^2} + \cdots + \frac{R_t}{(1+i)^t} + \ldots \tag{16}$$

This complicated equation for the value of the land can be greatly simplified if we are permitted to assume that the farm will yield the same rent R every year for evermore:, i.e., $R = R_1 = R_2 = \ldots$. Substituting into equation (16) we have

$$V = R(\rho + \rho^2 + \cdots + \rho^t + \ldots)\,, \quad \text{where} \quad \rho = \frac{1}{(1+i)} < 1\,. \tag{17}$$

Since the sum of a geometric series is $S = 1+\rho+\rho^2+\cdots+\rho^n \ldots = 1/(1-\rho)$,[6] we have

$$V = \frac{R}{(1-\rho)} - R = R\left[\frac{(1+i)}{i-1}\right] \tag{18}$$

$$= \frac{R}{i}\,. \tag{19}$$

Example

If the interest rate is $i = 5\%$, potential purchasers of a farm yielding an annual rent of $R = \$10{,}000$ for ever more will be willing to pay a maximum of $V = R/i = \$10{,}000/0.05 = \$200{,}000$.

While this result may not be intuitively obvious, on reflection it makes perfect sense. Note that if a purchaser had to borrow \$200,000 from the bank to make the purchase, the annual interest cost would be \$10,000, which means that the entire rental value of the land would be paid each year to the bank to cover the interest expense. At any price below \$200,000 there would be a positive return over and above the interest cost plus all other economic costs that are reflected in the cost function. In such circumstances, competitive bidding among potential purchasers will push the price of the land up to \$200,000. The price will go no higher, because then the buyer's annual interest costs would exceed the rental income, which is obviously a losing proposition. Now consider a potential purchaser who

[6] To find the sum of the geometric series let us first consider the truncated series:

$$S \;\; = 1 + \rho + \rho^2 + \cdots + \rho \;\; . \tag{i}$$

Multiplying by ρ yields

$$\rho S \;\; - \rho \;\; ^{+1} = \rho + \rho^2 + \cdots + \rho \tag{ii}$$

On substituting (ii) into (i) we have

$$S \;\; = 1 + \rho S \;\; - \rho \;\; ^{+1} \text{ or } (1-\rho)S \;\; = 1 - \rho \;\; ^{+1}\,. \tag{iii}$$

Dividing by $1-\rho$ yields a convenient formula for the sum of the truncated geometric series

$$S \;\; = \frac{(1-\rho \;\; ^{+1})}{1-\rho}\,. \tag{iv}$$

Also,

$$S_\infty = \lim_{\to\infty} \frac{1-\rho \;\; ^{+1}}{1-\rho} = \frac{1}{1-\rho}\,, \text{ provided } -1 < \rho < 1\,. \tag{v}$$

In subsequent chapters we will encounter a number of other applications of the equation for the sum of a geometric series.

has $200,000 saved and does not have to borrow from the bank. The wealthy purchaser will not pay more than $200,000 for the asset because a return of $10,000 can be earned by depositing the $200,000 in the bank at 5% interest. The annual interest cost is the opportunity cost of tying capital up in the land, and that cost cannot exceed the annual rent.

Note that equation (19) implies a rather unexpected result: If interest rates should rise, the value of the farmland will decline.[7] The inverse relationship between the value of land and interest rate arises because a higher interest rate means that the opportunity cost of committing funds to the purchase of the land has increased.

6.3 Monopoly versus competition

The bucolic farm is an icon for the good life, even if the farmer is struggling in the face of competitive market pressures to pay the mortgage and survive. But monopolists, unlike the farmer, are not loved. Writing in the 4th century BC, Aristotle, the great Greek philosopher, condemned the single seller as "unjust." The public objects that monopolists are "robber barons" earning *unfair* profits — this is a matter of *fairness* or *equity*. Most economists have a different objection to monopoly. The economists' complaint is that monopolies contribute to an *inefficient* allocation of resources because they restrict output — they do not produce enough. This is a critical distinction.

- If the problem with a monopolist were only that it earns unfair profits, a tax on the monopolist would seem to be an obvious remedy for this equity issue.
- If the problem with monopoly is that it contributes to an inefficient allocation of resources because it restricts output, then a tax on the monopolist may be unlikely to help and indeed may compound the problem by contributing to a still further reduction in output and even greater inefficiency.

Economists arguing against monopoly are inclined to focus on the efficiency costs. We must compare the model of the monopolist developed in Chapter 5.5 with the model of competition.

[7] This inverse relationship holds more generally, even if the payments are not the same every year, as can be seen from equation (16).

6.3.1 *Inefficient monopoly versus marginal cost pricing*

As was explained in Chapter 5.5, the profit maximizing monopolist, like the competitive firm, produces where marginal revenue equals marginal cost, $dR/dq = dC/dq$. But the fact that the demand curve facing the monopolist is downward sloping implies that $p > dR/dq$. Therefore,

$$p > \frac{dR}{dq} = \frac{dC}{dq}\,. \tag{20}$$

The behavior of the profit maximizing monopolist in the simplest situation of constant average total cost, $C = kq$, is illustrated on Figure 6.9, which elaborates on Figure 5.13. Since marginal revenue equals marginal cost at point e, output of q units maximizes profit. Point a on the demand curve reveals that output q can be sold at price p_m. Profits per unit of output are $p_m - k$, so total profits are $\pi = (p_m - k)q$, or the rectangle (p_m, a, e, k) on the graph. Consumer surplus S_c is represented by triangle (d, a, p_m) with area $(d - p_m)q/2$.

The case for marginal cost pricing

The fact that the profit-maximizing monopolist does not produce the level of output required to maximize the total benefits was recognized by Jules Dupuit [1804–1866], a French engineer, who introduced the concept of consumer surplus in a pioneering contribution to economic theory published in the middle of the 19th century. One possible objective in regulating railroads, electric utilities and other monopolies might be to maximize the total benefit $B(q)$ from producing the commodity, the sum of consumer surplus plus profits: i.e.

$$B(q) = S_c(q) + \pi(q)\,. \tag{21}$$

This is an apparently unbiased objective in that it does not matter how the benefits from producing the commodity are divided between the consumers and the monopolist, just so long as the sum of consumer surplus plus profits is maximized.

Figure 6.9 helps to visualize the case for marginal cost pricing for the simple case in which total costs are $C(q) = kq$. Our profit maximizing monopolist sets quantity where the marginal revenue curve crossed the marginal cost curve, charging price p_m. At this price a consumer who was almost willing to buy a unit of the commodity does without, even though it would have yielded almost p_m in satisfaction; but the cost of providing that unit of the commodity would have been only the marginal cost. Hence the loss from not producing that unit is almost $p_m - k$.

Marginal cost pricing will lead to zero economic profit for the monopolist and a consumer surplus triangle of (d, b, k), which is clearly larger than the sum of consumer surplus plus profits area of (d, a, e, k) when the monopolist was successfully maximizing profit. The difference is the excess burden generated by the monopolist given by the triangle (a, b, e) in Figure 6.9. The *excess burden* (aka deadweight loss) arising from monopoly is analogous to the tax generated excess burden triangle discussed in Chapter 3. No one gets the excess burden, neither the monopolist nor the customers. It is lost because the monopoly sets output below the efficient level of output. In contrast, competition is efficient because each profit maximizing firm is a price taker; hence, maximizing profits for the competitive firm requires adjusting output to the point where marginal cost equals price. Note that for the special situation on the graph, constant average costs and a linear demand function, the monopoly produces only half the efficient level of output![8] For this constant average cost case, the excess burden is equal to the consumer surplus realized under monopoly — it equals half of profits.

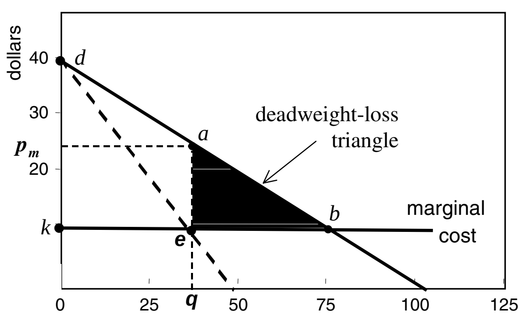


Fig. 6.9. Inefficient monopoly
The monopolist sets output at point e where marginal revenue equals marginal cost. p is the maximum price at which this quantity can be sold. Profit per unit is the excess of that price over average total cost, or the distance $p\ -k$ on the graph. Total profit is quantity sold times profit per unit, or the area of rectangle $p\ , a, e, k$ on the graph. Consumers lose the consumer surplus trapezoid $p\ , a, b, k$. The excess of the lost consumer surplus trapezoid over the gain in profit rectangle is the area of the deadweight-loss triangle a, b, e.

[8] Recall that when the demand function is linear the marginal revenue curve is twice as steep as the demand function.

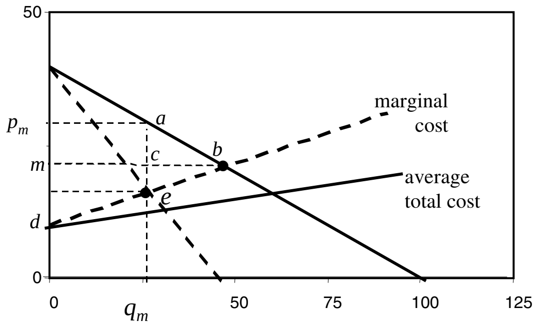


Fig. 6.10. Inefficient monopoly, increasing costs
By charging price p the monopolist attains maximum profits represented by trapezoid p $, a, e, d$. If instead the government prohibited charging more than m, the profit-maximizing monopolist would adapt by producing where price was equal to marginal cost at point b. As a result, consumers would realize a gain in consumer surplus represented by trapezoid p $, a, b, m$. The monopolist would lose profit rectangle p $, a, c, m$ on its initial output of q , but would realize on the increased output a partially offsetting gain in increased revenue over increased cost of c, b, e. The consumer surplus gain from marginal cost pricing exceeds the loss of the monopolist by the triangle a, b, e. This is the deadweight loss triangle (aka excess burden) that results from monopoly pricing.

The relative magnitudes of profit, consumer surplus, and the lost excess burden work out slightly differently on Figure 6.10, which involves increasing costs, even though the demand function is still linear. But now monopoly output is more than half of the efficient level and the excess burden is less than half of profit. This example shows that monopoly profits are not a precise indicator of the magnitude of the dead-weight loss that monopolists generate.

The first step in solving analytically for the level of output q that maximizes the benefit $B(q)$ is to note that consumer surplus is

$$S_c(q) = \int_0^q [p(\tau) - p(q)]d\tau = \int_0^q p(\tau)d\tau - R(q)\,. \tag{22}$$

This result together with the fact that $\pi(q) = R(q) - C(q)$ implies on substitution into (21) that the benefit from producing the commodity is

$$\begin{aligned} B(q) &= S_c(q) + \pi(q) = \int_0^q p(\tau)d\tau - R(q) + R(q) - C(q) \\ &= \int_0^q p(\tau)d\tau - C(q)\,. \end{aligned} \tag{23}$$

Differentiation now yields as a necessary condition for a maximum:

$$\frac{dB(q)}{dq} = p(q) - \frac{dC(q)}{dq} = 0\,. \tag{24}$$

This proves a fundamental proposition:

> Setting output at the point where price equals marginal cost is necessary for maximizing the sum of consumer surplus plus profit.

The excess burden arising from the shortfall of the monopolist's output is

$$B(q_e) - B(q_m) = S_c(q_e) - S_c(q_m) + \pi(q_e) - \pi(q_m)\,, \tag{25}$$

where q_e is the efficient level of output.

Now the reduction in consumer surplus is

$$S_c(q_e) - S(q_m) = [p(q_m) - p(q_e)]q_m + \int_{q_m}^{q_e} [p(\tau) - p(q_e)]d\tau\,. \tag{26}$$

The first term to the right of the equality is the part of the consumer surplus increase that would arise from the reduced price on the units that were purchased from the monopoly and the integral represents the consumer surplus on the additional output that would be produced under marginal cost pricing. Also, the change in profits is $\pi(q_e) - \pi(q_m) = (q_e - q_m)p_e - [p(q_m) - p(q_e)]q_m - C(q_e) + C(q_m)$, where the first two terms reflect the change in revenue and the last two are the change in costs. Further,

$$C(q_e) - C(q_m) = \int_{q_m}^{q_e} \left(\frac{dC}{dq}\right) dq\,,$$

where dC/dq is marginal cost. Rearranging terms we obtain

$$B(q_e) - B(q_m) = \int_{q_m}^{q_c} \left[p(\tau) - \frac{dC(\tau)}{dq}\right] d\tau\,. \tag{27}$$

This reduction in well-being is represented by the triangle a, b, e on Figure 6.10. It results from the fact that the output of the monopolist is too small — it is the excess burden (deadweight loss) inflicted by the monopolist. The existence of this excess burden means that when markets are monopolized, they fail at the task of allocating resources efficiently.

Public utility regulation and the fair rate of return

Governments regulate, and there are two basic strategies for protecting the public from monopoly. One procedure is to discourage mergers of competing corporations or the buying out of competitors when such activities will lead to an excessive concentration of production. And the government has on occasion broken firms up into smaller units. For some commodities this strategy is not feasible because the technology would make multiple firms grossly inefficient. The distribution (but not the generation) of electric power to homes has long been regarded as a natural monopoly, for it would be prohibitively expensive to have more than one set of utility wires supplying electricity to each house. For the same reason, local telephone companies are natural monopolies, although the development of the cellular phone has placed limits on the extent of the monopoly privilege. Natural monopolies arise when the technology is such that having more than one supplier to a market would involve tremendous expense. They arise when the technology involves very high fixed costs relative to marginal cost.

In the U.S. most states have a Public Utility Commission charged with the mission of making sure that natural monopolies do not exploit their advantage to the detriment of the public. They customarily regulate the price of telephone service and the distribution of electricity but they no longer regulate cable television and in many states the generation of electricity has been deregulated. Usually the regulators do not attempt to micro-manage the detail of pricing decisions but rather strive to keep the monopolist from realizing more than a "fair" rate of return on investment. The question of what constitutes a fair rate of return is a contentious issue often leading to conflicting testimony from expert economists at rate hearings. If too low a rate is set, the monopolist will be discouraged from undertaking worthwhile investments in new capacity or more efficient machinery. If too high a rate of return is set, the monopolist will be encouraged to undertake excessive investment, charge too high a price and restrict output.

6.3.2 *Pollution and other externalities*

The argument that monopoly leads to underproduction relies on the assumption that the monopolist's cost function fully reflects the costs of providing the commodity. This assumption will be violated if the production of the commodity generates pollution costs that are not borne by the manufacturer but instead are inflicted upon the general public. Such costs might include the health effects of air pollutants, the discomfort created by foul odors or noise, the damage to paint caused by toxic chemicals, etc. These costs are called *externalities* because they are not taken into account (internalized) by either producers or consumers. Because they do not have to pay for the externalities they generate, producers do not consider these costs in deciding on how much of the commodity to manufacturer and consumers count only the market price and neglect the externalities in deciding how much to buy.

Because neither the producer nor the consumer have to shoulder pollution costs, the market mechanism does not properly take pollution costs into account in determining the level of output. Let $C_p(q)$ denote the total cost of the externality, such as pollution costs. Then equation (23) must be modified to incorporate these costs:

$$B(q) = S_c(q) + \pi(q) - C_p(q) = \int_0^q p(\tau)d\tau - C(q) - C_p(q)\,. \tag{28}$$

At the level of output that maximizes benefit we must have

$$\frac{dB(q)}{dq} = p(q) - \frac{dC(q)}{dq} - \frac{dC_p(q)}{dq} = 0\,. \tag{29}$$

The last term in this equation is marginal pollution cost. We may say that efficiency requires that output be set so that price is equal to the sum of marginal production plus marginal pollution costs.

When the pollution externalities are significant, the competitive market solution — producing at the point where marginal (private) cost equals price — is no longer efficient. The forces of competition lead to $p = dC(q)/dq$, price equaling marginal cost. But output under competition will be too large because market participants neglect the marginal pollution costs imposed on third parties. One remedy is to impose an excise tax on the commodity equal to the marginal pollution cost incurred in its production; i.e., $t = dC_p(q)/dq$. This argument is subject to several qualifications.

First of all, it may be a mistake to subject a polluting monopolist to a tax on the marginal pollution costs it imposes on third parties. It might seem that fairness would require that the monopolist, like the competitive producer, should be subjected to a tax on the pollution that it generates. But efficiency may argue against taxing the polluting monopolist. The reduction in output below the competitive level that occurs under monopoly constitutes a push in the right direction, but there is nothing to insure that the monopoly determined level of output is close to the level that would maximize benefit function $B(q)$ of equation (28). It is certainly possible that even without the tax, the monopolist seeking maximum profits may be restricting output below the optimal level prescribed by equation (29). If so, the tax would contribute to a further reduction in output, which would cause a still greater inefficiency loss. Regulators in the United States are required to demonstrate that the benefits of proposed regulations will exceed compliance costs.

Second, instead of imposing a tax equal to the marginal pollution cost generated by the production of the commodity, it may be better to tax pollutants directly. We could tax the sulfur dioxide, oxides of nitrogen, carbon dioxide, and other pollutants generated by the electric power plant. This might encourage the electric power generator to switch to cleaner fuels or to install scrubbers in its smokestacks.

Third, outright regulation and emission quotas are alternatives to the taxing of pollution activity. For example, the U.S. Acid Rain Program required that electric utilities reduce their sulfur dioxide emissions to 50% of the 1980 emission levels either by installing scrubbers in the smokestacks of their coal-fired generators or by switching to low sulfur fuel. Instead of requiring each coal fired plant to achieve the same reduction, an *emission permit trading* procedure was adopted. Here is how emission-permit trading works: A utility that over fulfills its quota may sell the excess as a tradable emission permit. A utility that is finding it particularly costly to reduce emissions may purchase a tradable emission permit on the open market. The price of permits adjusts in the market to equate demand and supply, which meant that the target reduction in pollution is achieved. But the trading of permits reduces total compliance cost because the greatest share of the emissions reduction is achieved at plants where pollution control is relatively less costly. Trading under the Acid Rain Program creates financial incentives for electricity generators to look for new and low-cost ways to reduce emissions.

6.3.3 *Innovation, patents and dynamic efficiency*

Dynamic efficiency

In a classic defense of monopoly, Harvard Professor Joseph Schumpeter [1883–1950] forcefully defended monopoly in his popular *Capitalism, Socialism and Democracy* [1942]. He argued that the traditional case against monopoly has to do with *static* efficiency. Schumpeter argued that large firms contribute to *dynamic efficiency* in that they have the resources required to innovate the new production processes and new products required for economic progress. Yet it is not necessarily the largest firms in an industry that are the most dynamic innovators. The Schumpeterian hypothesis constitutes an ingenious defense of monopoly, but controversy persists as to whether dominant firms encourage or discourage innovation and progress.

Patent protection

The Constitution of the United States instructs Congress to create a patent system. The holder of a patent receives exclusive monopoly rights on the production of an invention for a period of time, usually twenty years. In addition to awarding *patents*, the government protects intellectual property rights by granting copyright and trademark protection. Patents have an obvious function, for in the absence of the reward of monopoly rights for invention and creativity, who would bother to spend the time developing a better mousetrap or writing innovative computer software?[9] Granting the inventor monopoly rights is not an ideal solution because it leads to the pricing of the creation above its marginal cost of production, which is inefficient. And where is the evidence that twenty years rather than fifteen or thirty year patents would provide the proper balance between appropriate incentive for invention versus the wastes of monopoly pricing? The patent dilemma is particularly heart rendering in the case of lifesaving pharmaceuticals. Without patent protection, profit motivated firms would not invest in the development of new lifesaving drugs. With patent protection the wonder drugs are all too frequently priced out of reach of many who will die without them.

Stephen Shavell, and Tanguy van Ypersele advocate a reward system for innovations as an alternative to patent protection of intellectual property rights.[10] Under a reward system, innovations would pass directly into the public domain rather than being subject to patent protection. Instead of being rewarded with patent protected monopoly power, innovators would be compensated directly by the government. For example, a pharmaceutical company developing a new drug would receive compensation from the government but the drug would sell at a competitive price because other

[9] There are exceptions, such as the development of the Linux computer operating system, but such altruistic counter examples are too infrequent to rely on as the sole source of innovative activity.

[10] "Rewards versus intellectual property rights," *Journal of Law and Economics*, October 2001, pp 525–546.

pharmaceutical companies would be able to freely market the drug without the payment of royalties. Thus the reward system would provide incentives for innovation without conferring monopoly power to the innovator.

Alternatively, the rewards might be optional, the innovator having the choice of accepting the government's reward offer or rejecting it and patenting the product. Innovators would have nothing to lose from the establishment of such a system, because they could always reject the government's reward if they thought it inadequate and rely upon a patent to protect their interest. But if they chose to accept the reward, consumers would benefit from the elimination of the excess burden (deadweight loss) generated by patent-protected monopoly pricing. Or to put it another way, widespread adoption of the innovation would not be discouraged by monopoly pricing.

The principal weakness of this proposal, Shavell and van Ypersele point out, is the difficulty that would be involved in calculating the appropriate reward for the innovator. How large a reward should the government pay to the developer of a Linus computer operating system or to a pharmaceutical company developing a Viagra? Perhaps the government's reward should take the form of a royalty payment based on the total sales of the innovation. Further, tax payers who did not consume the product may object that it is unfair to use their tax dollars to finance the provision of a commodity that they choose not to consume.

6.3.4 *Product differentiation and price discrimination*

The name brand product sells for \$1.29 at your supermarket. The house brand product on the next shelf sells for 89 cents. Which will you buy? Some customers report that they always buy name brand because they want the very best for their family. Other customers, wishing to stretch their dollars as far as possible, always try the house brand first. If the truth were known, sometimes the generic brand and the name brand are made by the same manufacturer in the same plant with the same ingredients, the two products differing only in packaging and price. Does it make sense for a manufacturer to sell the same product at two different prices? The manufacturer may explain that the difference in price arises from differences in advertising and marketing expenditures or quality control. But it turns out that even if there are no differences whatsoever on the cost side, *price discrimination* contributes to profit maximization.

Suppose as the simplest possible example that the market for a firm's product is composed of two distinct types of customers, one market segment

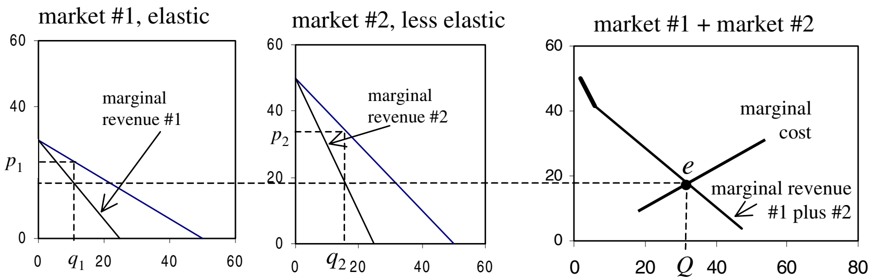


Fig. 6.11. Price discrimination
Profit maximization requires that output in each market segment be at that level at which marginal revenue equals marginal cost. To find the solution graphically, we construct a market graph on the right with a MR curve obtained by summing the quantity supplied in the individual markets at each level of marginal revenue. Total output is where this summed MR curve intersects the marginal cost curve. We market q_1 in market #1 at price p_1 and q_2 at price p_2 in market #2, which means that the marginal revenue in each market is equal to the marginal cost at the level of output identified on the right-hand panel by the intersection of marginal cost with the combined marginal revenue curve.

with inverse demand curve $p_1(q_1)$ and the second market segment with $p_2(q_2)$. Suppose for simplicity that the total cost function for the firm is $C(q_1 + q_2)$, implying that both products cost the same amount to produce. Then profits of our enterprise will be

$$\pi(q_1, q_2) = R_1(q_1) + R_2(q_2) - C(q_1 + q_2)\,, \tag{30}$$

where $R_1(q_1) = p_1(q_1)q_1$ and $R_2(q_2) = p_2(q_2)q_2$. The situation is illustrated on Figure 6.11 where $q = q_1 + q_2$. Note that the customers generating demand curve d_2 are much less sensitive to price.

If the manufacturer has succeeded in setting output for each market so as to maximize total profits, then the partial derivatives of (30) with respect to q_1 and q_2 must both equal zero; i.e., the necessary conditions for profit maximization are

$$\frac{\partial \pi(q_1, q_2)}{\partial q_1} = \frac{dR_1(q_1)}{dq_1} - \frac{dC}{dq_1} = 0$$

and

$$\frac{\partial \pi(q_1, q_2)}{\partial q_2} = \frac{dR_2(q_2)}{dq_2} - \frac{dC}{dq_2} = 0\,. \tag{31}$$

These two conditions imply that prices must be set so that in each market-segment, marginal revenue equals marginal cost. Furthermore, the marginal costs will be the same for both goods if they are identical. All this means that profits are not being maximized unless $MR_1 = MR_2 = MC$. On reflection, this makes perfect intuitive sense, because if the marginal revenue were higher in one market, say #1, more revenue could be obtained without any change in total output by selling one more item in market #1 and one less in market #2. Since total output and hence costs would be unchanged, the increased revenue translates directly into increased profit.

Some examples of price discrimination

- *Pharmaceuticals*: Americans pay more for pharmaceutical products, often much more than citizens in Canada or Britain or France or South Africa. In part the price discrimination may be explained by differences in shipping costs and regulatory burdens, but in part it results from differences in price elasticity. Some critics in the United States argue that the government should stop the outrageous practice by imposing price controls. But if they were forced to charge the same price everywhere, how would the pharmaceutical companies finance their research? And wouldn't they have to abandon their compassionate pricing polices for indigent customers and developing countries? With discriminatory pricing, critics respond, compassionate pricing is not only good public relations; it is profitable as long as the discounted price is above marginal cost. And why should those American citizens requiring pharmaceuticals have to shoulder the cost of developing new products for the whole world?
- *Senior citizen discounts*: Because senior citizens have ample time to seek out bargains, a retailer may find that they constitute a particularly price sensitive market segment. Charging a lower price in this more price elastic market segment contributes to higher profit.
- *Saturday night stopovers*: Empty seats in an airplane constitute a major loss for airlines because the marginal cost of flying with a passenger in that empty seat is negligible, given that the airplane is scheduled to make the trip. Load management is the art of offering incentives to fill otherwise empty seats without reducing the revenue collected from other passengers. Airlines use a variety of strategies in order to discriminate between business travelers and families on vacation. They know that those on vacation are likely to be more sensitive to fare differences than those whose travel expenses are being paid by their employer. Because business travelers typically wish to avoid being away from home on weekends, a reduced airfare for a Saturday Night Stopover may prove attractive only

to tourists. Similarly, advance fare purchase discounts are less attractive to business travelers, who must preserve flexible travel schedules, than they are for tourists who can often schedule their vacations far in advance.

- *College "scholarships"*: Podunk college proudly claims that its need-based scholarship program demonstrates the compassionate nature of the institution. Podunk doesn't discriminate in admissions, but it does have trouble filling its enrollment and tuition revenue targets. The Admissions Office offers partial tuition scholarships (never referred to as tuition discounts) for high need students whose demand for a Podunk college education is much more price sensitive than the demand of wealthier students who do not qualify for financial aid. The size of Podunk's scholarship offers are guided by information from the Educational Testing Service about the financial needs of individual applicants.

Price discrimination works only if the customers can be successfully partitioned into different market segments. National boundaries help. But there can be leakage between segments even when they are demarcated by national boundaries. For example, early in the history of the AIDS epidemic, ACT-up activists noted that DZT, an important drug, sold for $60 in Canada and $26 in England but was marketed in the United States at $100. They confounded the discriminatory pricing practice of the pharmaceutical manufacturer by forming "buyers clubs" to import this and other drugs into the United States from abroad. Later, senior citizen groups organized bus trips to Canada to purchase prescription drugs at prices substantially below those prevailing in the United States.

Even when there is a fair amount of leakage from one market into another, price discrimination may still be profitable. For example, a refrigerator manufacturer may find that the traditional offer of an especially low-price in August attracts many price sensitive customers who carefully shift the time of their purchase to take advantage of the August sale. There is leakage because some price insensitive customers who happen to buy in August also get the discount, but roughly 11/12 of the price insensitive customers buy when the sale price is not in effect.

First degree price discrimination

Carried to its logical extreme, the price discriminating monopolist will sell each unit of the good to the consumer who will pay the most for it. Return to Figure 3.5, which developed the concept of consumer surplus by considering the sale of a good book at a price of \$60. The customer who valued the book at \$100 enjoyed \$40 of consumer surplus! Suppose, however, that the seller told this customer that she had to pay \$99.99. Similarly, the customer who values the book at \$95 is told that he had to pay \$94.99 and so on down to the customer who is just willing to pay the marginal cost of producing the book, say \$20. Then each consumer will buy the book and the monopolist will have successfully extracted almost all the consumer surplus. This strategy will work only if (1) the monopolist can accurately estimate the value of the commodity to each customer and if (2) resale of the commodity by the purchasers can be prevented.

In the case of books, a customer will only want one copy. But suppose we are talking about another type of commodity, such as shoes or wine or cigars. A customer who would pay \$100 for one pair of a favorite brand of shoes might be willing to pay \$75 for a second pair. A monopolist knowing this could offer to sell that customer one pair at \$99.99 but a second pair for \$74.99. Or our monopolist might make an all or nothing offer to this customer: two pairs for \$174.99 or no sale. Either way the monopolist extracts almost all the consumer surplus from each customer. This is known as first degree price discrimination.

First-degree price discrimination may appear to be a particularly vile form of monopoly, but looks can be deceptive. Unlike the simple monopolist who charges every customer the same price, the discriminating monopolist will find it profitable to supply any customer who is willing to pay at least marginal cost. This means that output will be at the efficient level! With price discrimination in the first degree there is no efficiency loss from under production but there may still be a question of equity.

6.4 Monopolistic competition

The model of monopolistic competition is a blend of certain essential features of competition with several features of monopoly. Like the models of competition and monopoly, the theory of *monopolistic competition* assumes that each firm is trying to maximize profits. But the model of monopolistic competition differs in key respects from both monopoly and competition:

- Monopolistic competition differs from monopoly in that there are many firms in the industry.
- Monopolistic competition differs from pure competition because it abandons the assumption that all firms in the industry produce exactly the same commodity.

Product differentiation, the fact that in many industries firms produce similar but not identical products, is the key feature of the monopolistic competitive model. Here are some examples:

- The rental housing market in Middletown is not truly competitive, at least as economists define the term, because it is unusual for any two apartments to be exactly alike. They differ in location as well as in size and amenities.
- Neighborhood supermarkets are differentiated by location and ambience and marketing strategy.
- Gas station operators may tell you that they operate in a hotly competitive environment, but it is not a purely competitive market in the sense in which economists use the term because consumers do not perceive all gas stations as selling precisely the same products. And even if consumers recognized that all gasoline brands are essentially the same, location and convenience would differentiate gas stations.
- Barbershops differ in location, the style of the cut, waiting time and personality.
- Colleges compete for students. Harvard may compete with Yale, Swarthmore and the University of Massachusetts for students, but the educational services provided by these institutions are not identical.

The fundamental implication of product differentiation is that the demand curve facing the individual firm is not horizontal, as with pure competition, but downward sloping, like the demand curve confronting a monopoly. In contrast to a truly competitive environment, a firm selling in a monopolistically-competitive market will not lose all its customers if it slightly raises its price because at least some and perhaps many customers will be willing to pay the higher price rather than consume a substitute product that is similar but not identical to their preferred brand. Because of customer loyalty, the monopolistic competitive firm will not be called upon to service the entire market if it lowers its price slightly below that charged by its competitors. This means that a firm selling in a monopolistic competitive industry does not have to accept a price determined by market forces. It is a *price setter*; like a monopoly, rather than a price taker. Among other complications, the downward sloping demand curve means that the monopolistic-competitive firm typically finds that marginal revenue is below price, like a monopoly. This is in marked contrast to the competitive model, which has every firm facing a horizontal demand curve so that price and marginal revenue are equal.

The model of monopoly does not capture a key feature of monopolistically competitive markets. A true monopoly is the single seller. Because it is the only firm in the industry, the monopoly does not have to worry about the prices charged by other firms in the industry. In contrast, a firm operating in a monopolistically competitive environment is affected by the pricing decisions of other firms selling similar if not identical products. Under monopolistic competition the firm must worry that other firms selling similar products may cut prices, which will cause it to lose some but not all of its customers, unless it also cuts its own price.

6.4.1 *Demand in a monopolistic competitive industry*

Modeling the monopolistic competitive environment requires a more complicated demand function. While we could specify that the demand function faced by the ith firm is $q_i(p_1, p_2, \ldots p_n)$, where p_j is the price charged by the jth firm, a simpler way to proceed was developed by Edward Chamberlin in his treatise on *Monopolistic Competition*, published in 1933. Chamberlin achieved simplification by assuming that the quantity sold by firm i depends only on its own price, p_i and on the average price $\overline{p}$ charged by the other $n-1$ firms in the industry. It is reasonable to suppose:

$$q_i(p_i, \overline{p}, n), \text{ with } \frac{\partial q_i}{\partial p_i} < 0\,, \frac{\partial q_i,}{\partial \overline{p}} > 0\,, \text{ and } \frac{\partial q_i}{\partial n} < 0\,. \tag{32}$$

Thus the sales of the ith firm will decline if it raises its price while all other firms hold their prices unchanged. Conversely, the ith firm will be able to sell more if it holds to its old price when on average the price charged by other firms in the industry moves upward. Furthermore, if additional firms enter the industry, the ith firm is likely to sell less, given prices. All this makes intuitive sense. For example, we would expect Toyota to lose customers if it raises prices when other auto companies do not; but Toyota would gain sales if it held the price line at a time when auto prices generally were moving upward; and Toyota will lose customers if additional firms decide to enter the industry.

As a numerical example, consider the demand function facing a typical monopolistic competitor, call it the ith Corporation:

$$q_i = \frac{100(11 - p_i + 0.75\overline{p})}{n^{0.5}}\,. \tag{33}$$

For instance, if there are $n = 100$ firms in the industry and the average price is $\overline{p} = \$10.00$, the ith company will sell $q_i = 10(11 - p_i + 7.5) = 185 - 10p_i$.

In these circumstances the firm will sell 85 units if it also charges \$10. If the average price increases to $\overline{p}$ = \$16.00, the ith Corporation will be able to sell $q_i = 230 - 10p_i$. This demand function is reasonable in the sense that it satisfies the conditions specified in (32); specifically, $\partial q_i/\partial p_i = -100/n^{0.5} < 0$, $\partial q_i, /\partial \overline{p}) = 75/n^{0.5} > 0$, and $\partial q_i/\partial n = -q/2n < 0$.

The essential features of this demand function have been recorded on the left-hand columns of Table 6.2 and plotted on Figure 6.12. The focus is on the short-run, holding the number of firms fixed at 100. As indicated, our firm can sell more the higher $\overline{p}$, which is the average price prevailing in the industry. For example, if p_i = \$6 and $\overline{p}$ = \$6 our firm will sell 9.5 units. But if $\overline{p}$ = \$12 while our firm continues to charge p_i = \$6 our firm would be able to sell 14. If $p_i = \overline{p}$ = \$12, our firm can sell 8. This rather complicated demand relationship is plotted on the two dimensional graph by drawing three distinct demand curves labeled $d_{\overline{p}=\$6}$, $d_{\overline{p}=\$12}$, and $d_{\overline{p}=\$18}$. Each of these demand curves shows how the sales of our firm will respond to changes in the price p_i it charges, given the specified average price prevailing in the industry, denoted by $\overline{p}$. The steeper curve, labeled *D-D′*, sums up another important point about the demand relationship. It shows how much the ith firm could sell if all the other firms, charged the same price it is charging (e.g., $\overline{p} = p_i$ = \$6.00, $\overline{p} = p_i$ = \$12.00, or $\overline{p} = p_i$ = \$18.00). Of course, if more firms were to enter the industry, all of these demand curves would shift to the left because sales would fall in the face of the increased competitive pressure.

How well a firm does in this market depends in part on the pricing policy of the other firms in the industry. In the monopolistic competitive model it is assumed that the representative firm sets output so as to equate

Table 6.2. Sales, revenue, costs and profits of the ith company.

$\overline{p}$	p	q	$R(p)$	$C(Q)$	Profit
6	3	12.5	38	114	−77
6	6	9.5	57	102	−45
6	9.75	5.8	56	87	−31
12	6	14.0	84	120	−36
12	12	8.0	96	96	0
12	18	2.0	36	72	−36
18	14.25	10.3	146	105	41
18	18	6.5	117	90	27
18	22	2.5	55	74	−19

$q = 100(11 - p + 0.75\overline{p})/n^{0\ 5}$, $n = 100$

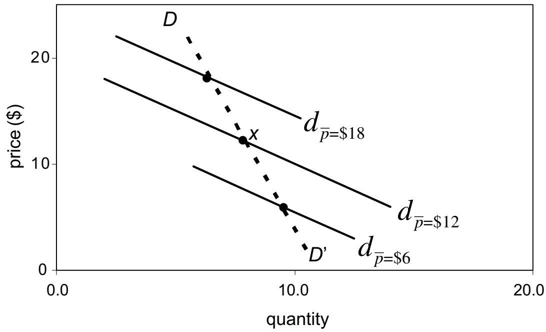


Fig. 6.12. Demand function, monopolistic competition
The $d- = \$18$ demand function shows how much the ith firm can sell as a function of the price it charges, if the average price charged by all the other firms is \$18. Similarly the $d- = \$12$ and $d- = \$6$ demand curves show how much it can sell if the other firms on average charge \$12 or \$6.

The DD' demand curve shows how much our firm could sell if all the other firms were to charge the same price it is charging. For example, point x on the $d- = \$12$ demand curve shows how much the ith firm can sell if it charges \$12 along with all the other firms. That is why the DD' demand curve goes through the \$12 point on the $d- = \$12$ demand curve.

marginal revenue with marginal cost under the simplifying assumption that the firm is so small that changes in its own price will not affect the prices charged by other firms in the industry.

6.4.2 *Equilibrium under monopolistic competition*

The next graph completes the picture by adding the average total cost curve. It is drawn for the total cost function[11]

$$C(q_i) = 64 + 4q_i\,. \tag{34}$$

If the average price charged by all the other firms in the industry is \$12, the best the ith Corporation can do is charge \$12, selling 8 units. This leaves our firm with zero economic profit, but any other point on the $\overline{p} = \$12$ demand function is worse, imposing a loss because average cost will exceed price. As can be seen from Table 6.2, the situation would be worse if the

[11] This form of the cost function, considered in Chapter 5.5, is invoked for its simplicity. Note that this cost function is incompatible with the model of competition because a price-taking firm striving to maximize profit would either produce nothing because $p \leqslant 4$ or else an infinite amount.

average price were $\overline{p} = \$6$, for then the $d_{\overline{p}=\$6}$ demand curve is relevant and the best p_i is \$9.75, but this leaves our unhappy firm with a loss of \$31. Obviously, if $\overline{p}$ were to remain at \$6.00 our firm would have no alternative in the long-run but to go out of business. Inspection of the table reveals that if all firms were to charge \$18, they would all enjoy positive profits. But this is not an equilibrium situation. Acting under the assumption that the other firms would not change from the price of $\overline{p} = \$18$, the ith Company will attempt to increase its profits by cutting its price to \$14.25. Unfortunately, other firms also have an incentive to lower price.

There is an equilibrium where no firm has an incentive to change price, given the assumption that it is so small a force in the market that others will not respond by changing their price when the ith firm adjusts its price. For the numerical example with $n = 100$ the equilibrium occurs at $\overline{p} = p_i = \$12$ — given $\overline{p} = \$12$, the best the ith firm can do is also to charge $p_i = \$12$. Unfortunately, with $n = 100$ firms in the industry, average cost equals price and the representative firm is making zero economic profit.

For the long run, it is usually assumed that there is free entry into the industry, meaning that new firms can enter the industry with essentially the same demand and cost conditions. If firms were losing money, some firms will withdraw from the business, which improves the profit position of remaining firms. If firms are enjoying positive economic profits, new firms will be attracted into the market, which serves to push down the

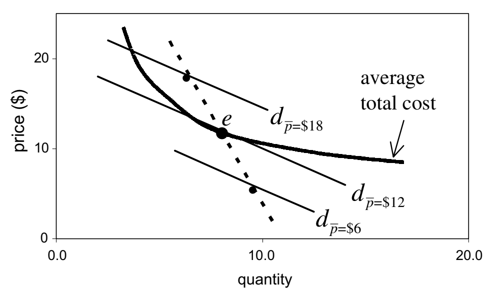


Fig. 6.13. Equilibrium in a monopolistic-competitive market
Monopolistic competitive equilibrium has two essential properties.

- No firm can increase its profits by changing the price it charges, given the price that all the other firms are charging. This is the case for this firm because price would be below average cost, signifying losses, if the firm were to move along the $d- = \$12$ curve by charging either more or less than $p = \$12$, given that all the other firms are charging \$12.
- Profits are zero (p = average total cost) because of free entry and exit.

sales of all firms, given price, and eventually leads to the elimination of economic profit. The situation stabilizes when economic profit is driven to zero, which is the case illustrated on Figure 6.13. To sum up, the long-run monopolistic competitive equilibrium has the following properties:

1. The demand curve facing the firm is downward sloping because of product differentiation.
2. Because the demand curve facing the firm is downward sloping, marginal reveue is less than price.
3. Free entry and exit of firms from the industry drives profits to zero in the long run.
4. In the long run the demand curve confronting the representative firm is tangent to the average total cost curve because the maximum profit is zero.
5. In long-run monopolistic competitive equilibrium, each firm's average total costs are a declining function of output because the average cost curve is tangent to the downward sloping demand curve.

6.4.3 *Is monopolistic competition efficient?*

It may reasonably be argued that the monopolistic competitive model is more realistic than either pure competition or monopoly. Few businessmen believe they face a perfectly horizontal demand curve. Few business firms are free from the competitive pressure of rival firms selling similar products. But it has often been argued that the monopolistic-competitive equilibrium involves an inefficient allocation of resources because the fact that the average cost curve is declining in long-run monopolistic-competitive equilibrium is said to imply that the same industry output could have been produced with fewer resources.[12]

Consider the long-run equilibrium illustrated on Table 6.2 for $p_i = \overline{p} =$ \$12: Each of the 100 firms in the industry is producing 8 units of output, or a total of 800 units. Since the average cost is \$12 per unit, the total industry cost of producing that output is $100 \times 8 \times \$12 = \$9{,}600$. But according to equation (34), a single firm could produce the 800 units of industry output for a total cost of only \$3,264. Thus the monopolistic competitive solution involves an excess resource expenditure of \$6,336! Or to put it another way, the total costs of \$9,600 would suffice to produce

[12]British economist Joan Robinson championed this viewpoint in *Imperfect Competition*, first published in 1933.

an output of $(9{,}600 - \$64)/4 = 2{,}384$ units, almost three times as much. Doesn't this suggest inefficiency?

This numerical example is not a special case. Since the demand curve facing the representative firm is tangent to the average cost curve under monopolistic competition, the average cost curve is downward sloping. That is why it must always be possible to produce the same industry output at lower total cost than that achieved under monopolistic competition. The objection is that monopolistic competition leads to too many gas stations, too many drug stores and too many brands of soap. The proliferation of excessive numbers of retail outlets and product brands wastes resources, causes higher prices, and may pollute the environment.

Is excessive product variety really another form of market failure justifying government intervention in the marketplace? The argument might serve to justify government regulation of the number of gas stations, liquor stores, and banks in a community. Indeed, bankers try to limit the entry of new banks into their community by warning about the dangers of "over banking." And liquor store owners are not unhappy when the local ministry tries to keep new liquor stores from being established in town. But the argument against monopolistic competition should not be overdrawn. A variety of different brands can be a blessing for it allows the consumer greater choice. Greater product variety catering to the diversity of consumer tastes may be worth the additional resources and higher price. All this means that it is not necessarily the case that monopolistic competition causes excessive product variety; indeed, it is possible for a monopolistic competitive equilibrium to involve too little variety.

The monopolistic competitive market will work out differently if a price floor can be enforced. Such floors may be established by government authority. For example, the State of Connecticut's Commission of Consumer Protection used to prohibit stores from selling refrigerators and other appliances at less than wholesale plus 6%. And in several states, liquor control commissions enforce minimum prices for wine and liquor. Price floors have also been established by private agreement among sellers, such as the Organization of Petroleum Exporting Countries (OPEC). If there is free entry, a price floor will only temporarily yield positive economic profit. New firms attracted into the industry by positive profits will cut into the sales of existing firms until a long-run equilibrium is established with zero profit. Thus the unintended consequence of price floors is to establish greater product variety at a higher price.

6.5 Oligopoly

The simplest model of monopolistic competition assumes that there are so many firms in the market that it is reasonable for each seller to conjecture that the average price prevailing in the market will not be influenced by its own pricing decision. This might be reasonable for the rental housing market in New York City. But in many industries the market is dominated by only a few sellers. For example, in the United States the four largest companies ship 86% of all electric light bulbs and 84% of cars are produced by the four largest car manufacturers. When there are only a few firms in the industry, each firm must price strategically. That is to say, each firm must worry about how other firms in the industry will respond to its pricing decisions.

Definition

An *oligopoly* is an industry that is dominated by a small number of sellers.[13]

Because oligopolistic markets are hard to analyze, it is helpful to focus attention on the special case of duopoly, which refers to an industry dominated by two firms.

6.5.1 *Case 1: Duopoly with identical products*

Suppose that two firms produce exactly the same product. Only one price can prevail in the marketplace. To see why, suppose that one firm charges a higher price than the other. No one will buy from the high-price firm unless the low-price firm cannot meet demand because of a capacity constraint. But if the demand for the output of the low-price firm exceeds capacity, it cannot be maximizing profits because it could have sold the same quantity at a higher price. All this means that the assumption that consumers and producers are rational implies that the law of one price must prevail when identical products are sold in the same market.

The French economist Auguste A. Cournot analyzed this problem in 1838.[14] A simple numerical example will illustrate Cournot's explanation of how such a market works. To be concrete, suppose that the demand for

[13] Oligopoly is from the Greek *oligos* meaning few + pō*lein* for seller. Similarly, monopoly is formed with *monos* meaning single and duopoly with the Latin *duo* meaning two.

[14] A. A. Cournot's *Recherches sur les principes mathématiques de la théorie des richesses*, published in 1838, is one of the first serious works in mathematical economics.

the industry output is

$$Q(p) = 200 - p \tag{35}$$

and that the cost function for each firm is

$$C_i(q_i) = 10q_i\,, \quad i = 1, 2\,. \tag{36}$$

Then the price that will prevail in the marketplace will be $p = 200-(q_1+q_2)$ and total profits of the first firm will be:

$$\pi_1 = [200 - (q_1 + q_2)]q_1 - 10q_1\,. \tag{37}$$

Similarly,

$$\pi_2 = [200 - (q_1 + q_2)]q_2 - 10q_2\,. \tag{38}$$

The problem for Firm #1 is to determine the best quantity to market, but that obviously depends on what Firm #2 decides to do. And the problem for Firm #2 is to determine the best quantity that it should bring to market, but this depends on what Firm #1 decides to do. What will happen?

The *Cournot solution* to the oligopoly problem is to assume that each firm acts under the mistaken conjecture that regardless of what quantity it produces, the other firm will continue to market the same quantity as it is currently selling. To find out how much Firm #1 will market under this conjecture, let us find the profit maximizing level of output by setting the partial derivative of equation (37) with respect to q_1 equal to zero:

$$\frac{\partial \pi_1}{\partial q_1} = 190 - 2q_1 - q_2 = 0\,. \tag{39}$$

Hence,

$$q_1 = \frac{(190 - q_2)}{2}\,. \tag{40}$$

This *reaction function* of Firm #1 tells us how much Firm #1 will market as a function of the quantity that Firm #2 is currently selling. Figure 6.14 and the accompanying table clarify the situation. The output of Firm #2, which is taken as given, is plotted on the abscissa and that of Firm #1 on the ordinate. The heavy dashed line labeled $R_1(q_2)$ represents Firm #1's reaction function. It will help in interpreting this function to consider the iso-profit curves plotted on the same graph. Each of Firm #1's iso-profit curve shows output combinations yielding the specified profit levels. The $\pi = 0$ iso-profit curve includes all the output combinations yielding zero

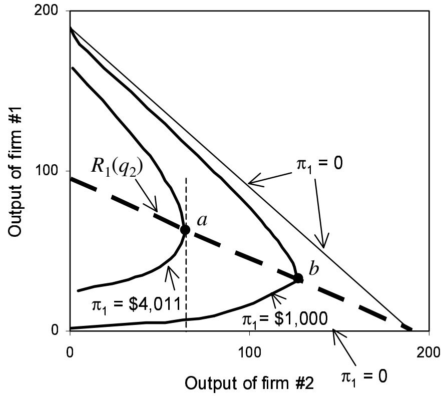


Fig. 6.14. A duopolist's reaction function, identical products
Each point on the $\pi_1 = \$4{,}011$ iso-profit curve yields profits of $4,011 to firm #1. At point a firm #1 would obtain maximum profits of $4,011, given that firm #2 is producing 63. At point b, the best firm #1 can do is make profits of $1,000, given the output of firm #2.

profits. This includes all points $\langle q_2, 0\rangle$ because the firm's profit will be zero if it sells nothing, regardless of the output of Firm #2. It also includes all points on the line with slope -1 for which $q_1 + q_2 = 190$, for then the price will be 10, which only covers average cost and leaves nothing for profit. At the opposite extreme, point $\langle 0, 95\rangle$ is the best possible point from Firm #1's viewpoint. At this point profit for Firm #1 is $9,025. As Table 6.3 reveals, at this level of output Firm #1 maximizes its total profit, given that Firm #2 produces nothing at all; i.e., Firm #1 is fully exploiting a monopoly position.

In order to appreciate the dynamic implications of this model, suppose that Firm #1 had been enjoying a monopoly of the market, maximizing its profits by selling 95 units. When Firm #2 enters the market, it will sell 47.5 units in order to maximize its profits, providing it assumes that Firm #1 will not change its level of output.[15] This is point b on Figure 6.15, where Firm #2's reaction function is plotted as the reverse image of Firm #1's reaction function. But then Firm #1, making the same assumption, will market 71.3, as specified by equation (40). As shown by successive rows near the bottom of Table 6.3 (rows numbered 1 to 7), the market will

[15]Note that Firm #2 has the same reaction curve as Firm #1 (i.e., equation (36) with the subscripts reversed).

Table 6.3. Cournot duopolists (identical products).

	q_1	q_2	q_1+q_2	price	Firm #1 revenue	Firm #1 costs	Firm #1 profit	Firm #2 revenue	Firm #2 profit	Joint Profit
Three examples:										
	50.0	50.0	100.0	100.00	5000	500	4500	5000	4500	9000
	48.0	48.0	96.0	104.00	4992	480	4512	4992	4512	9024
	95.0	95.0	190.0	10.00	950	950	0	950	0	0
Some points on firm #1's $1,000 iso profit curve										
	158.7	25.0	183.7	16.30	2587	1587	1000	408	158	1158
	132.4	50.0	182.4	17.55	2324	1324	1000	878	378	1377
	105.5	75.0	180.5	19.48	2055	1055	1000	1461	711	1711
	77.0	100.0	177.0	22.98	1770	770	1000	2298	1298	2298
	40.0	125.0	165.0	35.00	1400	400	1000	4375	3125	4125
Some points on firm #1's reaction function: $q_1(q_2)$										
	95.0	0.0	95.0	105.00	9975	950	9025	0	0	9025
	63.3	63.3	126.7	73.34	4645	633	4011	4644	4011	8022
	47.5	95.0	142.5	57.50	2731	475	2256	5463	4513	6769
	0.0	190.0	190.0	10.00	0	0	0	1900	0	0
Convergence to equilibrium										
1	*95.0*	*47.5*	142.5	57.50	5463	950	4513	2731	2256	6769
2	*71.3*	*59.4*	71.3	69.38	4943	713	4230	4119	3525	7756
3	*65.3*	*62.3*	127.7	72.34	4725	653	4072	4510	3887	7959
4	*63.8*	*63.1*	126.9	73.09	4665	638	4027	4611	3980	8006
5	*63.5*	*63.3*	126.7	73.27	4650	635	4015	4636	4003	8018
6	*63.4*	*63.3*	126.7	73.32	4646	634	4012	4642	4009	8021
7	*63.3*	*63.3*	126.7	73.33	4645	633	4011	4644	4011	8022
Joint profit maximization, market equally divided										
	47.5	***47.5***	95.0	105.00	4988	475	4513	4988	4513	9025

gradually converge to a new equilibrium where each firm is selling 63 1/3 units. Profits have deteriorated as a result of the entry of the new firm, but the fall in price is an obvious boon for consumers. This new equilibrium is an example of a Nash Equilibrium — each firm is maximizing its profits, given what the other firm is doing.

One possible response to the deteriorating profit condition is for the two firms to agree to divide the market. If each firm sells 47.5 units, industry profits will once more be maximized at $9,025. Such collusion may be illegal; but even if the duopolists are unconstrained by the antitrust laws, both will be tempted to increase their output in a marketing war, thereby pushing the system back toward the Nash equilibrium. An alternative strategy is consolidation, one firm buying out the other in order to have complete control of the market.

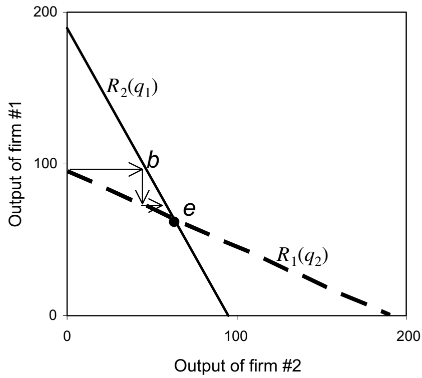


Fig. 6.15. Duopoly dynamics
$R_1(q_2)$ and $R_2(q_2)$, the reaction functions of firms 1 and 2, are both plotted on the graph. They are symmetrical because the cost and demand conditions are the same for the two firms.

Point e, with coordinates $\langle 63.3, 63.3\rangle$, is where the reaction functions intersect. As the example on the bottom of Table 6.3 illustrates, if initially firm #1 is a monopolist selling 95 units, when firm #2 enters the market it will want to sell 47.5 units, under its assumption that firm #1 will not change its output. Firm 1 reacts by cutting its output to 71.3, which leads firm #2 to increase its output to 59.4, and so on as the system converges to point e. Point e is a stable equilibrium: because the system will iterate to point e from any initial set of outputs, provided that each firm persists in the naïve assumption that the other firm will not change its output.

6.5.2 *Case 2: Duopoly with product differentiation*

Instead of assuming that the duopolists produce identical products, Joseph Bertrand [1822–1900], an eminent French mathematician, considered the case in which the products are differentiated. Unlike the preceding case of identical products, Bertrand's assumption of product differentiation means that the two goods may sell at different prices. The problem is similar to the case of monopolistic competition, except that now there are only two rather than a large number of firms in the industry. Each firm sets its own price, but how much it will sell depends in part on the price charged by the other firm. It makes sense for duopolists to focus on price because this variable is more easily observed than the quantity sold by one's competitor. Bertrand solved the problem under the assumption that each firm acts under the mistaken conjecture that the other firm will not change its price.

Suppose that Firm #1 faces the demand function

$$q_1 = \max\left(100 - p_1 + \frac{p_2}{2}, 0\right), \quad p_1 \geq 100 + \frac{p_2}{2}. \tag{41}$$

If $p_1 \geq 100 + p_2/2$, sales will be zero because our firm will have priced itself completely out of the market. The revenue of Firm #1, assuming it does not price itself out of the market, will be

$$R_1(p_1, p_2) = p_1 q_1 = 100p_1 - p_1^2 + \frac{p_2 p_1}{2}. \tag{42}$$

If Firm #1 has total cost function $C(q_1) = 10q_1$ then its profits will be

$$\begin{aligned} \pi_1(p_1, p_2) &= R_1 - C_1 = 100p_1 - p_1^2 + \frac{p_2 p_1}{2} - 10\left(100 - p_1 + \frac{p_2}{2}\right) \\ &= 110p_1 - p_1^2 + \frac{p_2 p_1}{2} - 5p_2 - 1000. \end{aligned} \tag{43}$$

Bertrand assumed that Firm #1 acts under the mistaken conjecture that Firm #2 will not change its price. Then $\pi_1(p_1, p_2)$ will be maximized, given p_2, by setting output at that level where

$$\frac{\partial \pi_1(p_1, p_2)}{\partial p_1} = 110 - 2p_1 + \frac{p_2}{2} = 0. \tag{44}$$

Therefore,

$$p_1 = 55 + \frac{p_2}{4}. \tag{45}$$

This reaction function shows how the price Firm #1 charges is affected by the price charged by the other firm. It is the curve labeled $p_1 = R_1(p_2)$ on Figure 6.16. But we cannot determine what p_1 will prevail until we know p_2.

To keep things simple, suppose that Firm #2 is the mirror image of Firm #1. That is to say, suppose it has demand function

$$q_2 = 100 - p_2 + \frac{p_1}{2} \tag{46}$$

and total cost function $C(q_2) = 10q_2$. Then by symmetry, our second firm's reaction function will be the mirror image of Firm #1's.

$$p_2 = 55 + \frac{p_1}{4}\,. \tag{47}$$

This reaction function is plotted as $p_2 = R_2(p_1)$ on Figure 6.16.

There is a unique set of equilibrium prices, p_1^e and p_2^e, that can prevail in this market, as identified on Figure 6.16 by intersection point e where the two reaction functions intersect. At this point Firm #1 is on its reaction curve; because it is maximizing its profit, given p_2^e (the price charged by Firm #2), Firm #1 has no incentive to change price. Furthermore, Firm #2 is on its reaction curve, given that Firm #1 is charging price p_1^e; therefore, this firm will not change its price. The equilibrium is found by solving equations (45) and (47) simultaneously:

$$p_i^e = 55 + \frac{p_2}{4} = 55 + \frac{(55 + p_1/4)}{4} = 73\frac{1}{3}\,. \tag{48}$$

Obviously, $p_2^e = p_1^e$. Substituting into the demand equation reveals that $q_1 = 100 - 0.5 \times 73\ 1/3 = 63\ 1/3$. At this level of output, revenue is \$4,644, costs are \$633 1/3 and profits $\pi_1 = \pi_2 =$ \$4,010 2/3.

This argument requires that each firm continue to assume that changes in its own price will not elicit a price reaction from its competitor. Could both firms make more profit if they were to wise up and agree to charge a higher price? While such collusion might be illegal, the firms may try to avoid detection by holding their discussions in secret. Alternatively, one

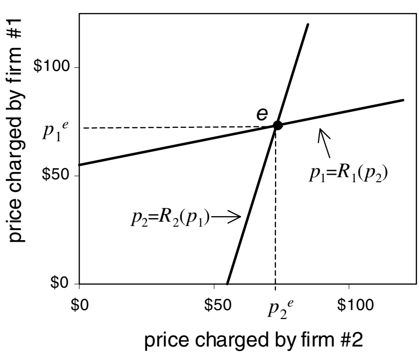


Fig. 6.16. Reaction functions, differentiated duopoly
Firm #1's reaction function, $p_1 = R_1(p_2)$ reveals the price it will charge as a function of the price charged by the other firm. Similarly, Firm #2's reaction function reveals the price it will charge as a function of the price charged by the other firm.

Point e identifies the equilibrium prices that will be charged by the two firms. Table 6.4 illustrates how the system converges to equilibrium.

firm might buy out the other in order to be able to set both p_1 and p_2. To find out the effect of such anticompetitive activities, consider the problem of maximizing joint profits:

$$\begin{aligned}\pi(p_1, p_2) &= \pi_1(p_1, p_2) + \pi_2(p_1, p_2) \\ &= 110p_1 - p_1^2 + 0.5p2p_1 - 5p_2 - 1000 \\ &\quad + 110p_2 - p_2^2 + 0.5p_2p_1 - 5p_1 - 1000\,. \end{aligned} \tag{49}$$

For joint profit maximization we must have

$$\frac{\partial \pi(p_1, p_2)}{\partial p_1} = 105 - 2p_1 + p_2 = 0 \tag{50}$$

and

$$\frac{\partial \pi(p_1, p_2)}{\partial p_2} = 105 - 2p_2 + p_1 = 0\,. \tag{51}$$

Hence, the price with collusion or merger will be $p_1 = p_2 = \$105$. Then $q_1 = q_2 = 47.5$ and the profits realized by each colluder will be \$4,513. This is a much happier result for both firms, but each party to the collusion will have an incentive to cheat.

Fortunately for consumers, duopolists find it difficult to maintain a favorable pricing structure. Table 6.4 shows how the pricing structure will deteriorate if each firm persists in the simplistic notion that the other firm will continue to charge its current price. Initially each firm is charging $p_1 = p_2 = 105$, which means that joint profits are being maximized. But Firm #2 recognizes that it can increase its profits by cutting the price to \$81.25, given that $p_1 = 105$ (this is a point on Firm #2's reaction curve). But Firm #1 responds in turn by cutting its price to \$75.31, which maximizes its profits *if* Firm #2 continues to charge \$81.25. Successive rounds of

Table 6.4. Duopoly pricing ∼ differentiated products.

	p_1	p_2	q_1	q_2	C_1	C_2	π_1	π_2	$\pi_1 + \pi_2$
Joint Profit Max	105.00	105.00	47.50	47.50	475.00	475.00	4,513	4,513	9,025
#2 undercuts	105.00	*81.25*	35.63	71.25	356.25	712.50	3,384	5,077	8,461
#1 undercuts	*75.31*	*73.83*	61.60	63.83	616.02	638.28	4,023	4,074	8,097
#2 undercuts	*73.46*	*73.36*	63.23	63.36	632.25	633.64	4,012	4,015	8,027
#1 undercuts	*73.34*	*73.34*	63.33	63.34	633.27	633.35	4,011	4,011	8,023
Nash Equilibrium	*73.33*	*73.33*	63.33	63.33	633.33	633.33	4,011	4,011	8,022

price-cutting will drive the price down to $p_1 = p_2 = \$73.33$. The downward movement will continue until an equilibrium is reached in which neither firm can improve its position, given what the other firm is charging.

This argument rests on the assumption that firms make rather naïve conjectures about the pricing strategy of their competitors. There are numerous variations on this theme. One possibility is that a price chiseling firm may try to keep its competitor in the dark by holding the list price at \$105 while secretly offering its customers price concessions, secret rebates, or hidden kickbacks. A more astute competitor may try to educate its rival by pushing its price to \$105, hoping that the other party will be smart enough to follow suit. Or if the antitrust authorities don't object, one firm may buy out the other in order to be able to preserve the joint profit maximizing price.

6.5.3 *Game theory*

Game theory, despite its name, is a serious and complex subject. Because the discipline examines the behavior of decision makers whose decisions affect each other, "Interactive Decision Theory" might be a more accurate description of this line of research. The term "Game Theory" stems from the resemblance of interactive decision problems to such parlor games as bridge, poker, monopoly and chess. Developed early in the 20th century, it became a serious topic for economists when mathematician John von Neumann [1903–1957] and economist Oskar Morgenstern [1902–1977] published in 1944 a most significant book, *The theory of games and economic behavior*.[16] The revolution was slow in coming, but now this book is regarded as a major economic classic. Game theory has been applied to the study of a wide range of phenomenon, including the cold war arms races, military strategy, and voter behavior. Oligopoly behavior is an important economic application of the theory of games.

[16] Oskar Morgenstern had immigrated from Vienna to the United States, where he taught at Princeton University. His co-author, Hungarian immigrant John von Neumann, was a brilliant mathematical physicist who made important contributions to growth theory in economics and to computer science as well as game theory. He participated in the development of the atomic bomb and served on the Atomic Energy Commission. His daughter, Professor Marina von Neumann Whitman, was appointed by President Nixon to serve on the President's Council of Economic Advisers.

The prisoner's dilemma

A simple but non-economic example illustrating what game theory is about is provided by the famous prisoners' dilemma:

> Two nefarious criminals, Abel and Baker, have been arrested for a serious felony. The prosecutor, worried that he may not have enough evidence to convict, hopes to persuade them to confess. He isolates them in separate cells in order to prevent them from communicating with each other. He tells Able that if Baker confesses and he does not, he will get 10 years. However, if he confesses but Baker does not, Able will get only 1 year. He also tells Able that if they both confess they will each get four years time. If neither confesses, both will get two years on a lesser charge. Baker is presented with the same options.

The following *payoff matrix* provides a convenient summary of the situation. Abel's two options are listed in separate rows; Bakers options are in columns. There are four payoff cells of data corresponding to the four possible outcomes. The first entry in each payoff cell indicates the payoff to Abel; the second *italicized* entry in each cell is Baker's payoff.

Table 6.5. The prisoners dilemma payoff matrix.

	Baker Confesses	*Baker Denies*
Abel Confesses	4 years, *4 years*	1 year, *10 years*
Abel Denies	10 years, *1 year*	2 years, *2 years*

Clearly, each criminal will minimize jail time if neither confesses. As indicated by the lower right-hand cell of the matrix, each will get two years. But if Abel denies the crime, he runs the risk that his partner may confess in the hopes of serving only one year, in which case Abel is in for ten, and he knows that there is no honor among thieves. What to do?

This payoff matrix illustrates two fundamental game theory concepts:

- A *minimax strategy* is one that minimizes your maximum loss.
  At worse, Abel could get 4 years if he confesses; if he does not confess, at worse he would get 10 years. If Abel is a minimaxer, he will confess in order to have the smallest maximum loss of four years. If Baker is also a minimaxer, he too will confess and they will both get 4 years in jail.

- In a *Nash equilibrium*, neither player can improve his position, given what the other player is doing.[17]

As can be seen from the matrix, both players pleading not guilty is not a Nash equilibrium because at least one of them can do better by confessing to the crime. Both players confessing does constitute a Nash equilibrium. If your partner in crime confesses, you had better have confessed as well, because 4 years is better than 10.

It seems clear that the prosecutor has artfully constructed the payoffs to the criminals so as to give them every inducement to confess. While the Nash equilibrium is sub-optimal, it takes an extreme act of faith on the part of each prisoner to end up with both refusing to confess and thereby escaping with only a two-year sentence.

Table 6.6. Duopoly pricing payoff matrix, differentiated products.

	Firm #2's price = $105	*Firm #2's price = $73.33*
Firm #1's price = $105	$4513, *$4513*	$3008, *$5014*
Firm #1's price = $73.33	$5014, *$3008*	$4011, *$4011*

Oligopolistic pricing

As our first economic application of game theory, consider the two duopolists who were selling differentiated products in Section 6.5.2. Table 6.6 summarizes the payoffs, where we simplify the problem by assuming only two possible prices. If Firm #1 pursues a maximin strategy, it will set its price at $73 because then it is guaranteed a profit of at least $4,011. Similarly, the maximin strategy for Firm #2 is to charge $73. This payoff matrix has a Nash Equilibrium in the lower right-hand corner where each firm charges a price of only $73.33, which is less profitable than if both charged $105. Will our two firms be any more successful than prisoners Able and Baker at avoiding the Nash equilibrium?

There is an essential difference between this problem and the prisoner's dilemma. That problem was a one-time thing, for it is most unlikely that our two prisoners will find themselves in the same position again. However,

[17] This concept is named after John Nash, who earned the Nobel Prize for his fundamental contributions to game theory. *A Beautiful Mind*, the movie version of his biography by Sylvia Nasar describing his life long struggle with schizophrenia, won four Academy Awards in 2002.

the duopolists make pricing decisions over and over again — they are playing a *repeated game*. In a repeated game there is some hope for signaling one's rival that one is willing to maintain price discipline. Firm #1 may continue to charge $105 in the hope that Firm #2 will imitate, but this may lead Firm #2 to play its rival for a sucker and continue to undercut. Sometimes a *tit-for-tat strategy* is tried: Firm #1 always charges the current period whatever price Firm #2 charged in the immediately preceding period. Once Firm #2 recognizes that Firm #1 is playing tit-for-tat, it should obviously push its price to $105.

Advertising strategy

A major brewery is considering the advisability of adjusting its advertising strategy. Market research has revealed that in this industry advertising steals customers from other firms in the industry but has little effect on total industry sales. That is to say, the primary effect of our firm's advertising is on its market share. If both firms simultaneously cut their advertising expenditure, each would sell almost the same quantity; and the dramatic reduction in advertising expenditure would mean that their profits would increase substantially. The danger is that Brand B will grab a substantial increase in market share by continuing to spend big bucks on advertising.

The payoff matrix on Table 6.7, summarizes the situation. Observe that this matrix is asymmetric — game theory does not require that the two decision-makers must be identically situated. Also, Brand B has a *dominant strategy*: no matter what advertising strategy Brand A follows, Brand B will make more if it has high advertising expenditure. If Brand A has low expenditures, brand B will make the most of the situation with high spending. Or if Brand A advertises high it will again pay Brand B to advertise high. This means that Brand A should assume that Brand B will follow its dominant strategy and spend substantial sums on advertising; and if Brand B is a high advertiser, Brand A must advertise high as well.

This outcome, while predictable, is far from optimal. If both firms were to cut advertising, each firm would make greater profits. But this happy

Table 6.7. Liquor advertising payoff matrix (millions of $).

	Brand B ∼ Low Ad Budget	*Brand B ∼ High Ad Budget*
Brand A Low Ad Budget	7, *7*	3, *10*
Brand A High Ad Budget	6, *3*	4, *5*

situation is likely to be unstable because it is not a Nash equilibrium: either firm can do better, given what the other is doing, with a high advertising budget. One possible resolution will be to lobby the government to prohibit the advertising of all alcoholic beverages on television! The politicians might please the clergy as well as increase profits for the breweries, but the television industry would miss the advertising revenue.

6.6 Antitrust action

Antitrust laws are designed to regulate monopoly and limit anti-competitive activities in order to make the market mechanism function more effectively. In the United States the aggressive actions of the so-called "robber barons" — such as financier J. P. Morgan, petroleum king John D. Rockefeller, and rail magnate W. H. Vanderbilt — led to the passage of the first major antitrust legislation in the last years of the 19th century. In 1887 Congress established the Interstate Commerce Commission to regulate the railroads. This was quickly followed by the passage in 1890 of the of the Sherman Antitrust Act. The Clayton Act of 1914 was designed to clarify the provisions of the Sherman Act and to establish the Federal Trade Commission. It also attempted to protect the organizational activities of labor unions by exempting collective bargaining agreements from antitrust scrutiny. Today both the Department of Justice and the Federal Trade Commission attempt to control anti-competitive activities. Their efforts are supplemented by antitrust action at the state level.

The Sherman Antitrust Act was designed to "protect trade and commerce against illegal restraints and monopolies." It contains two provisions of fundamental importance:

Section 1: Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is hereby declared illegal.

Section 2: Every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations, shall be deemed guilty of a misdemeanor.

Violating the Sherman Act is a serious crime. Corporations violating the act have had to pay substantial fines. Corporate officials participating in such crimes may have to pay fines and serve jail time.

The language of the Sherman Act was ambiguous and open to changing interpretation by the courts. A brief review of some of the more interesting court cases reveals how antitrust policy in the United States has been influenced by the changing attitudes of congress, by who occupies the White House, and by the decisions of the courts:

1895 The U.S. Supreme Court ruled that although the E. C. Knight Company controlled 98 percent of the sugar refining industry, it had not violated the Sherman Act because the company did not engage in interstate commerce in that the processing was done locally.

1911 The Supreme Court ruled that John D. Rockefeller's Standard Oil, which controlled 91% of the petroleum refining industry, was guilty of violating the Sherman Act. The remedy was to break up the company into 39 smaller firms. That is why we now have Standard Oil of Ohio, Amoco, Chevron, Exxon, etc.

1911 The American Tobacco Company, which controlled between 75% and 90% of the market for tobacco products, was also judged guilty of violating the antitrust laws and broken up into smaller companies that would compete against each other.

1914 Congress established the Federal Trade Commission to investigate the business practices of companies engaged in interstate commerce. The FTC can issue cease and desist orders when it finds behavior in violation of the law.

1975 Goldfarb v Virginia Bar Association: When the Goldfarb's went to purchase a house, each of the 36 lawyers they telephoned quoted the identical charge of 1% of the purchase price for handling the sale. In Virginia at that time, and indeed throughout most of the United States, either a state or a county bar association imposed a minimum fee schedule. The minimum charges for arranging a simple divorce, drawing a will, or representing the buyer of a house were spelled out in complete detail. Bar association grievance committees often focused their energies on disciplining attorneys who habitually priced below the fee schedule. All this was changed in 1973 when the U.S. Supreme Court ruled in the case of Goldfarb v Virginia Bar Association that the minimum fee schedules constituted a restriction on interstate commerce in violation of the Sherman Antitrust Act. As a result of that court decision, the market for legal services became much more competitive. Lawyers complained that clients shopped by phone for the least expensive lawyer, and Adam Smith's invisible hand pushed down legal fees.

1982 AT&T corporation (aka "Ma Bell") at one time had a virtual monopoly on local and long distance telephone service in the United States Under the 1982 decision. AT&T had to divest itself of all local telephone operating companies (the Baby Bells'), forming 22 separate companies.

1991 The U.S. Department of Justice's Antitrust Division sued MIT and eight "Ivy League" colleges and universities for engaging in a conspiracy to restrain price competition for students receiving financial aid. The schools replied that the Sherman Act did not apply to them because they were not-for-profit institutions. Further, they argued that their cooperative behavior facilitated the allocation of financial aid to students who were most in need. The Ivy League schools signed a consent decree agreeing to stop the challenged activity. While a consent agreement may settle a case without admission of guilt, it carries the force of law with respect to future actions. MIT refused to sign and was found guilty of violating the Sherman Act, but the decision was overturned when appealed to a higher court.

1996 Archer-Daniels-Midland Co., the Decatur, Illinois grain processing giant, paid the U.S. government a $100 million antitrust fine in 1996 for rigging prices of lysine and citric acid. It was fined $45 million by the European Competition Commission for the same offense. Its total legal tab, including lawyer's fees and criminal fines, totaled $250 million. Three ADM executives served prison time.

2000 U.S. District Judge Thomas Jackson concluded a two-year trial by ordering the break up of Microsoft into two companies. One would be responsible for operating systems, including Windows, and the other would focus on applications, including Word and Excel and Internet Explorer. Microsoft said it would appeal the case all the way to the Supreme Court, which might have taken years. On January 3, 2000, Microsoft stock had traded at $116.56. Over a 52-week period its price ranged from a high of $119.93 to a low of $60.37. The decision was overruled on appeal.

2000 The Federal Trade Commission ruled that the major record companies had illegally inflated the price of CD's by withdrawing promotional payments from any retailer that advertised records below the minimum advertised price (MAP). The FTC estimated that from 1996 to 1999 customers had paid some $480 million more as a result of the MAP price maintenance strategy. Five major music companies signed a consent decree in which they agreed to stop penalizing retailers who sold their products too cheaply.

Multinational corporations must worry about antitrust action abroad. The European Competition Commission reviews all major mergers with impact in the 15 nation European Union. It claims jurisdiction over any mergers that will have an impact in the European market, even if the companies are based outside the European Union. It has the authority to ban sales within the European Union of any company that tries to evade its authority. In a longstanding antitrust case the Commission has charged Microsoft with a variety of monopolistic practices. In addition to forcing changes in the way Microsoft does business the Commission could impose a fine equal to as much as one tenth of Microsoft's global sales.

Over the years the courts have struggled with the task of interpreting the antitrust statutes. While some presidents have chosen to vigorously prosecute antitrust cases others have chosen to neglect the issue. Since professional economists specializing in corporate finance and antitrust disagree about how serious a problem is created by monopoly and on what antitrust policy would be most appropriate, it is not surprising to find that over the years antitrust policy has been far from consistent.

Summary

1. This chapter has examined four different market types, all involving profit maximizing firms:
   - #1. Competitive markets are characterized by many price-taking firms, all producing the same product. Price is determined by the market so as to equate quantity demanded with supply. Free entry and exit drive economic profits to zero in long-run equilibrium.
   - #2. A monopoly is a single price-setting seller dominating the market.
   - #3. Under monopolistic competition, there are many price setting firms selling differentiated products.
   - #4. A firm dominated by a few large sellers is an oligopoly. A duopoly has only two firms.
2. Monopoly is inefficient because the profit maximizing level of output, where marginal cost equals marginal revenue, is less than the efficient level of output, where marginal cost equals price. Since producing an additional unit of output would cost less in the way of scarce resources

than the value of the additional output to the public, why not produce it?

3. A number of economists, following late Harvard Professor Joseph Schumpeter, argue that monopolistic pricing is the cost we pay for technological progress. Monopolies, it is said, contribute to *dynamic efficiency* by encouraging invention and innovation — but the evidence on this proposition is mixed.
4. Monopolistic competition is a blend of certain essential features of the models of competition and monopoly. Unlike competition, the presence of product differentiation means that each firm under monopolistic competition is a price setter facing a downward sloping demand curve. Unlike monopoly, there are many firms in the industry.
5. Cournot assumed that each firm in an oligopolistic industry would set its price under the mistaken assumption that other firms in the industry would not react by changing their prices. Bertrand argued that with product differentiation each firm would set quantity under the assumption that the other firms in the industry would not react by changing their output. Either way, the consumer benefits if the firms persist in their unrealistic assumptions rather than colluding to maximize joint profits.
6. Game theory constitutes an arsenal of mathematical techniques that can be used to analyze such diverse topics as conflicts among nations and the behavior of firms in oligopolistic industries. A payoff matrix is a simplifying pictorial device used by game theorists for conveniently showing the payoffs from the game (the benefits and costs of participation). A dominant strategy is one that is best for you no matter what your opponent does. A Nash equilibrium is a situation in which each player is doing the best they can, given the decisions made by the other participants in the game.
7. The United States Congress passed the Sherman Anti-Trust Act [1890] and the Clayton Act [1914] in order to cope with monopoly. The history of the antitrust laws over the last century shows that enforcement has been uneven, depending in part on the attitude of the courts and perhaps also in part upon whether a Democrat or a Republican occupies the Whitehouse. The courts have broken up some very large enterprises, such as U.S. Steel, Standard Oil, American Tobacco, and ATT. They have imposed substantial penalties on many major corporations that have been convicted of violating antitrust laws.

Key Concepts

Exercises

1. All firms in a competitive industry have long run total costs of

$$C(q) = 20 + 4q + q^2\,.$$

a. Determine the level of output at which average cost is at its minimum. Determine marginal cost and average cost at this level of output.
b. If there is free entry and exit from this industry, what will be the long run competitive price and how much will our representative firm sell?
c. If the demand curve facing the industry is $Q = 1000 - 10p$, what will be total industry sales at the price you determined in a. How many firms will there be in the industry in long run equilibrium?
d. If a \$1.00 per unit tax is imposed, what will happen in the long run to the profits of each firm that remains in the industry? What will happen to the price paid by the consumer? What will happen to the number of firms in the industry?

2. Prove for any cost function $C(q)$ that if average cost is minimized at output q^*, then marginal cost equals average cost at that level of output.
3. An electric power distributor facing demand function $p(q) = 20 - 2q$ has total cost function $C(q) = q + q^2$.

a. If the monopoly is free to maximize profits, what price will it charge and what quantity will it sell? Determine total profits and consumer surplus.
b. Draw a graph showing the firm's marginal cost, average total cost, the demand curve, total profit and consumer surplus.
c. Suppose you sit on the Public Utility Control Commission. A consumer advocate recommends that the price be set so as to equate marginal cost with price (more precisely produce at the level where $p(q) = MC$). Determine output, price, consumer surplus and profit.
d. Does the regulation suggested in c serve to maximize the sum of profits plus consumer surplus?
e. Suppose that the Public Utility Commission is abolished and that instead the government imposed a tax of \$2 per unit sold by the monopoly. How would the tax influence the price and quantity sold by the monopolist? Determine how the tax will affect consumer surplus and profits.
f. Is the \$2 tax better than the public utility regulation solution in terms of maximizing the net social gain = sum of consumer surplus + after tax profits plus tax revenue = consumer surplus + firm's sales revenue + government tax revenue − total costs? Discuss, considering both the efficiency and the equity issues raised by the tax.

4. Pollution: Suppose that the Public Utility described in question 3 generates pollution costing the general public about \$2.00 per unit produced. Congressman Bullhorn suggests that the utility be required to pay a tax of \$2.00 per unit produced in order to offset the social costs of pollution.

   a. How will the tax affect marginal cost and average cost?
   b. What will be the level of output produced by the monopolist if the tax is imposed.
   c. Should the government impose the tax?
      Hint: You can suppose the objective is to maximize the net social gain = after tax profits + consumer surplus − pollution costs + government tax revenue = consumer surplus + firm's sales revenue + government's tax revenue − production costs − pollution cost.

5. Product Differentiation and Price Discrimination: Suppose that a firm sells its product in differentiated markets with demand functions $p_1 = 10 - q_1/2$ and $p_2 = 20 - q_2/4$. Production costs are $C(q_1, q_2) = 2(q_1 + q_2)$. Find the profit maximizing level of output and prices.
6. Monopolistic Competition: As explained in Chapter 6.4, the ith firm sells its product in a monopolistic competitive market with demand function

$$q_i = 10(11 - p_i + 0.75\overline{p})/n^{0.5}\,,$$

   where p_i is the price it charges, $\overline{p}$ is the average price charged by all the firms in the industry, q_i is annual sales, and n is the number of firms in the industry. There is free entry and exit from the industry. Production costs for each firm in the industry are $C = 64 + 4q_i$; i.e., there are fixed costs of \$64 and marginal cost = average variable costs = \$4.

   After the newly elected governor has imposed a licensing fee of \$64 on each firm, the number of firms declines to 25.

   a. If all the other competing firms are charging $\overline{p} = \$12$, what is the best price for the ith firm to charge in order to maximize its profits? With this price will the ith firm make positive after-tax profits after paying the licensing fee?
   b. Would the ith firm be better off if every firm in the industry charged \$18.00? Explain.
      Hint: Calculate the resulting quantities and profits if every firm charges \$18.00.

If all the other firms are charging \$18.00, what is the best price for the ith firm to charge?

c. Would this industry be in long run monopolistic competitive equilibrium with $p_i = \overline{p} = \$18$ and $n = 25$? Would this industry be in long run monopolistic competitive equilibrium with $p_i = \overline{p} = \$12$ and $n = 25$? Explain.

7. Joan Robinson objected in her book on *Economics of Imperfect Competition* [1932] that the market mechanism tends to generate too much product variety, leading to excessive production costs and distribution expense. Does the solution you found to question 7 suffer from this problem? That is to say, if you were a central planner could you find a more cost effective way of supplying the same total output to consumers at a lower price? Could a change in the number of factories increase welfare? Explain.
8. Game Theory: Congratulations, you have been promoted to President of the Ajax manufacturing company. Ajax sells in a duopolistic market in competition with the Bjax company. Duopolists Ajax and Bjax are both considering an expansion of their capacity. The following payoff matrix shows how their profits (in millions of dollars) depend on their decisions (the first number of each pair is the profit of Ajax; the second *number* is the profit of Bjax):

Table 6.8. Profits and capacity.

	Bjax has low capacity	*Bjax has large capacity*
Ajax has low capacity	\$8, *\$5*	\$5, *\$4*
Ajax has large capacity	\$9, *\$4*	\$4, *\$3*

a. Does the Ajax company have a dominant strategy? Does the Bjax company have a dominant strategy? Explain.
b. Is there a Nash equilibrium for this game? Explain.
c. Will you expand your plant capacity, Mr. President? Explain your decision.

9.* Construct a simple numerical example showing that the prohibition of discriminatory pricing might, under certain circumstances, lead to a reduction in Consumer Surplus.

7

Distribution: Who Gets What?

7.1 Overview

This chapter will look at how income is distributed, both among individuals in the United States and throughout the world. The first task will be to learn about various statistical measures used to describe the distribution of income. Then we will look at the evidence. Later we will examine the process by which market forces allocate rewards among workers and investors. Finally we will look at how government welfare programs and tax policy influence the distribution of income in the United States.

7.2 Measuring inequality

Suppose you were assigned the task of cutting a pie into five pieces for five people. Suppose you divided the pie as shown on Figure 7.1, giving one individual 50% of the pie and another less than 4%. No one would consider such a division to be fair, but this is the way that income is distributed in the United States. In 1997 the wealthiest 20% of the population received 50.1% of total income while the 20% at the bottom of the economic ladder received only 3.5%.

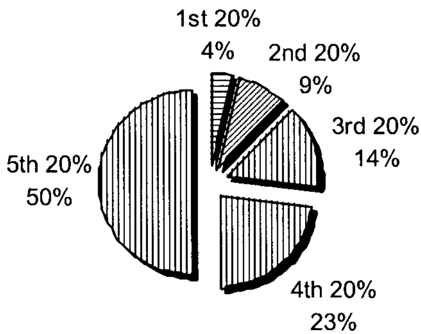


Fig. 7.1. Dividing the income pie.

The data on Table 7.1 have much to reveal about the wellbeing of the inhabitants of 82 countries for which data are readily available. These countries constitute about 85% of the world's population. Some of the concepts reported on the table are fairly easy to explain. The first column of data reports each country's population, the second reports life expectancy at birth, and the third presents the infant mortality rate, which is the number of deaths during the first year of life per 1,000 live births. These data indicate that in the majority of the regions of the world life expectancy is much shorter and infant mortality is much higher than in the United States. But in seven countries people on average do live longer than in the United States. And even more countries have a lower infant mortality rate than the United States. While these measures of social wellbeing are readily understood, the concepts presented in the remaining columns of the table require considerable explanation.

Table 7.1. Some international comparisons of well-being.

	Country	Population (million)	Life expectancy	Infant mortality	GNP/pc	GNP/n ppp	CON/pc ppp	Bottom 20%	Top 20%	Top20/ B20	Gini (%)
1	Tanzania	28.8	51	115	140	620	546	188	1,239	6.6	36.5
2	Madagascar	13.1	52		200	640	582	169	1,456	8.6	41.5
3	Rwanda	7.8		111	80	740	644	312	1,259	4.0	27.8
4	Nigar	8.7	46		230	770	631	237	1,392	5.9	34.5
5	Guinea-Bissau	1.0	38	148	240	820	738	77	2,173	28.0	53.7
6	Zambia	9.2	47	106	350	860	722	141	1,820	12.9	44.7
7	Nigeria	108.0	52		280	1,190	940	188	2,317	12.3	43.4
8	Nepal	20.9	54	101	200	1,230	959	437	1,895	4.3	28.9
9	India	913.6	62	90	320	1,280	870	370	1,854	5.0	32.3
10	Kenya	26.0	59	67	250	1,310	812	138	2,522	18.3	54.4
11	Bangladesh	117.9	57	103	220	1,330	1,131	531	2,142	4.0	27.2
12	Cote d'Ivoire	13.8	56	95	610	1,370	795	270	1,752	6.5	35.5
13	Uganda	18.6	42	118	190	1,410	1,199	407	2,882	7.1	39.0
14	Mauritania	2.2	51	119	480	1,570	1,256	226	2,920	12.9	40.7
15	Senegal	8.3	50	81	600	1,580	1,248	218	3,657	16.7	51.7
16	Lesotho	1.9	61	81	720	1,730	1,488	208	4,471	21.5	53.7
17	Nicaragua	4.2	67		350	1,800	1,710	257	4,720	18.4	49.0
18	Honduras	5.8	66	49	600	1,940	1,416	269	4,064	15.1	50.3
19	Zimbabwe	10.8	58	48	500	2,040	1,306	261	4,067	15.6	54.2
20	Ghana	16.6	58	83	410	2,050	1,722	680	3,633	5.3	32.6
21	Guinea	6.4	44		520	2,061	1,690	254	4,242	16.7	45.3
22	Pakistan	126.3	60	97	430	2,130	1,512	635	3,002	4.7	29.8

Table 7.1. (Continued)

	Country	Population (million)	Life expectancy	Infant mortality	GNP/pc	GNP/n ppp	CON/pc ppp	Bottom 20%	Top 20%	Top20/ B20	Gini (%)
23	Bolivia	7.2	60	83	770	2,400	1,896	531	4,569	8.6	40.5
24	China	1,190.9	69	38	530	2,510	1,079	335	2,369	7.1	36.4
25	Philippines	67.0	65	41	950	2,740	1,945	632	4,650	7.4	39.1
26	Kazakstan	16.8	68		1,160	2,810	1,686	632	3,406	5.4	31.5
27	Moldova	4.3	68		870	3,002	2,372	818	4,921	6.0	33.2
28	Sri Lanka	17.9	72	18	640	3,160	2,402	1,069	4,719	4.4	28.9
29	Latvia	2.5	68		2,320	3,220	1,707	819	3,132	3.8	26.0
30	Lithuania	3.7	69		1,350	3,290	2,500	1,013	5,263	5.2	32.2
31	Jamaica	2.5	74	15	1,540	3,400	2,346	680	5,572	8.2	39.6
32	Guatemala	10.3	65	60	1,200	3,440	2,958	311	9,319	30.0	56.9
33	Morocco	26.4	63	57	1,140	3,470	2,360	779	5,462	7.0	37.7
34	Indonesia	190.4	63	74	880	3,600	2,196	955	4,469	4.7	30.6
35	Peru	23.2	65	53	2,110	3,610	2,527	619	6,368	10.3	43.1
36	Egypt, Arab Rep.	56.8	62		720	3,720	3,013	1,311	6,192	4.7	30.7
37	Dominican Rep	7.6	70	54	3,080	3,760	3,008	632	8,377	13.3	48.3
38	Romania	22.7	70		1,270	4,090	2,536	1,166	4,412	3.8	24.6
39	Jordan	4.0	70	29	1,440	4,100	3,075	907	7,703	8.5	41.5
40	Ecuador	11.2	69		1,280	4,190	2,933	792	7,714	9.7	44.4
41	Belarus	10.4	69		2,160	4,320	2,203	1,223	3,624	3.0	20.8
42	Bulgaria	8.4	71	17	1,250	4,380	2,803	1,163	5,508	4.7	29.5
43	Estonia	1.5	70		2,820	4,510	2,165	714	5,012	7.0	37.7
44	Russian Fed	148.3	64		2,650	4,610	2,305	426	6,200	14.5	47.2

Table 7.1. (Continued)

	Country	Population (million)	Life expectancy	Infant mortality	GNP/pc	GNP/n ppp	CON/pc ppp	Bottom 20%	Top 20%	Top20/ B20	Gini (%)
45	Tunisia	8.8	68	38	1,790	5,020	3,112	918	7,205	7.8	38.5
46	South Africa	40.5	64		3,040	5,130	3,027	499	9,580	19.2	55.9
47	Columbia	36.3	70	23	1,670	5,330	3,998	720	11,153	15.5	49.2
48	Brazil	159.1	67	58	2,970	5,400	3,294	346	11,117	32.1	60.6
49	Poland	38.5	72	15	2,410	5,480	3,507	1,631	6,418	3.9	26.1
50	Algeria	27.4	68		1,650	5,628	3,208	1,107	7,394	6.7	37.0
51	Panama	2.6	73	21	2,580	5,730	3,495	350	10,451	29.9	54.3
52	Costa Rica	3.3	77	14	2,400	5,774	3,464	693	8,782	12.7	44.2
53	Hungary	10.3	70	16	3,840	6,080	4,378	2,079	8,011	3.9	25.9
54	Slovenia	2.0	74		7,040	6,230	3,427	1,628	6,493	4.0	27.1
55	Slovak Republic	5.3	72		2,250	6,580	3,487	2,075	5,475	2.6	18.7
56	Thailand	58.0	69	27	2,410	6,970	3,834	1,073	10,101	9.4	44.4
57	Mexico	88.5	71	36	4,180	7,040	4,928	1,010	13,626	13.5	48.2
58	Venezuela	21.2	71	34	2,760	7,770	5,594	1,007	16,336	16.2	51.4
59	Malaysia	19.7	71	15	3,480	8,440	4,473	1,029	12,011	11.7	46.3
60	Chile	14.0	72	17	3,520	8,890	5,601	980	17,082	17.4	53.7
61	Czech Republic	10.3	73		3,200	8,900	5,162	2,710	9,653	3.6	25.4
62	Korea, Rep of	44.5	71	16	8,260	10,330	5,475	2,026	11,552	5.7	32.9
63	Spain	39.1	77	8	13,440	13,740	8,656	3,592	15,841	4.4	27.2
64	Israel	5.4	77	9	14,530	15,300	9,333	2,800	18,479	6.6	32.6
65	New Zealand	3.5	76	9	13,350	15,870	9,522	2,428	21,282	8.8	37.9
66	Finland	5.1	76	6	18,850	16,150	9,206	2,900	17,306	6.0	31.0

Table 7.1. (Continued)

	Country	Population (million)	Life expectancy	Infant mortality	GNP/pc	GNP/n ppp	CON/pc ppp	Bottom 20%	Top 20%	Top20/ B20	Gini (%)
67	Sweden	8.8	78	6	23,530	17,130	9,422	3,769	17,383	4.6	28.1
68	United Kingdom	58.4	76	7	18,340	17,970	11,501	2,645	25,474	9.6	38.6
69	Australia	17.8	77	8	18,000	18,120	11,416	2,511	24,087	9.6	36.7
70	Italy	57.1	78	8	19,300	18,460	11,445	3,891	23,463	6.0	32.9
71	Netherlands	15.4	72	7	22,010	18,750	11,438	4,689	21,102	4.5	27.9
72	Germany	81.5	76	7	25,580	19,480	11,298	3,954	22,766	5.8	32.3
73	France	57.9	78	7	23,420	19,670	11,999	3,360	25,137	7.5	34.8
74	Denmark	5.2	75	8	27,970	19,880	10,338	2,791	19,952	7.1	32.6
75	Canada	29.2	78	7	19,510	19,960	12,176	3,470	24,473	7.1	33.5
76	Norway	4.3	78	8	26,390	20,210	10,509	3,258	19,284	5.9	30.0
77	Belgium	10.1	76	8	22,870	20,270	12,567	4,964	22,621	4.6	27.2
78	Japan	125.0	79	5	34,630	21,140	12,261	5,334	22,990	4.3	27.7
79	Singapore	2.9	75	6	22,500	21,900	8,760	2,234	21,418	9.6	41.5
80	Hong Kong	6.1	72	7	21,650	22,554	13,307	3,593	31,271	8.7	39.2
81	Switzerland	7.0	78	7	37,930	25,150	14,839	3,858	33,090	8.6	37.2
82	United States	260.6	77	9	25,880	25,880	17,598	4,136	36,869	8.9	36.2
	Average	57.3	66.9	42.4	6,697	7,526	4,597	1,382	9,823	9.5	37.9
	Maximum	1190.9	79.0	148.0	37,930	25,880	17,598	5,334	36,869	32.1	60.6
	Minimum	1.0	38.0	5.0	80	620	546	77	1,239	2.6	18.7

Source: Michael C. Lovell, "Inequality Within and Among Nations," *Journal of Income Distribution*, 1998, pp. 5–44.

7.2.1 *Three measures of economic wellbeing*

GNP/pc

GNP stands for Gross National Product, which is a measure of the market value of all the goods and services produced in a country during the year. The table reports GNP per capita, which is the output per person produced by the economy.

In order to permit meaningful comparisons among nations, the GNP per capita figures are not reported in the currency of the relevant country; after all, GNP figures in pounds for England are not directly comparable to the GNP of Japan measured in yen. In order to permit meaningful comparisons, all the GNP/pc figures have been translated into United States dollars, as a common denominator so to speak. For example, the value in yen of Japan's per capita output has been translated from yen into dollars. The GNP/pc numbers were translated using the foreign exchange rate, just as a Japanese tourist traveling to the United States might convert yen into dollars. The conversion is made using an average of exchange rates over several years in order that the conversion will not be distorted by temporary fluctuations in the foreign exchange market.

GNP/n ppp

This is a more sophisticated measure of output per capita. It differs from *GNP/pc* in that the conversion into dollars is made by a much more difficult but more accurate procedure.

The primary problem with using exchange rates to convert foreign output per capita figures into dollars is that the exchange rate is not a very precise yardstick for converting purchasing power values from one country's currency into another's. For goods that are traded internationally, such as oil and wheat and coffee, the foreign exchange rate may provide a reasonable basis for converting purchasing powers from one currency into another. But for goods that are *not* traded among nations, such as haircuts and fire protection and milk, the foreign exchange rate may not provide a good basis for comparing their values in different countries. An alternative but more difficult measure of per capita wellbeing, the data in the *GNP/n ppp* column, is estimated by University of Pennsylvania scholars by conducing surveys designed to compare the cost of purchasing a representative market basket of goods and services in different countries.[1] It is interesting to note from the table that this more precise measure indicates that per capita output for the poorer countries is not quite as low as was indicated by *GNP/pc*. Also, the United States, which was ranked substantially behind both Switzerland and Japan by *GNP/pc*, leads the world in per capita output when ranked by *GNP/n ppp*.

CON/pc ppp

This is consumption spending per capita using the same purchasing power conversion factors used for the *GNP/n ppp* comparisons. Consumption is substantially smaller than GNP because in some countries much output may be diverted from consumption to meet military and other government requirements and because the part of GNP that is invested in physical plant, machinery and so forth is not immediately available for the consumer. So per capita consumption may be a better measure of current economic wellbeing than per capita GNP.

7.2.2 *Income by quintiles*

The remaining columns on the table provide information relevant to the study of the distribution of income.

Bottom 20%

The average income of those in the bottom 20% of the income distribution.

Imagine, for a moment, that we had everyone in the country form a single line. We could order them randomly or alphabetically. But instead let us line them up by their incomes, starting with the lowest of the homeless on the left and with Microsoft's Bill Gates at the extreme right of the line. Now let us divide the line into five equally sized groups. Statisticians call such groups *quintiles*. Then *Bottom 20%* is the average income received by the lowest quintile, the bottom 20% of the population when the population is ordered by income — this is indeed the bottom of the economic ladder.

[1]The purchasing power parity data are known as the *Penn World Tables* because they were developed by Alan Heston, Irving Kravis, Robert Summers at the University of Pennsylvania: http://pwt.econ.upenn.edu/. The problem of making meaningful comparisons of income levels in different countries at any point of time is similar to the problem of making meaningful comparison of price levels and output measures over time in a particular country. The details of that procedure will be discussed in Chapter 8.

This concept may well be the best widely available measure of the position of the least advantaged, the key concept of Rawls' *Theory of Justice* discussed in Chapter 4.2.5. The table reveals that the least advantaged, or at least the bottom 20%, do better economically in the Netherlands, Belgium and Japan than they do in the United States.

Top 20%

The average income received by those in the top 20% of the income distribution.

It is clear from the table that the top 20% of the United States population do substantially better than the upper crust in any other country. Indeed, the data suggest that members of a Bangladesh household in the top quintile of that country's income distribution could decisively improve their economic situation by immigrating to the United States even if they found themselves slipping into the bottom quintile of the U.S. income distribution — thus the United States distribution unambiguously dominates that of Bangladesh. However, it is clear from the table that such unambiguous dominance is rather rare; in particular, the United States unambiguously dominates only 17 countries. In 65 of the 82 countries, at least one fifth of the population has a higher standard of living than the least advantaged bottom fifth in the United States.

T20%/B20%

A simple measure of inequality is obtained by dividing the income of those in the top quintile by the income of those in the bottom quintile. For the United States we have T20%/B20% = \$36,869/\$4,136 = 8.9. This measure of income inequality is reported for 82 countries in the next to last column of Table 7.1. Inspection of the *T20%/B20%* column reveals that the income distribution is remarkably unequal in Brazil, Panama, Guatemala, Guinea-Bissau and Kenya. Among high-income developed countries, the United Kingdom, Australia and the United States have a high level of inequality as measured by this ratio.

Table 7.2 presents more detailed data by quintiles for the United States, with one row for each of the five quintiles. The second column presents the average income of each quintile. The next to last entry in the second column Table 7.2 of \$58,208, is the average income of all households. The median reported at the very bottom of this column, \$42,228, is the middlemost income. More precisely, considering once again the line we obtained by

Table 7.2. Household income by quintiles in the United States, 2001.

	Average income	% of Pop	% of total income	Cumulative % of pop P	Cumulative % of income $L(P)$	Difference $P - L(P)$
1st 20%	10,096	20.0%	3.5%	20.0%	3.5%	16.5%
2nd 20%	25,668	20.0%	8.8%	40.0%	12.3%	27.7%
3rd 20%	42,301	20.0%	14.5%	60.0%	26.8%	33.2%
4th 20%	67,114	20.0%	23.1%	80.0%	49.9%	30.1%
5th 20%	145,862	20.0%	50.1%	100.0%	100.0%	0.0%
Totals:			100.0%			107.5%
Average	$58,208					35.8%
Median	$42,228					

Source: U.S. Bureau of the Census, http://www.census.gov/hhes/income/income01/inctab1.html.

arranging all the households in a row from the poorest to the most wealthy, the *median income* would be the income of the household in the precise center of the line. Half the households have income above the median income of $42,228 and half have income below the median.[2] We will use the remaining columns of the table in developing a Lorenz curve for the United States.

7.2.3 *Lorenz curves*

The Lorenz curve is a clever graphical procedure revealing the degree of inequality in a country's income distribution. Let p denote percentile of the population when ranked by income (not by SAT score), let $y(p)$ be the income received by those in that percentile and y denote the total income received by the entire population. Then the Lorenz function is

$$L(P) = \frac{y(1)}{y} + \cdots + \frac{y(P)}{y} = \frac{\sum_{p=1}^{P} y(p)}{y}\,. \tag{1}$$

This concept is clarified on Table 7.2, which derives the Lorenz function for the United States, but using quintile data because percentile figures are not readily available. The following concepts are reported in successive columns on that table.

[2]If there is an even number of people, we can take the average of the income of the middle two.

% of Pop

The proportion of all the households in each of the five quintiles, which is obviously 20%.

% of Total Income

The proportion of the total income that is received by each quintile. Thus the table tells us that the bottom 20% of the population received 3.5% of total income. Figure 7.1 was based on the figures in this column.

Cumulative % of Pop

Obtained by cumulating the proportion of the population that is in the indicated quintile or one with lower income. Thus the entry for the 2nd quintile of 40% represents the 20% in the 1st plus the 20% in the 2nd quintile, etc.

Cumulative % of Income

Similar to Cumulative % Pop, but obtained by cumulating the proportion of total income that is received by those in the indicated quintile or a lower one. For example, the 12.3% figure for the 2nd quintile is the sum of figures for the first and second quintiles in the % income column, 3.5% + 8.8%.

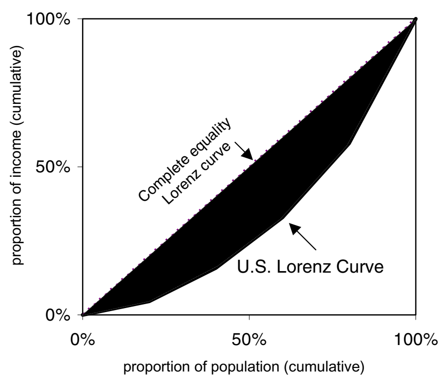


Fig. 7.2. Lorenz curve, U.S. data
The Lorenz curve would be a 45 degree line connecting the lower-left with the upper right corner of the graph if income were equally distributed. How far the Lorenz curve sags below the line of complete equality depends upon how unequally income is distributed.

The Gini coefficient equals twice the shaded area between the Lorenz curve and the 45^0 line.

Lorenz function

The Lorenz function for the United States, which is plotted on Figure 7.2, is based on the data in the last two columns of the Table 7.2. The 45^0 line connecting point $(0\%, 0\%)$ with point $(100\%, 100\%)$ on the graph indicates the shape that the Lorenz curve would assume if, hypothetically, income were equally distributed, every household receiving the same income and each quintile receiving precisely 20% of the total income pie.

7.2.4 *International comparisons*

Figure 7.3 compares the Lorenz curves for India and for the United States. While these two countries differ dramatically in many respects, including per capita income, the closeness of the Lorenz curves suggests that their income distributions are quite similar. Figure 7.4 shows that the income distribution of the Slovak Republic is much more egalitarian than that of the United States, while Brazil has much greater inequality.

World inequality

Figure 7.5 shows the Lorenz curve for the World and purports to answer two interesting questions:

1. If some magic act could eliminate income inequality within every country while leaving unchanged each country's per capita income, how big a step would this make toward the elimination of worldwide inequality of incomes?

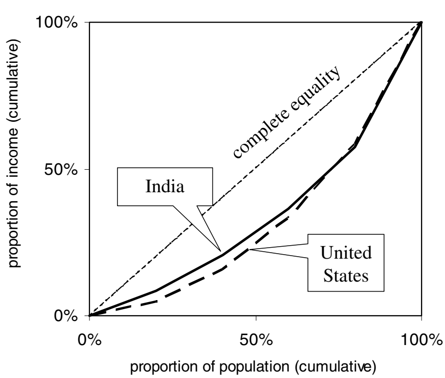


Fig. 7.3. India and U.S. Lorenz curves
The degree of income inequality is about the same in the United States and India.

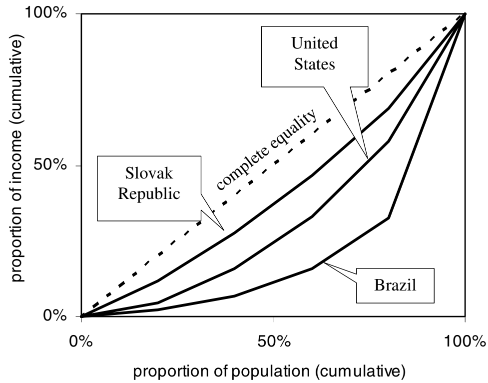


Fig. 7.4. A sample of Lorenz curves
Income is more equally distributed in the Slovak Republic than in the United States. Brazil has substantially more income inequality.

2. Alternatively, if income differences were eliminated among countries, so that every country had the same per capita income, but there was no change in the income distribution inside any country, how much inequality would remain?

The Lorenz curves on Figure 7.5 reveal that the elimination of differences in per capita income *among* countries would be a much greater step toward complete equality than the elimination of differences in income within every country. That is to say, among country inequality is much greater than within country inequality. The graph suggests that if one has compassion for the poor, one should be worrying more about inequality among countries than about inequality within the country where one happens to reside.

7.2.5 *Gini coefficients*

While the Lorenz curve provides a revealing picture of the extent of income inequality, this graphical device cannot be used conveniently in comparing income inequality among a large number of countries. The Gini coefficient is an attempt to summarize in a single number the evidence on income inequality. The concept is most easily explained by looking back at Figure 7.2. Consider the shaded area between the Lorenz curve and the curve representing complete equality. The *Gini coefficient* is twice this area between the Lorenz curve and the complete equality curve. For a hypothetical country with complete equality, the Lorenz curve would coincide with the

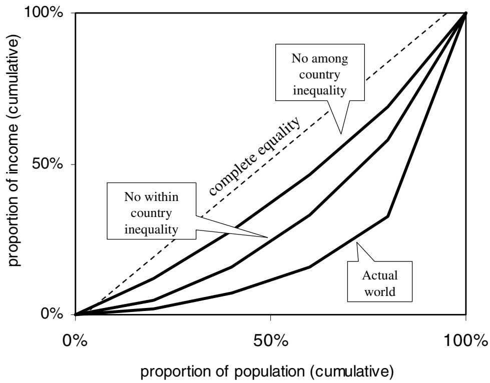


Fig. 7.5. World inequality
If income inequality could miraculously be eliminated within every nation, but today's differences in income among countries remained, there would be little reduction in total world inequality. But the elimination of income differences among nations without any change in within country inequality would cause a major shift of the Lorenz curve toward equality.

complete equality line, there would be no area to shade, and hence the Gini coefficient would be zero. Alternatively, the area would be 1/2 for a country in which one individual (a super Bill Gates) received all the income, for then the Lorenz curve would form a triangle with vertices at points $\langle 0, 0\rangle$, $\langle 100\%, 0\rangle$ and $\langle 100\%, 100\%\rangle$. Thus the Gini coefficient ranges from 0 for complete equality to 100% for complete inequality.

Gini coefficients for 82 countries are recorded in the last column of Table 7.1. The Gini coefficients are large for many Latin American countries, including Mexico and Brazil. The Gini coefficients for India and the United States are almost identical, as was suggested by the similarities of their Lorenz curves on Figure 7.3. Further, the Gini is higher for the United States than for the vast majority of industrialized countries.[3] Of course, inequality varies over time as well as from country to country. The U.S. Census Bureau estimates that during the last half of the 20th century the Gini coefficient declined to a low of 34.8% in 1968 but had climbed to a peak of 43.5% at the end of the century.

[3]The Gini coefficient can be used to measure a variety of different types of inequality. A typical class grade distribution has a Gini Coefficient of about 6% — thus grades are much more evenly distributed among students than income is in any country in the world. Perhaps the most unequally distributed characteristic of education is private university endowments, with a Gini coefficient of about 67%.

7.3 Rewards: Wages, productivity, exploitation and monopsony

While the Gini coefficient attempts to measure the actual degree of inequality, it cannot tell us how much inequality is tolerable or appropriate. Measuring *what is* cannot tell us *what ought to be.*

It is not obvious that the ideal society involves an equal distribution of income. Jeremy Bentham, we saw in Chapter 4.2, argued that the objective of society should be the maximization of the sum total of happiness. As was clear from the numerical example presented on Table 4.2, maximizing the sum of individual utilities might involve a quite unequal distribution of worldly goods or income. Alternatively, maximizing the position of the least advantaged, John Rawls' criterion, does not necessarily involve an equal distribution of income. As was illustrated by Figure 4.5, an all-out attempt at equalizing income might conceivably weaken work incentives to such an extent that output would drop to the point where everyone would be worse off. Too aggressive an attempt at income redistribution might conceivably make everyone worse off if a reduction in the after-tax income of the most successful had severe adverse effects on their incentive to work.

Why should a skilled surgeon receive more income than a common laborer? One answer is that the surgeon should earn more because of the considerable expense, both tuition and foregone earnings, incurred in preparing for this occupation. Would it be fair to fail to compensate the surgeon for the extra investment in human capital required to learn the craft. Moreover, some will argue that it is only fair that more be enjoyed by those who contribute more to the production process because they choose to work longer hours or because they are so fortunate as to be endowed with special skills. And as a practical matter, if such compensation were not paid fewer might enter the profession, causing a shortage of practitioners — incentives matter.

These arguments suggest that complete equality of income, a Gini coefficient of 0, might be neither fair nor practical, but they do not resolve the question of whether the current level of inequality is appropriate or manifestly unjust.

7.3.1 *Wage determination*

Before addressing the question of what constitutes a fair wage, it is necessary to investigate how the level of employment and the wage rate are determined in the marketplace. The wage a profit maximizing firm has to pay for labor is determined in part by the supply function of labor — let $w(L)$ be the inverse labor supply function indicating the wage that the firm will have to pay in order to hire L workers. If the firm is small relative to the size of the local labor market, such as might be the case for a New York law firm hiring a new secretary, the labor market may be quite competitive. In such circumstances the firm will be a price taker in the labor market, we will have $dw/dL = 0$, and we may write $w(L) \equiv w$ to indicate that the wage the firm has to pay is independent of the number of workers it considers hiring. The situation is quite different for a firm that is large relative to the size of its labor market, such as an auto manufacturer setting up a new assembly plant in Kentucky, for then the supply of labor may be much less elastic. Our auto manufacturer, because it has to hire so many workers, will have to attract workers over a sizable geographic area. The more workers it hires, the further they will have to commute. In such circumstances the firm will find that it has to offer a higher wage to induce workers living further from the factory to undertake the lengthy commute; i.e., $dw/dL > 0$, which means that the supply curve of labor is upward sloping. A firm with $dw/dL > 0$ is called a *monopsonist.* So is a firm that finds that it faces an upward sloping supply curve for raw materials or equipment or other inputs into the production process.

In preparation for thinking about the profit maximizing level of employment, consider a firm that faces inverse demand function $p(q)$, hires L workers at wage $w(L)$, utilizes K machines at a monthly rental cost $r(K)$, and has production function $q(L, K)$. Revenue, the product of price times output, will be a function of the number of workers hired and the quantity of machinery,

$$R(L, K) = q(L, K)p[q(L, K)]\,, \tag{2}$$

where $p[q(L, K)]$, the function showing how price is related through quantity to employment and capital, is obtained by substituting the production function into the inverse demand function $p(q)$. By the *marginal revenue product of labor* we mean the derivative

$$\frac{\partial R(L, K)}{\partial L} = \frac{\partial q}{\partial L}p[q(L, K)] + q(L, K)\frac{dp}{dq}\frac{\partial q}{\partial L}\,. \tag{3}$$

This is the change in revenue that will be generated by hiring an additional worker, taking into account both the increment in output that will be produced and the decline in price that will result when that increased output is brought to the marketplace. Note that we may rewrite (3) as

$$\frac{\partial R(L,K)}{\partial L} = \left\{ p[q(L,K)] + q(L,K)\frac{dp}{dq} \right\} \frac{\partial q}{\partial L}\,, \tag{3*}$$

where the expression in braces is marginal revenue.[4] Therefore, the marginal revenue product of labor is marginal revenue times the marginal product of labor.

Now the total cost of producing our output is

$$C(L,K) = Lw(L) + Kr(K)\,, \tag{4}$$

where $Lw(L)$ is the cost of all the labor utilized by the firm and $r(K)$ is the cost of using a machine. The derivative of total cost with respect to labor,

$$\frac{\partial C}{\partial L} = w(L) + L\frac{dw}{dL}\,, \tag{5}$$

is the *marginal labor cost*. The critical point to observe is that $\partial C/\partial L$, the cost of hiring an additional worker, is greater than the wage because when the additional worker is hired, all the workers already on the payroll must receive a matching wage increase — to pay your present employees less than your new hire would invite labor unrest and undermine productivity.[5]

The profits of our firm will be total revenue minus total cost or

$$\pi(L,K) = R(L,K) - Lw(L) - Kr(K)\,. \tag{6}$$

To find the level of employment that will maximize profits in the short run, given K, we obtain by differentiating equation (6) and substituting from (3) a necessary condition for profit maximization

$$\frac{\partial \pi}{\partial L} = \frac{\partial q}{\partial L} p[q(L,K)] + q(L,K)\frac{dp}{dq}\frac{\partial q}{\partial L} - \left[w(L) + L\frac{dw}{dL} \right] = 0\,, \tag{7}$$

[4] Recall from equation 21 of Chapter 5 that $dR/dq = p(q) + qdp/dq$; the expression in braces in (8) is more complicated because it explicitly states that output depends on L and K.

[5] Sometimes employers try to circumvent this difficulty by offering new workers only the prevailing wage but supplementing it with a signup bonus and transportation or moving allowance and other startup incentives.

or

$$\left\{p[q(L,K)] + q(L,K)\frac{dp}{dq}\right\}\frac{\partial q}{\partial L} = w(L) + L\frac{dw}{dL}\,. \tag{8}$$

Note the term on the left of the equality is the marginal revenue product of labor, or $\partial R/\partial L$ as defined by equation $(3)^*$. The term on the right of the equality is marginal labor cost as specified by equation (5). All this means that the profit maximizing employer will hire workers up to the point where

$$\begin{aligned} &MRP(\text{Marginal Revenue Product of Labor}) \\ &\quad = MLC(\text{Marginal Labor Cost})\,. \end{aligned} \tag{9}$$

Intuitively, the equation states that at the point of profit maximization the increment of revenue generated by hiring an additional worker must precisely equal the increment in costs arising from hiring that worker, including the additional wages that will have to be paid to workers already on the payroll.

Graphical analysis

The implications of this profit maximizing condition are conveniently summarized on Figure 7.6: Both the average labor cost function $w(L)$ and the marginal labor cost function dC/dL are plotted on the bottom panel of this graph. The positive slopes of these two curves reflect the fact that our monopsonist will have to pay higher wages the more labor it employs. The marginal product and average product of labor curves are also drawn on the graph. The profit maximizing level of employment is determined by the point e where the marginal revenue product of labor curve intersects the marginal labor cost curve in accordance with condition (8). The resulting wage rate is w^e.

7.3.2 *Marginal productivity theory of wages*

The marginal productivity theory of wages assumes that employers are profit maximizers and that both the product market and the labor market are competitive. With dp/dq and dw/dL both equal to 0, profit maximization condition (8) reduces to

$$p\frac{\partial q}{\partial L} = w(L)\,; \tag{10}$$

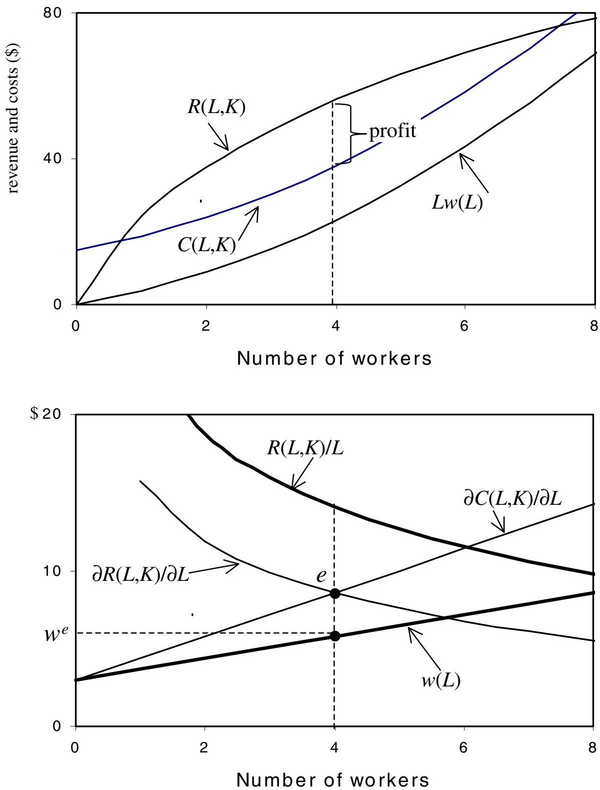


Fig. 7.6. Monopsony

The wage rate as a function $w(L)$ of the level of employment is specified on the bottom panel of the graph. The resulting total labor cost curve, $Lw(L)$, is plotted on the upper panel. Adding fixed costs yields the total cost function $C(L, K)$. The resulting marginal labor cost function $\partial C(L, K)/\partial L$ and average revenue per worker $R(L, K)/L$ are plotted on the lower panel. The profit maximizing level of employment, given capital stock K, is determined by the equality of marginal labor cost and marginal revenue product of labor at point e on the lower panel; our firm hires four workers. Now that we know the level of employment, we can read the wage rate, employment costs, revenue and profits from the graphs.

i.e., the workers are paid a wage that is equal to the value of their marginal product of labor. Or to put it another way,

$$\frac{w(L)}{p} = \frac{\partial q}{\partial L}\,, \tag{11}$$

which says that the real wage (the money wage adjusted for the level of prices) is equal to the marginal product of labor.[6]

The marginal productivity theory of wages, which may first have been clearly stated by Columbia University Professor John Bates Clark [1847–1938], relies on the assumption that both the labor market and the market in which the product are sold are indeed competitive. What happens if in fact markets are not competitive?

Monopoly, monopsony and exploitation

Famed English economist A. C. Pigou [1877–1959] argued that workers are *exploited* unless the wage they are paid is equal to the value of the increment in output made possible by the workers' labor. This means, he said, that workers should be paid the value of their marginal product; i.e., $w = p\partial Q/\partial L$. As indicated by equation (10), precisely this condition will prevail if the market in which workers are hired and the market in which the employer sells the product are both competitive.

The requirement that workers be paid the value of their marginal product is clearly violated in the case depicted on Figure 7.6. The wage received by the workers, w^e, is substantially less than the value of their marginal product for two reasons:

#1 The marginal labor cost is above the wage rate, which is necessarily the case for a profit maximizing monopsonist with $dw/dL > 0$.
#2 The value of the marginal product of labor is greater than marginal revenue product because marginal revenue is less than price, which is necessarily the case because our firm is not selling its product in a perfectly competitive market.

According to Pigou's definition, workers will be exploited unless *both* the labor market and the product market are competitive. Otherwise, the workers will receive less than the value of their marginal product. Pigou's argument implies that perfectly competitive product and labor markets are essential for fairness as well as for economic efficiency.

[6] The concept of the real wage is discussed at length in Chapter 8.5.1.

Pigou's intriguing definition of exploitation is not without its problems.[7] Suppose, for example, that a monopolist and a competitive firm hire workers in the same market, and therefore pay them the same wage. But the employees of the competitive firm will be paid the value of their marginal product while those working for the monopolist will receive less than theirs. Pigou would have to say that those who work for the monopolist are exploited while those who work for the competitive firm are not, even though they all receive the same wage for equal work effort.

7.3.3 *Labor's share and the Cobb-Douglas production function*

Paul H. Douglas and his collaborator Charles W. Cobb used their famous production function to address a related question of fairness: what determines the size of the share in the total output of the economy that will be allocated to workers and how much will be claimed by the capitalists? The income that workers receive is wL, the product of the wage times the number of hours of work supplied. The value of output is pQ, where Q is real output and p is the price level. *Labor's share* is the ratio of the income workers receive to the value of output, or wL/pQ. Cobb and Douglas observed that labor's share had been quite stable historically. They explained that labor's share was stable because it was determined by the nation's technology, arguing as follows. With their production function, $Q = L^{\lambda}K^{(1-\lambda)}$, the marginal productivity of labor is $\partial Q/\partial L = \lambda Q/L$, as was explained by equation (10) in Chapter 5, page 210. Under the assumption of perfectly competitive labor and product markets, the value of the marginal product equals the wage; i.e., $p\partial Q/\partial L = w$ or $w/p = \partial Q/\partial L = \lambda Q/L$. Multiplying this last equation by L/Q reveals that labor's share is

$$\lambda = \frac{wL}{pQ}\,. \tag{12}$$

Cobb and Douglas argued that labor's share had been stable because it was determined by the parameter λ of their production function. More than this, their estimated value of $\lambda = wL/pQ = 0.807$ (see equation 7, page 209) was remarkably close to the observed labor's share, exactly as their theory predicted.

[7] Karl Marx [1818–1883] had a fundamentally different but equally problematic definition of exploitation.

If, as Cobb and Douglas argued, labor's share is technologically determined, the scope for changing the distribution of income is severely limited and the perennial struggle between labor and their capitalist employers comes to naught. If the workers do succeed in pushing up money wages, the gains may be wiped out at the store when employers pass the cost increases through to their customers. However, their argument relied on the assumption that both the product market and the labor market are competitive; i.e. they were denying the existence of both monopoly and monopsony.

Quite apart from the questionable validity of the Cobb-Douglas assumption of competition, follow-up empirical research, stimulated in part by the pioneering Cobb-Douglas study but utilizing more elaborate functional forms for the production function, reveals that the share of output received by workers, far from being entirely technologically determined, may indeed depend in part on the ratio of capital to labor used in the production process. In a controversial elaboration of this argument, Stanford Professor Michael Boskin asserted that a tax policy that encourages capital accumulation by favoring investors will in the long run increase the ratio of capital to labor, which will lead to a rise in worker productivity and hence an increase in labor's share in a larger output.[8] Boskin later served as a key economic adviser to President George H. Bush, and his argument provided intellectual justification for the supply side economic policy of providing tax relief for investors. A decade later President George W. Bush successfully pushed for drastic tax cuts, arguing that reducing taxes on capital gains and dividends and the phasing out of the inheritance tax would shorten the recession and encourage long run growth.

7.3.4 *The minimum wage and union bargaining*

According to the competitive model, a decision by the government to raise the minimum wage will generate unemployment, as was explained in Chapter 3.6.2. Paradoxically, if the labor market is monopsonistic, dominated by a single large employer rather than competitive, an increase in the minimum wage may generate an *increase* in employment. Figure 7.7 explains why. The minimum wage changes a segment of the labor supply function as indicated by the horizontal segment on the lower panel

---

[8] Michael Boskin, "Taxation, Saving, and the Rate of Interest," *Journal of Political Economy*, April 1978, Part 2, pp. S3–S27.

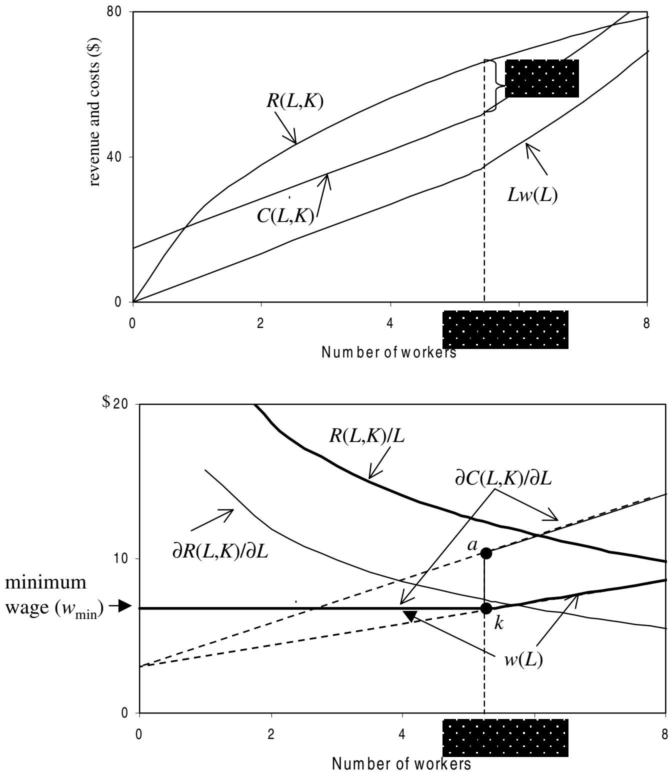


Fig. 7.7. The minimum wage increases employment under monopsony
The minimum wage creates a new labor supply curve on the bottom panel with a kink at point k corresponding to employment level $L(w_{\min})$. More than $L(w_{\min})$ workers can be hired only by paying above the minimum wage. As a result, the labor cost curve $Lw(L)$ in the top panel has a linear segment for employment below $L(w_{\min})$ followed by a kink at k. That is why its derivative, the marginal labor cost curve on the bottom panel, is identical to the minimum wage for employment levels before the kink at k; it jumps at point k to point a because the labor cost function $C(L, K)$ is not differentiable at $L(w_{\min})$. With the minimum wage, profits are maximized with 5.36 rather than 4 workers.

of the graph — even if employment is cut below point $L(w_{min})$ the wage cannot fall below $w_{\min}$. As a result, the Labor Cost curve on the upper panel has a linear segment and a kink at $L(w_{\min})$ because more than $L(w_{\min})$ workers can be hired only by paying them more than the minimum wage. Marginal labor cost equals $w_{\min}$ only up to point $L(w_{\min})$. Because the resulting labor cost function $L(w)$, plotted on the upper panel, is not differentiable at wage $w_{\min}$, the marginal labor cost function plotted on the bottom panel is discontinuous at this point, jumping from k to a. Beyond that level of employment the minimum wage floor is irrelevant because the market yields a higher wage — this means that at levels of employment above $L(w_{\min})$, both $w(L)$ and marginal labor cost (MLC) are the same as they were on Figure 7.6. The key point to remember is that at levels of employment below $L(w_{\min})$ the wage remains constant at $w_{\min}$, which makes marginal labor cost, $dC(q)/dL$, equal to $w_{\min}$. This is so because the firm is already paying its present employees $w_{\min}$, which means that it will not have to pay them more per hour when new workers are hired. As can be seen from the graph, the profit maximizing level of employment is now $L(w_{\min})$, which is larger than the before minimum wage level reported on Figure 7.6. Note, however, that if the minimum wage were raised very high there would be a reduction in employment.

Precisely the same argument also applies to labor unions. When a labor union succeeds by negotiation or a strike in pushing up the wage rate, some workers will be laid off if the labor market is competitive. But if the workers are employed by a monopsonist, the increase in the wage scale won by the union might lead to an increase in employment!

7.3.5 *Efficiency wage theory*

According to competitive theory, wages adjust to clear the market, supply equaling demand. But if supply equals demand, there is no unemployment! Well in fact there will be *frictional unemployment* because time is required for willing workers to find vacant job opportunities. According to theory, *involuntary unemployment* is created when the market is not allowed to do its work, wages being kept above the market clearing level by the minimum wage, labor union pressure, or some other rigidity.

The theory of efficiency wages provides an alternative explanation of how labor markets work. In competitive labor markets, workers are paid the value of their marginal product, but suppose that productivity depends in part on the level of wages? Efficiency wage theory asserts that a wise employer may find it worthwhile to pay workers more than it has to. If the workers only get the market-clearing wage, the jobs may be filled. But paying a higher wage and offering generous fringe benefits might increase worker productivity!

In the absence of a wage premium, absenteeism may be high because the workers won't worry about being fired. The worker who loses his job today because of high absenteeism or slacking off on the job will be able to obtain another job with a new employer tomorrow at the market clearing wage, so why not take the day off and go to the ballgame? Employers worry about absenteeism because it can have a quite disruptive impact on worker productivity, particularly on the assembly line. One way to fight absenteeism is to pay your workers more than the going wage. Workers receiving the premium wage will think twice about skipping work or slacking off on the job because if they are fired they will not receive the wage premium on their next job. Further, employers know that when a worker quits, they will have to hire and train a replacement. Time on the job may be required before the new worker gets up to speed. Paying workers a premium wage will save hiring and training costs by building worker loyalty and reducing excessive labor turnover.

Now suppose that the majority of employers in a labor market decides to combat absenteeism by paying more than the market-clearing wage. This does not mean that the worker who is laid off or fired or quits will be able to find another job paying a wage premium. With wages generally above the market clearing level there are more workers than jobs — unemployment prevails. Unemployment will make workers think twice about quitting. The fear of unemployment will encourage workers to avoid absenteeism and other types of behavior that increase the risk of being fired.

The key point of such arguments is that the productivity of workers may depend positively on their wage rate. A higher wage leads to higher output per worker. That is why this line of reasoning is known as *efficiency wage theory*. The theory helps to explain why the prevailing wage may be above the market-clearing wage; it explains why involuntary unemployment can prevail.

7.4 Government and the distribution of income

Inequality in the United States is tempered by tax policy and a variety of government programs. The extent of inequality is also influenced by government cash transfer payments, such as Social Security benefits, welfare payments, and veterans' benefits. Data on the distribution of income for the United States usually refer to money income, leaving out the value of employee provided health benefits and government provided Medicaid. Also neglected in the customary measures of the income distribution are realized capital gains, which are largely received by the well-to-do.[9] In years of booming stock markets, a measure of income including capital gains would reveal substantially more inequality than the U.S. data that was reported on Table 7.2; but when the stock market slumps, the capital losses are borne largely by the wealthy.[10] Table 7.3 shows that government cash transfers make a substantial contribution toward equality. The income tax also contributes toward equality in disposable income.

7.4.1 *Transfer programs*

Payments that the government makes to individuals that are not in exchange for current productive services are called transfer payments. Some transfer payments, such as Veterans Benefits, are made for past services. So are the pensions received by retired government employees. Social Security benefits are in large part but not entirely paid for in advance through payroll taxes. Unemployment benefits are also financed with payroll taxes. Welfare benefits are paid for from the general tax revenue, although in many states those who have received welfare benefits are expected to pay back the sums they have received if substantial earnings, an inheritance or winning the lottery makes it possible for them to do so.

Social Security

Social Security benefits are financed primarily through a payroll tax. Your employer is required to deduct 7.5% from your earnings to help finance social security benefits. More than this, your employer is required to pay a matching 7.5%. However, only the first $80,000 or so of income is subject to this tax.[11] Thanks to this cap, Bill Gates pays no more in Social Security taxes than a computer programmer making $80,000 per year, but they may both receive the same benefit check when they retire.

How much you will get back in retirement benefits when you retire depends on how much you have earned over your working lifespan. The Social Security Administration keeps track of each worker's earnings history

[9]As explained in Chapter 5.3.3, a realized capital gain is the excess of the price you sell an asset for over what you paid for it.

[10]If unrealized capital gains and losses were counted as income, Bill Gates would have fallen to the very bottom of the income distribution when the stock markets tanked in the 2001–2002 Bear Market. For most purposes, capital gains should probably not be counted as income to the extent that they are offset by inflation.

[11]The cap is adjusted each year to keep pace with inflation. It was $80,400 in 2001.

Table 7.3. Inequality in the United States, alternative definitions, 2001.

Income definitions:	Average	1st 20%	2nd 20%	3rd 20%	4th 20%	5th 20%
#1 Money income excluding capital gains	$58,208	3.5%	8.8%	14.5%	23.1%	50.1%
#2 Less government transfers (= #1 − #5)	53,860	1.1%	7.4%	14.8%	23.7%	53.1%
#3 Less government transfers Plus capital gains (= #2 + #6)	58,279	1.0%	6.9%	13.7%	22.4%	56.0%
#4 Money income + health insurance + cap gains − SS taxes	62,628	3.2%	8.2%	13.6%	21.9%	53.1%
#5 Government transfers (= #1 − #2)	4,348	33.4%	26.3%	11.8%	15.5%	13.1%
#6 Capital gains (= #3 − #2)	4,419	0.1%	0.3%	1.1%	6.8%	91.7%

Note: Each row reports for the specified income definition the average level of income and the distribution of income by quintiles.
Source: U.S. Census Bureau, http://www.census.gov/hhes/income/histinc/rdi8.html.

on its computers. The calculation of your monthly retirement benefit is a complicated process. First, the computer calculates your average indexed monthly earnings (AIME). The formula for AIME averages your income (up to the $80,000 cap) over the 35 years of highest actual earnings, adjusted for wage inflation (i.e., adjusted for changes in an index of the average level of wages). Second, the computer will calculate your primary insurance amount (PIA) from your AIME. If you are near the bottom of the income distribution your PIA maybe as high as 84% of your AIME, which means that your monthly Social Security Benefits on retirement will come quite close to replacing your earnings during your working years. If your income is close to the $80,000 cap, the monthly benefit may replace less than 45% of your income. If you are married, your spouse has the option of receiving 50% of your PIA or the benefit calculated on the basis of his own earnings history, whichever is larger.

What are the implications of these procedures? The system is progressive in that low income workers get back a much higher benefit relative to what they have contributed then do higher earners. This provision enables low earners to retire with a measure of decency. This transfer is financed by a regressive tax structure because of the cap of $80,000 on the income that is subject to Social Security taxes. Someone making $650,000 per year receives the same benefits as someone making $80,000, which seems fair in that they will both have paid the same taxes into the system. But it may be unfair in the sense that the high income earner might be expected to contribute more toward the financing of the high replacement ratio for low income workers than does an individual with an income of $80,000. The provision of 50% spousal benefits helps cover the higher expenses of retirement for a married couple, which seems reasonable. But the procedure subsidizes the traditional single-earner household in that they receive 50% more in benefits than a bachelor with the same earnings history who has made the same tax contributions during the working lifetime. Because the Social Security tax rates have increased substantially over the years, there is considerable intergenerational inequity. Today's retirees receive a bargain on their Social Security taxes relative to the much higher taxes that younger generations of workers will be paying over their working lifes.

There are serious questions about the financial viability of the Social Security System. Currently the system is running a substantial surplus in that each year more in Social Security tax revenue is collected than is paid out in benefits. The Social Security Surplus is used to reduce the total Federal budget deficit. But all this is projected to change around 2016 when the increasing number of retirements will lead to a substantial increase in benefit payout without a corresponding increase in tax revenue.

There are also some efficiency questions arising from the Social Security System. Harvard Professor Martin Feldstein argues that the Social Security System for providing retirement benefits tends to discourage thrift in the United States. Private saving for retirement is reduced below what it would be if workers did not rely on Social Security to provide support on retirement. The reduction in private saving is not offset by increased government saving because the Social Security tax does not collect enough to fully fund the costs of future retirement benefit obligations. The result, charges Feldstein, is more consumption in the present and less saving for the future. The reduction in saving reduces funding for investment in plant and equipment, which lowers the rate of productivity growth.

7.4.2 *Measuring poverty*

What poverty is cannot be appreciated by someone who has not lived there. Statisticians do attempt to measure it. Official government statistics on poverty are based upon a procedure worked out by Social Security Administration statistician Molly Orshansky in the 1960s. She estimated the *poverty threshold*, which is the level of income that is used in separating those in poverty from those who are able to live above it. If a family's money income is below the poverty threshold, then every member of that family is considered to be living in poverty. The first step of Orshansky's ingenious but simple procedure for estimating the poverty threshold was to obtain from the Department of Agriculture information on the minimum cost of a nutritionally adequate diet for families of various sizes. She then multiplied the cost of this diet by a factor of three to obtain the poverty threshold for families of each size. She chose three as the multiplier because evidence from household surveys indicated that the typical lower class family spent about 1/3rd of its income on food. The resulting poverty lines are adjusted from year to year in order to net out the effect of inflation. In year 2000 the poverty threshold for a single person under age 65 was $8,959. For a single parent with two children the threshold was $13,974.

The *poverty ratio* is the proportion of the population that has an income below the poverty threshold. Figure 7.8 presents information about the extent of poverty in the United States, as measured by this procedure. The poverty ratio has fluctuated over the years, rising when recession hits (indicated by the shaded regions) and falling along with unemployment

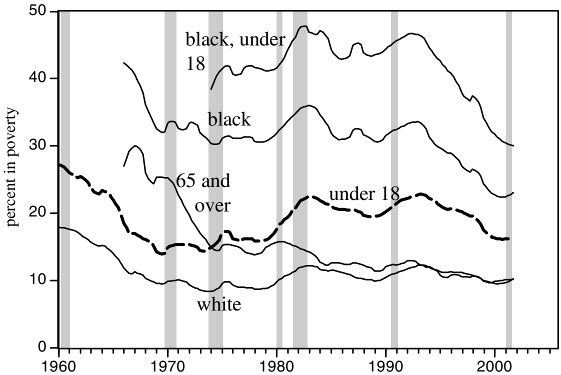


Fig. 7.8. Poverty rates
The expansion of Social Security and the introduction of Medicare led to a major reduction in poverty among the elderly in the late 1960s.

The reduction in poverty in the 1990s was due to the prolonged economic boom and welfare reform.

The shaded vertical stripes indicate periods of economic recession as identified by the National Bureau of Economic Research.

when the economy recovers. The data reveal a substantial reduction over the years in the proportion of the population living in poverty. Indeed, the proportion of the population of the United States living in poverty dropped in half from 1959 to 2000, most of the reduction taking place from about 1959 to 1973.

Poverty Gap

How much would it cost to lift every household in poverty up to the poverty threshold. The poverty deficit for anyone living in poverty is the shortfall of that person's income below the poverty threshold; i.e., poverty deficit = $\min(0, \text{poverty threshold} - \text{income})$. If we add up these individual poverty deficits for everyone living in poverty in the United States we obtain the *poverty gap*. The poverty gap has been estimated to be \$82.971 billion in 1997. That is a sizable sum, but our economy is large too. The poverty gap is only about 1% of GDP, the dollar value of all the goods and services produced in the economy. It would appear that only a small increase in tax rates would raise the funds required to cover the poverty gap and to eliminate poverty in America. Appearances are deceptive, however. If everyone were guaranteed a check equal to their poverty threshold it would obviously have an adverse effect on work incentives. The working poor with earned income below the poverty level would have no incentive to continue working. And many with incomes moderately above the guaranteed annual wage might decide that it would be better to take a cut in income to the poverty threshold in exchange for a life of leisure. Thus the task for policy makers is complex. We need a program for helping the needy that will not place a burden on taxpayers that is politically unacceptable and that will not have too negative an effect on work incentives.

The Aid for Dependent Children Program (AFDC), established in 1935, has served as the major source of financial support in the U.S. safety net. The program was administered by the states subject to guidelines established by the federal government. It was a matching program, for the federal government's financial support to each state was geared to how much the state decided to spend. Poorer states received three dollars of federal support for every dollar they decided to spend while wealthier states were matched dollar for dollar, receiving one dollar of federal support for each dollar of their own funds the state devoted to the program. At one time only single parent households were eligible for support. This provision was designed to focus the support on those most in need, but it also subsidized the breakup of families by encouraging unemployed fathers to move out so that their children would qualify for support. At one time benefits were reduced by 2/3rds of any earned income; in effect, this meant that recipients paid a 66% tax on their earned income. In 1979 this implicit tax was raised to 100%, which obviously discouraged those receiving income under this program from joining the work force. Figure 7.9 reveals that

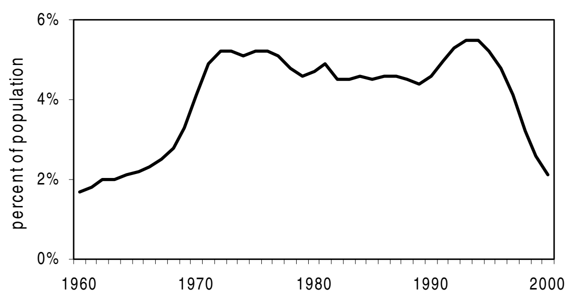


Fig. 7.9. AFDC/TANF
The case load of the Aid for Dependent Children program (AFDC) and its successor, Temporary Aid for Needy Families (TANF), has varied substantially over the years, stimulated by President Lyndon Johnson's War on Poverty and curtailed as a result of the remarkably low unemployment coupled with welfare reform during the Clinton years.

there have been dramatic shifts in enrollments in the Aid for Dependent Children Program. These changes reflect major shifts in public policy.

Lyndon Johnson's War on Poverty

President Lyndon Johnson declared that an all out "War on Poverty" would be the major thrust of his presidency, but he became mired in the war in Vietnam. Major changes in welfare programs did take place during Johnson's administration. In the mid 1960s he established the Department of Housing and Urban Development, Medicaid to provide medical services for the poor, and Medicare to provide health insurance for the retired. The work-study program was established to help provide jobs to help needy college students finance the costs of their education. Note that while such programs as Medicaid obviously help the recipients, these programs may not directly affect the official poverty ratio because that concept is based on money income before taxes are deducted and does not include non-cash benefits.

Clinton ends welfare as we know it

Another fundamental change in welfare was implemented in 1996 when President Clinton moved "to end welfare as we know it" by signing into law the Personal Responsibility and Work Opportunity Reconciliation Act. This act shifted the focus of welfare programs from income assistance to employment assistance. Placing people in a job became the primary goal of welfare programs. The Aid to Families with Dependent Children (AFDC) program was replaced with Temporary Assistance to Needy Families (TANF), which imposed strong work requirements and time limits on eligibility for assistance. The act also shifted the focus to providing block grants to the states instead of matching grants. Block grants provide a fixed amount of money to each state and give the states a wider degree of discretion as to how the funds are spent. As can be seen from Figure 7.9, there was a dramatic decline in welfare enrollments during the Clinton presidency. The drop in welfare enrollments was supported by the exceptionally low unemployment rates, which meant that jobs were available even for those with few skills and little work experience.

7.4.3 *Tax Policy*

As was revealed by Table 7.3, taxes have a significant impact on income inequality. A tax is said to be proportional if the ratio of the tax collected to income is the same for rich and poor alike. Neither the Gini inequality ratio nor the T20/T80 ratio is affected by a proportional tax. A tax is said to be *progressive* if the proportion of income that is collected increases with income. It is *regressive* if the tax to income ratio is greater for the poor than for the rich. The gasoline tax, sales taxes, and the excise tax on telephone services are examples of regressive taxes. The federal income tax appears to be progressive because the rates are much higher for people in the upper income brackets. But while the rates are progressive, in practice the presence of a variety of exemptions and credits partially offset the progressive rates.

Tax Expenditure

Taxes are, of course, the way in which the government raises money to finance its expenditures. But in addition to providing direct financial assistance, the government can act indirectly to support favored activity by exempting certain types or sources of income from taxation. For example, the government could use tax revenue to provide a child care allowance to enable parents to work. But instead it allows working parents to claim a child care credit of up to $1,000 for the amount they spend on a child's care expenses. In year 2000 the Child and Dependent Care Credit added up to $24.65 billion of tax revenue that was not collected. This $24.65 billion did not show up directly in the government's budget as an expenditure and it was $24.65 billion that did not show up in tax revenue. Similarly, a deduction for charitable contributions can be claimed when computing ones personal income tax — this exclusion added up to $2,730 billion in 2000. The exclusion of mortgage interest on owner-occupied houses cost the internal revenue service $55.1 billion. Step-up of basis of capital gains on death, discussed in Chapter 5.3.3, cost $27.09 billion. Tax expenditures are defined as revenue losses attributed to provisions of the tax laws which allow a special exclusion, exemption, deduction or credit in computing your tax liabilities. They are estimated to have cost Uncle Sam about a third of tax revenue in 2000.[12]

Earned Income Tax Credit (EITC)

The negative income tax, discussed in Chapter 4.4.4, is a reform that has been more popular with economists than with politicians or the public.

[12]The estimates are compiled annually by the Congressional Budget Office and published in the *Statistical Abstract of the United States.*

But the Earned Income Tax Credit, a feature of the tax code initiated during the Nixon administration and expanded greatly in 1993, does capture something of the spirit of the negative income tax. The Earned Income Tax Credit provides for a reduction in taxes for low-income earning families, but differs in two major respects from the negative income tax proposal. First, the EITC is only available for those who work and does not directly support those who are unemployed and the poor who are unable or uninterested in work. Second, the support in effect matches earned income but only up to a point. For example, in 2001 a family with one child and an income of $7,100 would receive a credit of $2,428. If their income was less the credit would be reduced proportionately. If their income was above $7,124 but below $10,750, the credit would stay at $2,428, but the credit is gradually reduced for incomes above $10,750 and is zero for an income of $28,300 or more. Thus the credit supplements hourly wages up to the $7,100 break point but reduces hourly earnings if the family's income is above $10,750 but below $28,300. This is all very complicated, but the IRS will compute the EITC for low-income earners requesting assistance.

Summary

1. Markets generate income inequality, but some more than others: In the United States the top 20% of the population receive almost 50% of total income. The bottom 20% receive less than 4%. While in Brazil and Mexico income is much less equally distributed than in the United States, income is less equally distributed in the United States than it is in most industrialized countries, including Japan, England, Germany, and France.
2. There are a number of different measures of income inequality: One of the simplest is calculated by first ranking the population by income and then dividing it into equally sized groups, called quintiles. We then measure income inequality by calculating the ratio of the average income received by those in the top 20% of the population (upper quintile) divided by the average income received by those in the bottom 20% (bottom quintile). This ratio, T20%/B20%, is 32.1 for Brazil.
3. The Lorenz curve is an ingenious graphical procedure that clearly displays the degree of income inequality. Table 7.2 clarified the concept by showing how it is constructed with income for the Untied States.

4. A comparison of Lorenz curves on Figure 7.3 revealed that while India and the United States have dramatically different levels of per capita income, the degree of income inequality is about the same. Lorenz curves plotted on Figure 7.5 suggest that eliminating income inequality within every country in the 82 nation sample would lead to only a modest reduction in world income inequality. Inequality among nations is much more pronounced than inequality within countries.
5. The Gini Coefficient is a measure of income inequality. It would be zero for a country in which everyone received the same income; it would be 100% for a country in which one person (a super Bill Gates) received the entire income. It equals twice the area between the Lorenz curve and the 45^0 line that would be the Lorenz curve *if* everyone were to receive the same income, as was illustrated on Figure 7.2.
6. Why are some workers paid so much more than others? The easy answer is to say that it depends on the demand and supply of workers, but it is necessary to look behind the demand and supply curves. The factors determining labor supply were considered in Chapter 4.4.4. In this chapter we looked at the demand side of the labor market, learning that under competition the wage is equal to the value of the worker's marginal product. This is the marginal productivity theory of wages invoked by Cobb and Douglas in arguing that their production function explained the constancy of labor's share, which is the proportion of national income that is received by workers. Efficiency wage theorists argue that in practice employers find it advisable to pay more than the market-clearing wage in order to reduce absenteeism and on-the-job shirking.
7. A monopsony is a firm that faces an upward sloping supply curve for labor or some other input. When our monopsonist decides to hire more workers it will bid up the wage rate. In order to avoid worker unrest, it will have to pay workers already on the payroll the higher wage. Because this means that the cost of hiring an additional worker, the marginal labor cost, may be substantially above the wage rate, the workers will be paid less than the value of their marginal product; this means that by Pigou's definition, those who work for a monopsonist are exploited. There are two surprising implications of the monopsony model: (a) An increase in the minimum wage may lead to an *increase* in employment. (b) When unions push up wages, the result may be higher employment.

8. The distribution of income is decisively affected by government tax and transfer programs. Social security is largely financed by a payroll tax. Currently the Social Security system collects more in taxes than it pays out, and the surplus is used to help cover the government's deficit. But this situation is projected to reverse around 2018 when an increase in retirements will lead to a shortfall of Social Security tax revenue below expenditures.
9. The official poverty threshold is three times the estimated cost of a nutritionally adequate diet. The poverty ratio — the proportion of the population with incomes below the poverty line — has dropped in half over the last forty years. The poverty gap, the sum required to raise the income of every family above the poverty level, is not all that large, but just guaranteeing everybody an income of at least the poverty level would have a serious adverse effect on work incentives.
10. President Lyndon Johnson's "War on Poverty" led to the establishment of Medicare, providing health care for the elderly, and Medicaid, which provides health care for the poor. Three decades later, during the Clinton administration, the focus of welfare shifted from providing income assistance to providing work assistance. The Johnson administration saw a major increase in the proportion of the population on welfare while the Clinton administration witnessed an abrupt decline.
11. The Earned Income Tax Credit, which supplements the income of the working poor by reducing their income tax bill, is related to the Negative Income Tax proposal, a reform that has received support from many economists. One major difference is that the Earned Income Tax Credit does not provide support for the non-working poor.

Key Concepts

Earned Income Tax Credit, 328
efficiency wages, 318
exploitation, 314
Gini coefficient, 307
GNP per capita (GNP/pc), 301
infant mortality rate, 296
involuntary unemployment, 318
labor's share, 315
Lorenz curve, 306
marginal labor cost, 311
marginal revenue product of labor, 310
median income, 304
monopsonist, 310
poverty gap, 324
poverty ratio, 323
poverty threshold, 323

Exercises

1. Five citizens live in the mythical kingdom of Econoland. Their incomes are \$10, \$20, \$30, \$40 and \$50.
    a. Plot the Lorenz curve for Econoland. Show your computations. Label everything.
    b. Estimate the approximate value of the Gini coefficient from your graph.
    c. Does the distribution of income in Econoland appear to be more equal, less equal or about the same as the distribution of income in the United States?
2. Suppose that the wage that a monopsonist has to pay its workers as a function of the number of workers employed is

$$w(L) = w_0 + w_1 L\,.$$

   The firm's production function is

$$q(L, K) = \alpha L^{\lambda} K^{\lambda'}\,.$$

   Assume initially that K is fixed at 1.
    a. Determine total labor costs and marginal labor costs as functions of L.
    b. Determine total revenue and marginal revenue as functions of q if the inverse demand function is $p(q) = a - bq$.
    c. Determine the marginal product and the marginal revenue product of labor functions.
    d. Find the level of employment that will maximize the firm's profits.
    e. What will happen to the profit maximizing level of employment if K increases to 2?
3. Draw a graph similar to Figure 7.6, but assume that there is competition in the market in which the firm hires its labor and in the market where it sells its product; i.e., $\partial w/\partial L = \partial p/\partial q = 0$.

4. The Econoland welfare program pays each citizen \$30 per day. However, each citizen's benefit is reduced dollar-for-dollar by the income earned, up to \$30 per day. For example, someone earning \$20 per day will receive a \$10 welfare payment. Those earning \$30 or more will not receive any welfare payment.

   a. Draw an indifference map for a utility maximizing recipient who stops working after the welfare program is established.
   b. Draw an indifference map for a utility maximizing recipient who keeps working after the program is established.
   c. Can you draw an indifference map for a previously idle recipient who starts working after the program is established? Explain.

5. The old welfare program is replaced with a "negative income tax." $T = -\$40 + 0.3Y$ where Y is the before tax daily income. For example, someone with zero earned income will receive \$40 (a negative tax). Someone who earns \$80 will pay a tax of $T(80) = -40 + 0.3 \times \$80 = -\$16$. Therefore, the IRS will pay this citizen \$16 so total income will be \$96.

   a. Is the negative income tax "progressive" or "regressive"? Explain.
   b. What is the break even level of income for this tax?

6.# Compute $g = 2\sum p[P - L(p)]$ with the data in the last column of Table 7.2. Does g equal the Gini coefficient for the United States in 2001?

   Prove that this relationship holds in general, for *any* distribution of income.

8

Monitoring Economic Performance

8.1 Overview — Macro versus micro

Macroeconomics is the branch of economics that examines the overall performance of the economy. It is the study of economic aggregates on a national scale — growth, income, employment, and prices. Macroeconomists try to explain how the economy grows and fluctuates over time and why we have inflation. In looking at the broad picture, macroeconomists usually build upon the foundations provided by microeconomics, which is defined as the study of the behavior of firms and households and how they interact in the market place. Our study of microeconomics in the first half of this book has prepared us for the analysis of macroeconomic issues.

How do economists measure overall economic performance, the extent of unemployment and the speed of inflation? What can graphs and tables of economic data tell us about the economic history of the United States economy and current economic developments? Learning how economists appraise the economic condition is a first step toward finding out what determines how well the economy functions and how economic policy makers can, for good or ill, influence the performance of the national economy.

8.2 An economic report card

Table 1 constitutes an economic report card providing in successive columns a decade-by-decade snapshot of how well the United States has done in terms of a variety of indices of economic performance. This table and the accompanying charts on unemployment and inflation provide a convenient framework for a brief historic review of economic ups and downs. While some of the concepts presented on the table are easily understood, others will be clarified as we progress through this chapter.

- The 1930s depression decade was by far the worst, for the unemployment rate soared to an all time high of 25% by 1933 — one worker in four could not find a job! As can be seen from Table 8.2, recovery was painfully slow. And long before business was operating anywhere close to capacity the economy suffered from a sharp relapse into the severe 1937–1938 recession. Even by 1940 the unemployment rate was a disgraceful 15%.
- Sparked by the need to produce the necessities of war, the 1940s may well have been our economy's finest decade. Output expanded rapidly, even as the youngest and finest marched off to war. The dramatic rise in output was made possible by a remarkable increase in the number of women joining the labor force and the return of retirees to the workplace to help the war effort; and the army of unemployed was put to work. By 1944 the unemployment rate had dropped to an unprecedented 1.2%. In retrospect, perhaps the most surprising aspect of the wartime experience was that there was not more inflation — price controls and rationing worked to keep a lid on prices until after the end of the war. While

Table 8.1. Economic performance — the record of eight decades.

		1929	1933	1940	1944	1950	1960	1970	1980	1990	2000
REAL OUTPUT (billions 96 $)											
1	Real GDP (1)	822	603	981	1,714	1,687	2,377	3,578	4,901	6,708	9,224
2	GovtSpend on G&S (2)	110	118	182	1,196	367	661	931	1,021	1,387	1,573
3	Govt/GNP (3 = 1/2)	0.13	0.20	0.19	0.70	0.20	0.28	0.26	0.21	0.21	0.17
4	Net Exports of G&S	−11	−11	−2	−34	−10	−21	−64	10	−16	−399
5	Gross Private Investment	94	18	111	52	233	273	436	655	907	1773
PRICES											
6	CPI (1982–84 = 100)	17.1	12.9	14.0	17.6	24.1	29.6	38.8	82.4	130.7	172.2
7	Inflation rate (annual %)	0.0%	−6.8%	1.2%	5.9%	5.4%	2.1%	2.7%	7.8%	4.7%	2.8%
WORKER PRODUCTIVITY, Business Sector (1992 = 100)											
8	Output per hour		20.0	26.2	30.2	36.9	48.8	67.0	80.4	92.5	118.6
9	Real Compensation per hour					43.4	60.0	78.8	89.4	96.5	108.8
STOCK MARKET											
10	S&P Composite	26.0	9.0	10.7	12.4	18.4	55.9	83.2	118.8	334.6	4,126.8
UNEMPLOYMENT RATE (%, Civilian)											
11	All	3.2	25.2	14.6	1.2	5.3	5.5	4.9	7.1	5.4	4.0
12	Women					5.7	5.9	5.9	7.4	5.4	4.1
13	Black					9.0	10.2	8.2	13.1	10.1	6.7
FAMILY INCOME (2000 dollars)											
14	Median Income (all)					22,728	41,838	41,838	41,974	45,392	50,890
15	Black Median Income							27,313	25,960	27,506	34,192
16	Black as % of White Median Income							65.3%	61.8%	60.6%	67.2%
POVERTY											
17	% Persons in Poverty						22.2%	12.6%	13.0%	13.5%	11.3%
18	Black % in poverty							33.5%	32.5%	30.7%	22.0%
19	Female households, no husband present, % in poverty						42.4%	32.5%	32.7%	33.4%	24.7%

Sources: *Historical Statistics of the United States: Colonial Times to 1970*, U.S. Department of Commerce, 1974; Bureau of Labor Statistics.

the process of converting from a garrison to a peacetime economy was far from easy, the postwar recession of 1948–1949 was surprising mild, particularly when compared with the experience of the 1930s.

- The decade of the 1950s opened with the Korean War, an intense but relatively brief conflict followed by a minor post-war recession. In retrospect, the 1950s was a decade of remarkably successful economic performance. While there was much concern about inflation, in fact this decade turns out to have been less inflationary than the peacetime 1970s.
- In the early 1960s President Kennedy's Council of Economic Advisers, arguing that a deliberate effort should be made to push the economy closer to full employment without inflation, asked Congress for a tax cut designed to stimulate the demand for the nation's output. Kennedy could not get his tax-cut panacea through congress — the legislation was finally passed in 1964 during Lyndon Johnson's presidency. Later in the decade the economy was over-heated by excessive government spending demanded by the unanticipated expansion of the Vietnam War effort.
- Inflation was the big surprise in the 1970s. Early in his administration Nixon took measures to slow down the economy in order to stem the inflationary tide — the deflationary medicine had the unfortunate side effect of pushing the economy into recession, but the inflation forces were not contained. Then, in the fall of 1973, the Organization of Petroleum Exporting Countries (OPEC) engineered sharp increases in the price of oil that precipitated massive world-wide inflation.
- The 1980s opened with high inflation, which was brought under control through restrictive monetary policy. An unfortunate side-effect of the anti-inflationary medicine was a massive rise in the unemployment rate to the worse recession levels since the Great Depression of the 1930s.
- The last decade of the 20th century was a complete surprise, for professional economists no less than for financial investors. At the start of the decade many of the wisest minds anticipated that a stagnant United States economy would be surpassed by Japan. Quite the contrary, Japan's economy remained mired in severe recession throughout the decade while the American economy experienced unprecedented output growth in the longest boom in modern times coupled with spectacular stock market returns. And contrary to the conventional wisdom of economists, the United States was able to enjoy the lowest levels of unemployment since World War II without serious inflationary consequences.

Our brief chronology clearly demonstrated that no two decades are alike. Indeed, the experience of each succeeding decade has been a surprise not only to Wall Street speculators and the public generally but also to professional economic forecasters. The fundamental changes from one decade to the next have proved unpredictable not only because the movements of the economy are shaped in considerable measure by wars and other unanticipated political developments but also because the economy's underlying laws of motion are only imperfectly understood. At times established economic thought has proved inadequate to explain the unfolding of events and a fundamental rethinking has been required. Such was the case in the 1930s, when existing theory could not explain the protracted period of profound unemployment. Again, in the 1970s existing theory had difficulty accounting for the great inflation. And it is difficult to explain why the American economy managed to achieve 4% unemployment without inflation in the 1990s. Advances in economic knowledge are frequently the result of research sparked by the need to rethink the conventional economic wisdom in order to provide an explanation for unanticipated economic developments.

Table 8.1 provided only the briefest statistical summary of economic developments. Table 8.2 takes a close up view of the Great Depression of the 1930s. Table 8.3 reveals the way in which the American economy mobilized for all out war in World War II. These tables contain much useful insight, but they also involve technical concepts requiring clarification. The discussion in the remainder of this chapter will help you to interpret the data.

8.3 Unemployment

As a first approximation, the unemployment rate can be defined as the proportion of those willing and able to work who cannot find jobs. Chapter 1's Figure 1.2 looked at annual data on the unemployment rate over the past century. That graph clearly revealed that the unemployment rate has been subject to gross fluctuations, ranging from a low of 1.2% in World War II to the high of 25% in the Great Depression of the 1930s. The disastrous rise in unemployment in the 1930s was the worst in the history of our country, even exceeding the Crisis of 1893. While the United States has escaped depression levels of unemployment since World War II, the more detailed monthly data on Figure 8.1 makes it clear that in modern times the United States economy has been subjected to severe bouts of unemployment.

Table 8.2. The great depression of the 1930s.

(All figures in billions of dollars unless otherwise indicated.)

		1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	Change 1929–1933
1	Unemployment rate (%)	3.2	8.9	16.3	24.1	25.2	22	20.3	17	14.3	19.1	17.2	
2	Gross Domestic Product	103.7	91.3	76.6	58.8	56.4	66.0	73.3	83.7	91.9	86.3	92.0	−45.6%
3	GDP (1996 dollars) (Y)	822.2	751.5	703.6	611.8	603.3	668.3	728.3	822.5	865.8	835.6	903.5	−26.6%
4	Consumption	625.7	592.3	574.3	523	511	546.9	580.6	639.6	663.5	652.6	689	−18.3%
5	Investment	93.6	62.5	39.2	11.8	17.5	31.6	58.4	74.9	93.6	61.9	79.6	−81.3%
6	Net Exports	−10.5	−10.7	−10.5	−9.9	−11.0	−9.5	−18.0	−16.3	−15.1	−5.4	−5.5	4.8%
7	Government Purchases	110.1	121.3	126.6	122.4	118	133	137	158.9	153.2	164.6	179.7	7.2%
8	Disposable Income	585.8	542.2	519.7	449.8	437.0	462.0	505.2	565.9	585.5	547.6	590.3	−25.4%
9	GDP Deflator (1996 = 100)	12.5	12.1	11	9.7	9.5	10.3	10.6	10.6	11.2	10.9	10.8	−24.0%
10	Inflation Rate (%)		−3.2%	−9.1%	−11.8%	−2.1%	8.4%	2.9%	0.0%	5.7%	−2.7%	−0.9%	
11	CPI (1982–84 = 100)	17.1	16.7	15.2	13.7	13	13.4	13.7	13.9	14.4	14.1	13.9	−24.0%
12	Inflation Rate, CPI (%)		−2.3%	−9.0%	−9.9%	−5.1%	3.1%	2.2%	1.5%	3.6%	−2.1%	−1.4%	
13	Hourly earnings (manufacturing) $	0.56	0.55	0.51	0.44	0.44	0.53	0.54	0.55	0.62	0.62	0.63	−21.4%
14	Real Wage (1982–84 $)	3.27	3.29	3.36	3.21	3.38	3.96	3.94	3.96	4.31	4.40	4.53	3.4%
15	Stock Market (S&P, 1941–43 = 10)	26.02	21.03	13.66	6.93	8.96	9.84	10.6	15.47	15.41	11.49	12.06	−65.6%
16	Interest rate on business loans	5.8%	4.9%	4.3%	4.7%	4.3%	3.5%	2.9%	2.7%	2.6%	2.5%	2.1%	
17	Corporate Bond Yield	4.7%	4.6%	4.6%	5.0%	4.5%	4.0%	3.6%	3.2%	3.3%	3.2%	3.0%	
18	Real Rate of Interest	4.7%	6.9%	13.6%	14.9%	9.6%	0.9%	1.4%	1.7%	−0.3%	5.3%	4.4%	
19	Bank Reserves	2.4	2.4	2.1	2.4	2.6	4.0	5.7	6.7	6.9	8.7	11.5	8.1%
20	Required Reserves	2.3	2.3	2.0	1.9	1.8	2.3	2.7	4.6	5.8	5.5	6.5	−22.4%
21	Money Supply (M_1)	26.6	25.8	24.1	21.1	19.9	21.9	25.9	29.6	30.9	30.5	34.2	−25.2%
22	Real money supply: M_1/CPI	212.8	213.2	219.1	217.5	209.5	212.6	244.3	279.2	275.9	279.8	316.7	
23	Velocity (ratio of pY/M_1)	3.9	3.5	3.2	2.8	2.8	3.0	2.8	2.8	3.0	2.8	2.7	

Note: Line 16 is the rate charged by banks in major cities on short-term business loans.

Table 8.3. The economics of World War II.

		1939	1941	1944	1945	1946	1947	Growth/ Annum 1939–1945
1	GDP (real, 1987 dollars)	840.7	1,070.6	1,670.0	1,602.6	1,272.1	1,252.8	11.4%
2	GDP (nominal)	90.8	125.0	211.0	213.1	211.9	234.3	15.3%
3	Govt spending (G)	13.5	24.8	96.9	83.3	29.2	26.2	29.7%
4	Defense Spending	1.3	13.8	87.5	73.7	16.4	10.0	96.0%
5	Government Deficit (Federal)	3.4	5.6	48.7	48.7	17.0	2.9	
6	Government Debt (Federal)	48.2	57.5	204.1	260.1	271.0	252.0	32.4%
7	Debt/GDP	53.1%	46.0%	96.7%	122.1%	127.9%	107.6%	
8	G/GDP	14.9%	19.8%	45.9%	39.1%	13.8%	11.2%	
9	GDP Deflator(1972 = 100)	10.8	11.7	12.6	13.3	16.7	18.7	3.5%
10	Consumer Price Index	13.9	14.7	17.6	18.0	19.5	22.3	4.4%
11	Civilian Labor Force	55,230	55,910	54,630	53,860	57,520	60,168	−0.4%
12	Employment	45,750	50,350	53,960	52,820	55,250	57,812	2.4%
13	Unemployment	9,480	5,560	670	1,040	2,270	2,356	−30.8%
14	Unemployment Rate (%)	17.2%	9.9%	1.2%	1.9%	3.9%	3.9%	
15	Money supply: M_1	34.2	46.5	85.3	99.2	106.5	111.8	19.5%
16	M_2	49.3	62.5	106.8	126.6	138.7	146.0	14.4%
17	currency	6.0	8.4	21.2	25.3	26.5	26.6	27.0%
18	velocity: $v_1 = Y/M_1$	2.7	2.7	2.5	2.1	2.0	2.1	
19	$v_2 = Y/M_2$	1.8	2.0	2.0	1.7	1.5	1.6	
20	Interest rate: U.S. Gov Bonds	2.41%	2.05%	2.48%	2.37%	2.19%	2.25%	
21	T-bills	0.023%	0.014%	0.375%	0.375%	0.375%	0.375%	

Note: All dollar magnitudes in billions.

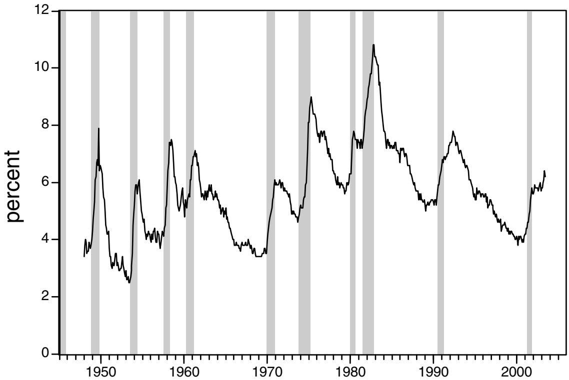


Fig. 8.1. Unemployment in the U.S. since World War II
The unemployment rate is the ratio of those who cannot find work to those who are willing and able to work.
Note: The shaded vertical stripes indicate periods of recession as identified by the National Bureau of Economic Research.

Observe that the movements in the unemployment rate do not resemble simple harmonic motion, such as the sinusoidal waves often displayed on an oscilloscope; quite the contrary, there is a sizable erratic or perhaps random component. There is considerable variation both in the magnitude of the occasional peaks in the unemployment rate and in the length of time between successive peaks. The unemployment rate usually falls to exceptionally low levels during wartime, which has led some radical critics to complain that the capitalistic system requires wars in order to keep the unemployment rate under control.

8.3.1 *Estimating unemployment*

When the U.S. entered the great depression of the 1930s there were no systematic estimates available of the number unemployed, except those provided by the decennial census. The Current Population Survey was developed to meet this need. Each month the Census Bureau interviews approximately 50,000 households in a Herculean effort to estimate accurately the number of unemployed.

The monthly Current Population Survey focuses on a fundamental question asked of every member of the household who is 16 years of age or older: "What were you doing last week, working, looking for work, or something else?" On the basis of responses to this and a variety of related questions each person aged sixteen and over is classified as being either in or out of the labor force. The labor force includes those who actually have jobs plus those who are unemployed. Excluded from the labor force are students who are not interested in working, spouses who are not currently interested in gainful employment, and senior citizens enjoying their retirement. Those in the labor force are classified as being either employed or unemployed on the basis of a finely specified set of criteria. The unemployed are those who had no employment, were available for work except for temporary illness, and had made specific efforts, such as contacting employers, to find employment during the last four weeks.

Examples

- A college junior looking for a summer job while home for spring vacation is not classified as unemployed. This student is not in the labor force because she is not immediately available for work.
- Neither a college undergraduate who works six hours a week in the computer center nor a factory worker cut back to a 15 hour week because the factory where she works has to trim its excessive inventories of unsold goods are counted among the ranks of the unemployed. The primary focus is on the number of jobs and not on the number of hours worked.
- A worker who is on temporary layoff but not looking for work because he is scheduled to return to work within two weeks is classified as unemployed. This worker is included in the labor force even though he is not actively looking for a new job.
- An auto-factory worker who found that the early retirement bonus offered by his employer was too good an offer to refuse will be classified as not-in-the-labor-force rather than unemployed.
- A steel mill worker who was laid off after nine years on the job is not counted as unemployed if she does not bother looking for work because she knows no jobs are available. If she were actively looking for a job but couldn't find one she would be counted as an unemployed member of the labor force.

Some of these distinctions are rather picayune; some are controversial. Some would argue that the last case, the "discouraged worker," should be counted among the army of unemployed. But to count as unemployed the discouraged workers who drop out of the labor force because they say they believe no work is available would risk making measured unemployment excessively sensitive to a subjective judgment about the availability of work that might well be incorrect. The fact that one is counted as employed even when one works only a fraction of the number of hours one would like to put in means that the unemployment rate may understate the severity of recession. Fortunately, the *Bureau of Labor Statistics* (BLS), publishes a number of alternatives to the basic measure of the unemployment rate in order that the effects of these distinctions can be fully appreciated.[1]

Key labor force concepts:

- Civilian non-institutional population: Number of civilians age 16 and over who are not incarcerated in jails or other institutions. Excludes resident members of the armed forces.[2]
- Labor Force: Those members of the civilian non-institutional population who are either employed or looking for work.
- Employed: Those members of the labor force who have jobs.
- Unemployed: Labor force − Employed
- Unemployment Rate: Unemployed/Labor Force
- Employment Population Ratio: Employed/Civilian non-institutional population
- Labor Force Participation Rate: The proportion of the non-institutional civilian population that is in the labor force, either having jobs or looking for work; i.e., labor force/population.
- Discouraged Workers: Those who have lost their jobs but do not look for work because they believe no appropriate jobs are available. Since they are not looking for work they are classified as not in the labor force rather than counted among the unemployed.

Data on key unemployment concepts are reported in Table 8.4. Note the marked contrast between the months of December 1982 and December 2000. December 1982 was the worst month of the 1982 recession — one worker in ten did not have a job. This "Reagan Recession" (few presidents escape having a recession named in their honor) was the most severe experienced since the 1930s.

[1]The data are published regularly in the *Monthly Labor Review* and on the Bureau of Labor Statistic's website, http://www.bls.gov/data/home.htm. Historical data is also available in the annual *Economic Report of the President*.

[2]Data are also available that include as employed members of the labor force the "resident armed forces" (i.e. those in military service who are stationed in the United States); obviously, this increases both the level of employment and the size of the labor force; it lowers the reported unemployment rate by about 1/10th of one percent.

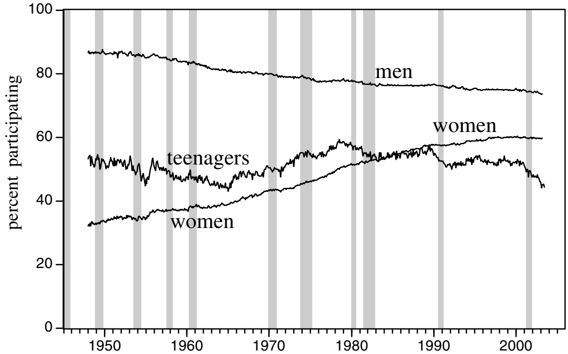


Fig. 8.2. Labor-force participation rates
The labor-force participation rate is the proportion of the population 16 years old or older that is working or looking for work.

Table 8.4. Key labor force concepts.

	Dec 1982	June 1991	Dec 2000
Civilian non-institutional population	173,199	187,977	210,743
Civilian labor force	110,962	124,652	141,544
Employed	99,055	118,238	135,888
Unemployed	11,907	6,414	3,230
Unemployment rate	10.7%	5.1%	4.0%
Employment/Population	57.2%	62.9%	64.4%
Labor force participation rate	64.1%	66.3%	67.2%

Source: Bureau of Labor Statistics.

Graphs help us to appreciate how the key unemployment categories have evolved overtime. Observe from Figure 8.2 that over the last half century there has been a steady rise in the female labor force participation rate while males have slacked off and teenagers have been erratic. As a result, Figure 8.3 reveals, there has been a gradual increase in the female proportion of the total labor force. Figure 8.4, which shows how different demographic groups are affected by unemployment, makes it clear that unemployment discriminates.

Scatter-plot

The black and white unemployment data on Figure 8.4 are re-plotted as a scatter-plot on Figure 8.5 in order to more clearly reveal the relationship between the two. Instead of having time plotted on the abscissa and two

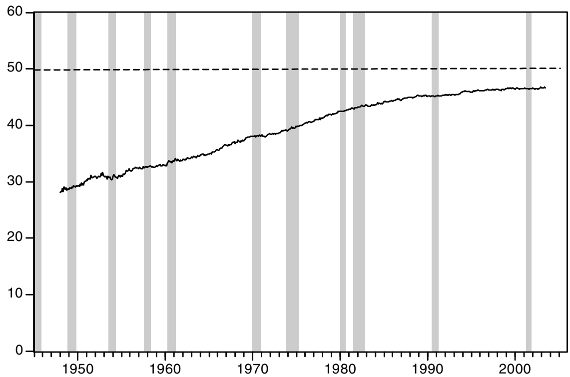


Fig. 8.3. Women as a percentage of the labor force
In the last half of the 20th century women as a percentage of the total labor force increased from 32% to 47%.

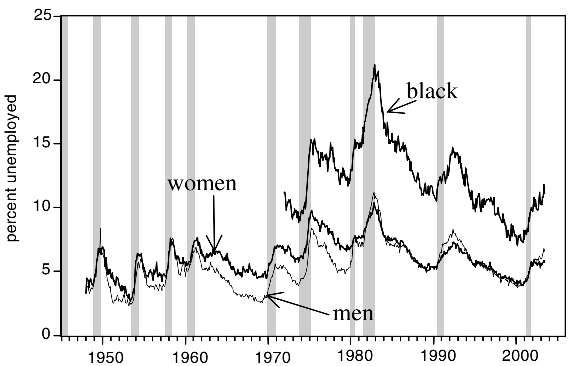


Fig. 8.4. Unemployment rates by sex and race

curves, one for black and one for white unemployment, this *scatter-plot* has one dot for each date. Each dot on Figure 8.5 reveals the unemployment rate of whites on the abscissa and blacks on the ordinate for a particular year, i.e., each point has coordinates (white unemployment, black unemployment). Observe that the points on this scatter-plot are closely clustered around a straight line because the two rates move very closely together. But the line has a slope of about two because black unemployment is always approximately twice the white rate. This graph makes clear that blacks and whites suffer together during periods of unemployment, but blacks suffer twice as much. Achieving prosperity benefits all, or as the economic

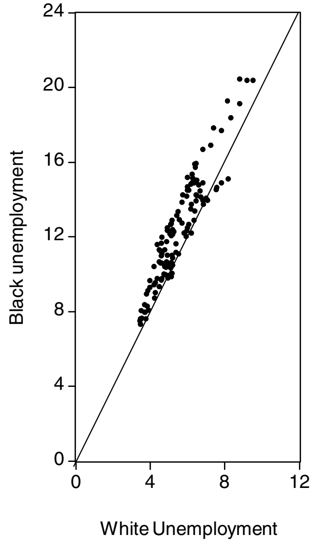


Fig. 8.5. Scatter-plot: black versus white unemployment rates
Each dot represents unemployment in a particular year: $(U \;\; , U \;\;)$. The line through the origin with a slope of +2 approximates the data points fairly well, suggesting that blacks tend to suffer at least twice as much as whites when unemployment climbs. When boom times role, whites and blacks both gain.

journalists are fond of saying, "a rising tide raises all boats" — but it is clear from the graph that some benefit much more than others. It is also true that a recession in the United States means depression levels of unemployment for Black Americans.

Conducting a large monthly survey to measure the characteristics of the labor force is an expensive operation, but the resulting numbers are extremely useful in appraising current economic conditions. An alternative measure, the number of laid off workers receiving unemployment benefits, would be much less expensive to calculate but is not worthy of serious consideration because of several limitations: First of all, the number of insured unemployed obviously is affected by changes in unemployment benefit eligibility rules, which vary both over time and from state to state. Second, the number receiving unemployment benefits excludes two very important groups: It excludes students graduating from school and other new entrants into the labor force who have not worked long enough to qualify for unemployment benefits. It also excludes the long-term unemployed who have exhausted their unemployment benefits.

8.3.2 *Fuzz on the data*

Producers of the evening news know they must reserve time on the first Friday of each month for a discussion of the unemployment rate — that is the day when the Bureau of Labor Statistics (BLS) releases the survey results for the previous month at a press conference in Washington. Whether down, up or holding steady, the latest unemployment rate is almost certain to grab the headlines. But are the unemployment numbers all that precise? Changes in methodology may affect the estimates; for example, the shift in January of 1995 from pencil and paper questionnaires to Computer Assisted Survey Information Collection (CASIC), by reducing survey error, may have increased the measured unemployment rate by 0.1% or 0.2%. Such one-time changes are not all that serious, but two significant sources of error that can cause delays in recognizing reversals in the pace of economic activity and contribute to mistaken economic policy demand discussion:

Sampling error

Because the unemployment estimate is based on survey evidence rather than a complete count of the population, the estimates are subject to sampling error — the estimate will depend on who happens to be selected in the sample for interviewing.[3] To keep the sampling error under control, the Census Bureau uses a sample of 50,000 households, which is several times the sample size of 1,000 or so commonly used by the public opinion pollsters in attempting to gauge public opinion and to predict election outcomes. The Census Bureau statisticians have calculated the probability that an observed change in the unemployment rate should be attributed to random sampling error rather than to an actual change in underlying economic conditions:

- There is about a 50% chance that a change from the preceding month of 1/10% or more in the unemployment rate estimate will be observed when in reality there has been no change in the actual unemployment rate.

[3] Every ten years the decennial census of population provides, in principle, a complete count of the unemployed. When the survey and census results are compared, the census count often turns out to be *less* precise than the sample! The problem is that respondent and enumerator errors are much higher with the decennial census, which is a major undertaking conducted by temporary employees who have not had as much experience as the staff that regularly conducts the monthly Current Population Survey.

This means that an observed increase in the unemployment rate from, say, 5.4% to 5.5%, may be just as likely to be due to chance measurement error as to an actual rise in the actual level of the unemployment rate.

- There is about a 15% chance that an observed change of two-tenths of a percent in the unemployment rate from the preceding month will be observed when there has actually been no change in the true unemployment rate.[4]

Because of the likelihood of false signals of recession, both economic policy makers and the public must be careful not to overreact to small month-to-month changes in the unemployment rate. Only if the movement persists for two or three months can it be concluded with much confidence that there is a genuine change in underlying labor market conditions rather than a false signal arising from noise in the data.

Seasonal adjustment

The numbers reported on the evening news on the first Friday of each month are not the raw ratios computed from the survey. Rather, the data have been *seasonally adjusted* in an attempt to remove the systematic movements in the data that characterize various seasons of the year. Seasonal adjustment removes the typical seasonal fluctuations so as to make it easier to spot changes in underlying economic conditions.

The nature of the seasonal adjustment process is clarified by considering the first two rows of Table 8.5, which contrasts the original unadjusted data with the seasonally adjusted numbers during the "Reagan Recession" of the early 1980s. Observe that the seasonally adjusted figure for December 1981 of 8.9%, is 0.6% higher than the unadjusted figure. The BLS reported a seasonally adjusted figure of 8.9% rather than 8.3% because

---

[4]Statisticians frame the problem in the following way: Let $d\ = U\ - U_{-1}$ denote the observed change in the estimated unemployment rate and δ the unobservable actual change. Let $\sigma\ =$ the standard deviation of d . Then the null-hypothesis that $\delta\ = 0$ should be rejected at the 5% level of significance only if $|d\ - \delta\ | > 1.96\sigma$, where $\sigma\ = 0.11\%$ when the unemployment rate is about 5%; i.e., we reject the null-hypothesis of no change in the unemployment rate only if the observed change is greater than $1.96 \times 0.11\% = 0.22\%$. For a smaller observed change we should conclude that the evidence does not establish that there is a change in the unobservable actual unemployment rate. Students studying statistics learn that the binomial theorem suggests that with independent random sampling the value of $\sigma^2 = 2\pi(1 - \pi)/n$, where π is the actual unemployment rate and n the sample size, but this is only an approximation because the Census bureau uses a complex stratified-cluster-rotational sampling procedure.

they estimated that the jobs generated for the holiday sales season typically make the unemployment rate fall artificially low in December of each year. The adjustment is in the opposite direction in January; the seasonally adjusted rate for January lying about 0.9% below the unadjusted rate because the end-of-holiday layoffs coupled with the customary tendency for bad weather to cause job cutbacks leads to a systematic tendency for the raw unemployment rate to rise each January by about 0.9%. Similarly, the raw survey data on unemployment generally show a sizable increase of almost 0.3% each June because of the influx of students into the labor force at the end of the school year. On the other hand, many students drop out of the labor force in September, which causes a systematic seasonal decline in unemployment.

The *Census-X12 (ARIMA)* computer program attempts to filter out these typical seasonal movements in generating the seasonally adjusted numbers that attract so much attention each month.[5] The task is complex because the seasonal movements evolve over time rather than being strictly constant.[6] The hope is that the seasonally adjusted figures will more accurately reveal underlying cyclical changes and provide a better indicator of recession or revival.

Unfortunately, the seasonal adjustment procedure is far from precise. The unemployment figures are subject to revision when additional information becomes available. When economic historians look back at recession episodes they have the luxury of being able to look at revised data that are much more precise than the preliminary unemployment numbers that were the best evidence available to economic policy makers and economic journalists at the time. But it is the preliminary numbers that guide policy

[5] The computer program is known as Census-X11 because it was originally developed at the Census Bureau. Subsequently the program was modified by using an autoregressive integrated moving average (ARIMA) procedure to extrapolate the series before processing it with Census-X12.

[6] The seasonal adjustment factors shifted when the due date of the Federal Personal Income Tax was changed from March to April 15th. The seasonal adjustment factors may also shift as a result of changes in business practices, such as a shift in the date in which new car models are typically introduced. They are also affected by demographic movements, such as a change in the proportion of teenagers in the population. The seasonal effect of the weather on employment in the construction industry has been dampened by a number of technological changes, including the development of concrete that can be poured when the temperature is below freezing. The Census-X12 program allows for a "moving seasonal" to capture such effects. However, when unemployment rises because of unusually severe weather the effect should not be counted as part of the "seasonal" because it is not a *typical* seasonal movement.

Table 8.5. Revisions of the seasonally adjusted unemployment rate.

	1981 Dec	1982 Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec	1982 Average
Initial Estimate:														
Seasonally Unadjusted	8.3	9.4	9.6	9.5	9.0	9.1	9.8	9.8	9.6	9.7	9.9	10.4	10.5	9.7
Seasonally Adjusted	8.9	8.5	8.8	9.0	9.4	9.5	9.5	9.8	9.8	10.1	10.4	10.8	10.8	9.7
Adjustment	0.6	−0.9	−0.8	−0.5	0.4	0.4	−0.3	0.0	0.2	0.4	0.5	0.4	0.3	0.0
Revised Estimate:														
Seasonally Adjusted	8.5	8.6	8.9	9.0	9.3	9.4	9.6	9.8	9.9	10.1	10.4	10.7	10.7	9.7
Revision	−0.4	0.1	0.1	0	−0.1	−0.1	0.1	0	0.1	0	0	−0.1	−0.1	0.0

Note: The adjustment row reports the seasonally adjusted minus the seasonally unadjusted unemployment rate. The revision row reports the revised less the unrevised seasonally adjusted unemployment rate.

makers in Washington and grab the headlines. To illustrate, the third row on Table 8.5 shows the revised seasonally adjusted unemployment data for the Reagan recession. The 0.4% decline in the unemployment rate from 8.9% in December 1981 to 8.5% in January of 1982, reported in the second row of Table 8.5, was welcome news when first released by the BLS. It was not until the revised figures were released in March that it became obvious that something had been fundamentally wrong with the preliminary seasonally adjusted data and that the economy was definitely not on the road to recovery. The revised numbers are more precise in part because the seasonal adjustment calculations can be refined on the basis of the additional evidence about typical month-to-month changes provided by more recent data on unemployment. Also, the survey evidence can be still more precisely processed when the next decennial census yields more accurate information about the proportion of the population in school and other demographic characteristics of the labor force.

The essential point to note is that the judgment of economic policy makers as well as the message of TV business commentators is based on imprecise evidence. Alarming short-term dips and pips in the preliminary unemployment rate sometimes turn out to represent only misleading "noise" in the data. Some studies suggest that a month-to-month increase in the unemployment rate of less than 2/10% or even 3/10% should not be regarded as evidence that the economy is sliding into recession until the change is confirmed by a consistent movement for two or preferably three months.[7] Nevertheless, economic journalists and policy makers in Washington are sometimes misled and may over-react to changes in the preliminary numbers.

8.3.3 *Okun's law, the GDP gap and the costs of unemployment*

What are the costs of unemployment? One consequence is the rise in expenditures on unemployment benefits, which contributes to a rise in the government's budget deficit. Income tax collections and other tax revenue sources drop off when the economy slows down in recession, which is a second factor contributing to budget deficits. While these factors do take their toll on the government's budget, they are *not* the major cost to the

[7] Lawrence R. Summers, "Measuring Unemployment," *Brookings Papers on Economic Activity*, 1981.

economy of a rise in unemployment. When more workers are unemployed the nation's output drops and the economy's productive potential is not realized. In recession the Gross Domestic Product (GDP) falls far below the potential output that the nation is capable of producing. The loss in output is a major cost of rising unemployment. More than this, higher unemployment contributes to social distress, increased poverty, and rising crime rates.

While serving on the staff of President Kennedy's Council of Economic Advisers, Arthur Okun [1929–1979] suggested a simple rule-of-thumb procedure, now known as Okun's Law, for estimating the *GDP gap*, which is the loss in output arising from excessive unemployment[8] According to Okun's Law, the GDP gap is:

$$Y^{gap} = \left(\frac{Y^p - Y}{Y}\right) = 2.5(U - U^*) \tag{1}$$

Notation:

Y	Symbol denoting GDP
Y^p	Potential GDP
$Y^p - Y$	GDP shortfall
Y^{gap}	$(Y^p - Y)/Y$ or the GDP gap
U	Unemployment rate
U^*	Benchmark Unemployment Rate
$U - U^*$	Excess unemployment

To illustrate, the 1980 *Economic Report of the President* argued that 5.1% was a reasonable estimate of U^*. Using this value of U^* we calculate for December, 1992, when the unemployment rate was 10.2%:

$$Y^{gap} = 2.5(10.2\% - 5.1\%) = 12.7\%\,. \tag{2}$$

That is to say, the economy's output is estimated to have been 12.7% below its rated capacity. In contrast, for June, 1991, when the unemployment rate

[8] "Potential GNP: Its Measurement and Significance," *Proceedings*, American Statistical Association, 1962. The late Arthur Okun [1929–1979] of Yale University developed the relationship bearing his name while serving on the staff of President Kennedy's Council of Economic Advisers. Later he chaired the Council under Lyndon Johnson; subsequently, he was a Senior Research Economist at the Brookings Institution in Washington, D.C.

was 5.1%, right on the benchmark rate, the equation implies that the Y^{gap} was zero.

Despite its simplicity, *Okun's Law* has survived much better than anyone had a right to anticipate when it was first enunciated in the 1960s, although a factor of 2 rather than 2.5 is now thought by many to provide a better link between the gap and excess unemployment. Controversy about Okun's Law has centered primarily on the appropriate value for the benchmark unemployment rate. When Arthur Okun developed his relationship, the Kennedy Council of Economic Advisers had announced a "4% interim unemployment target," which he used as the benchmark unemployment rate U^* in equation (2). It is obvious from the equation that the smaller the benchmark unemployment rate the larger the resulting estimated loss from unemployment.

Critics of the Kennedy Council objected at the time that any attempt to push the unemployment rate down to such a low level would generate substantial inflation. They pointed out that a spell of unemployment may be required to allow for the natural movement of workers between jobs. A laid off worker is all too likely to experience a period of unemployment before finding suitable employment. And a high unemployment rate helps combat inflation by making it more difficult for workers to push for excessive wage increases.

Reflecting the conflict of professional opinion, a wide range of different names has been given to the critical benchmark unemployment rate in Okun's equation. At various times it has been referred to as the "Natural Unemployment Rate", the "Full Employment Unemployment Rate," the "Non-inflationary Rate of Unemployment" (NIRU), and the "Non-Accelerating Inflation Rate of Unemployment" (NAIRU).[9] Whatever the name, it is fair to say that the vast majority of macro economists were totally surprised to find that in year 2000 the American economy was able to enjoy 4% unemployment without experiencing substantial inflation as a consequence.

[9]Perhaps the most optimistic assertion concerning the magnitude of this target rate was the statement of William Vickrey in his Presidential Address to the American Economic Association at the 1993 Annual Meetings in Anaheim, California. Vickrey argued that once proper tax reform and appropriate direct controls on inflation were imposed, the unemployment target could be held to 1.5%. It is fair to say that almost all economists oppose both direct controls and any attempt to target such a low unemployment rate.

8.4 Inflation arithmetic

Figure 1.3, back in Chapter 1, showed that the dollar bill is a rubber yardstick providing an imprecise measure for comparing values at different points in time. The price index increased from 18.0 in 1945 to 177.1 in year 2000 (1982–84 = 100), implying that one dollar at the end of World War II would buy almost as much as ten dollars in 2000 — prices increased by a factor of ten in the last half of the 20th century.[10]

Most but not all major inflations in the United States have taken place in times of war. The inflation of the 1970s into the 1980s was the most severe peacetime inflation in U.S. history. This inflation was precipitated when the Organization of Petroleum Exporting Countries (OPEC) achieved massive increases in the price of oil, starting with a jump in the price of a barrel of oil from \$3.50 to \$13.00 per barrel in 1973, at the time of the Arab-Israeli Yom Kippur War. And in 1979, at the time of the Iranian revolution, OPEC pushed the price of oil up to \$35 per barrel.

Until the mid-20th century, Figure 1.3 shows, periods of inflation were more or less balanced off by deflations — periods of falling prices — leaving little net change in the overall price level. Comparison of the evidence on price movements plotted on Figure 8.6 with the unemployment data that was plotted on Figure 1.2 suggests that a period of falling prices, far from being a happy time, is likely to be associated with severe recession or outright depression.

Relative to much of the world, inflation has been rather mild in the United States. A study based on a sample of 133 countries from 1960 to 1996 revealed that 1/5th of the countries had experienced episodes of inflation at rates above 100% per annum.[11] Table 8.6 summarizes the inflationary experience of a number of countries. The next section of this chapter will explain how price indices are constructed and how they are to be interpreted.

---

[10] Later in this section we will explain in detail how to interpret price indices. Then it will be clear that if the price index had increased from 18 to 180, \$1 in 1945 would have purchased as much as \$10 in 2001 because \$1/18 = \$10/180 The expression "(1982–84 = 100)", while obviously making no sense mathematically, specifies the units of measurement for the price index. As will be explained more clearly a bit later on, this price index has been scaled so as to average 100 over the years 1982 to 1984.

[11] Stanley Fischer, Ratna Sahay and Carlos A. Végh, "Modern Hyper- and High Inflations," *Journal of Economic Literature*, September, 2002, pp. 837–880.

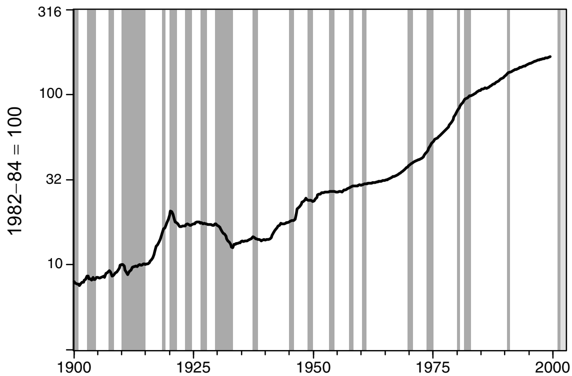


Fig. 8.6. Inflation in the 20th century
Consumer price index, semi-log (ratio) scale.

8.4.1 *Inflation compounds like money in the bank*

How rapidly inflation can erode the value of money is not always fully appreciated. The rules of compound interest apply to inflation as well as to money in your savings account:

> If you leave \$100 in the bank earning 10% interest, you will have the \$100 principal + \$10 interest = \$110 at the end of the first year; leave the \$110 in the bank and you will have \$110 + \$11 interest = \$121 at the end of the second year; after three years you will have \$121 + \$12.10 = \$133.10. After 10 years you will have $\$100 \times 1.10^{10} = \$259.37$ in the bank, thanks to the powers of compound interest!

Recall that the equation for compound interest is

$$A_t = P_0(1+i)^t\,, \tag{3}$$

where P_0 denotes the principal sum we deposited in the bank in year 2000 for t years at rate of interest i. By analogy with money growing in the bank at rate i, when the rate of inflation is constant at rate $\dot{p}$, the price level grows in accordance with the following equation, where p_t is now interpreted as the level of the price index in year t:

$$p_t = p_0(1+\dot{p})^t\,. \tag{4}$$

In general, the rate of inflation is $\dot{p}_t = (p_t - p_{t-1})/p_{t-1}$, but we write $\dot{p}$ without the t subscript when we are considering a constant rate of inflation.

Table 8.6. Inflations around the world.

Price index (1995 = 100)

Country	1971	1980	1990	2000
United States	26.6	54.1	94.1	113.0
Argentina			24.7	99.5
Germany	42.5	66.2	85.7	107.0
Hungary	7.9	11.7	32.4	201.4
Israel		0.2	54.7	136.0
Korea	10.2	40.3	74.1	121.5
Italy	9.0	31.2	78.3	112.8
Japan	34.3	76.3	93.4	101.5
Mexico	0.1	0.3	44.5	239.9

Rate of inflation (percent per annum)

	1971–1980	1980–1990	1990–2000	1971–2000
United States	7.4	5.7	1.8	3.8
Argentina			15.0	
Germany	5.0	2.6	2.2	2.4
Hungary	5.0	10.7	20.0	8.9
Israel		75.3	9.5	
Korea	16.5	6.3	5.1	6.6
Italy	14.8	9.6	3.7	6.7
Japan	9.3	2.0	0.8	2.8
Mexico	13.0	64.9	18.3	22.1

Source: *International Financial Statistics Yearbook*.

Table 8.7. United States wartime inflations.

(1982–84 = 100)

War	Start of War Year	Start of War CPI	End of War Year	End of War CPI	Percentage Increase	Annual Rate of Inflation
Civil War (North)	1861	9.0	1864	16.0	77.8%	21.1%
World War I	1915	10.1	1920	20.0	98.0%	14.6%
World War II	1939	13.9	1945	18.0	29.5%	4.4%
Korean War	1950	24.1	1954	26.9	11.6%	2.8%
Vietnam War	1965	31.5	1970	38.8	23.2%	4.3%

Table 8.7 reports on the rate of inflation in five periods of military conflict. The table reports the level of the price index at the beginning and at the end of each war, the percentage increase in prices during the war, and the annual rate of inflation. Some care is required in calculating the severity of inflation. Note that the *number* of points by which the price

index increased in the North during the Civil War, $16 - 9 = 7$, is a bogus measure of the extent of inflation. It makes the Vietnam War increase from 31.5 to 38.8 appear to be of the same magnitude as Civil War increase from 9 to 16. Instead, economists usually report the *percentage* increase in prices. The percentage increase in prices during the Civil War was $(16 - 9)/9 = 7/9 = 77.8\%$.[12] The Vietnam increase of only $38.8/31.5 - 1 = 23.2\%$ is small when compared to the Civil War inflation.[13]

The *average annual rate of inflation* in the North during the Civil War [the $\dot{p}$ in equation (4)] can be found by solving the compound interest rate equation for $\dot{p}$:

$$16 = (1 + \dot{p})^3 9\,. \tag{5}$$

This is analogous to the following interest rate problem:

Suppose you had placed \$9 in the bank at the beginning of the Civil War and had left the funds on deposit in the bank accumulating interest. At the end of the war your bank informs you that you have accumulated \$16. What rate of interest did you earn on you deposit? That is to say:

> If you had earned interest at annual rate $\dot{p}$ on \$9 in the bank, at the end of the War you would have had \$16 in nominal terms, or precisely the same real amount in terms of actual purchasing power.

There are a number of strategies for calculating $\dot{p}$:

#1: One way to solve for $\dot{p}$ is to take logs to the base e of equation (5):

$$\ln 16 = 3\ln(1 + \dot{p}) + \ln 9\,.$$

[12] Inflation in the Confederacy was much worse than in the North, prices on average rising about 20% per month or an annual rate of $(1 + 0.2)^{12} = 792\%$.

[13] A problem with this procedure for calculating rates of change is that it is *asymmetric*. For example, if prices rise from 9 to 16 followed by a fall back to 9 in the following year, the increase is reported as 77.8% while the fall in price is reported as -43.8%; and the average rate of inflation for the two years is $1/2\ (-43.8\% + 77.8\%) = 17\%$ even though the price level at the end of the two year period has returned to its initial level! There are two procedures for avoiding this absurdity: Sometimes the average of the two years is used in the denominator when computing percentage changes. For our example this yields a rate of inflation of $(16 - 9)/[(16 + 9)/2] = 56\%$. Since $(dp/dt)/p = d\ln(p)/dt \approx \ln(p\ ) - \ln(p\ _{-1})$, some recommend that one report the difference of the logs to the base e as the rate of inflation: $\ln(16) - \ln(9) = 57.5\%$. An advantage of these two procedures is that they are symmetric, showing the same rate of change whether prices are rising or falling. Except when the changes are large, however, it is customary to just divide by the initial value.

Therefore,

$$\ln(1+\dot{p}) = (\ln 16 - \ln 9)/3 = (2.77 - 2.20)/3 = .19179$$

Taking the antilogarithm with the calculator's e^x key yields $1+\dot{p} = e^{.19179} = 1.211$; more precisely, $1+\dot{p} = \exp[\ln(1+\dot{p})] = \exp(0.19179) = 1.211$. Thus $\dot{p} = 21.1\%$.

Notational convention: $\exp(x) = e^x$. Also, by definition, $e^{\ln x} = \exp(\ln x) = x$.

#2: With slightly less effort you can find the answer on any $10 calculator with a Y^x key: if from equation (3), $(1+\dot{p})^3 = 16/9$, then $(1+\dot{p}) = (16/9)^{1/3} = 1.2114$ or $\dot{p} = 21.14\%$.

#3: Alternatively, you can solve for the unknown $\dot{p}$ directly on many "Business Analyst" pocket calculators.

Historical Note

During most of World War II a complex system of price controls was invoked in an attempt to hold inflation in check. The price index had stood at only 13.9 (1967 = 100) in 1939, increasing during the war to 19.5 by 1946. When price controls were removed, the price level jumped to 24.1 by 1948. Many economists argue that the price controls had just postponed the inflation.

When inflation is extremely rapid, it is convenient to measure inflation at monthly rather than annual rates. *Hyperinflation* is a period in which prices rise by 50% or more per month (i.e., $(1+.5)^{12} - 1 = 12{,}875\%$ per annum) for several months. In one of Hungary's two post-World War II hyperinflations, prices shot up at a peak monthly rate of 41.9×10^{15}. During the last quarter of the 20th century there were 15 hyperinflations. From May 1989 to May 1990, prices in Argentina climbed at a 12% per month clip. This was severe but not hyperinflation.

8.4.2 *Plotting inflation*

There are several procedures for graphing inflation data. One may try to interpret appropriately data plotted on a standard time series chart, such as on Figure 1.3. Such graphs can be misleading. For example, Figure 1.3 erroneously made the Viet Nam war inflation appear to be as serious as the Civil War inflation. The optical distortion arises because the graph makes changes in the price level proportional to the change in the number of points of the index rather than the percentage change. One way to avoid such optical distortions is to plot the rate of change in the consumer price index from one period to the next. Thus Figure 8.7 replots the 20th century inflation data from Figure 8.6, but as percentage changes — each point shows the rate of change in the price index from the preceding year, $(p_t - p_{t-1})/p_{t-1}$. This new graph makes it clear that the unprecedented overall climb in the price level in the 2nd half of the 20th century resulted from the absence of periods of deflation rather than from more severe inflationary episodes. But a disadvantage to plotting percentage changes is that it makes it hard to disentangle how much of a change in the overall price level all those erratic fluctuations have added up to over the years.

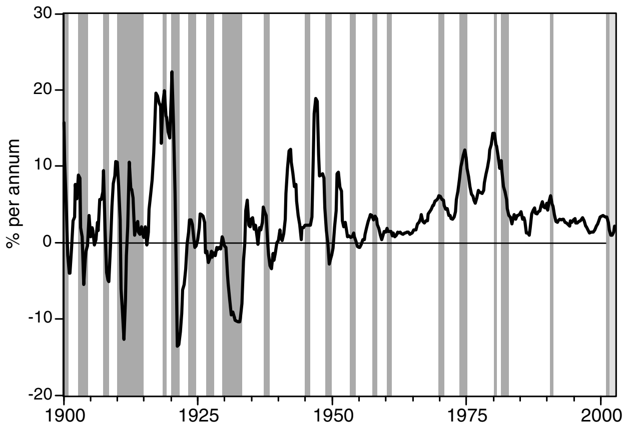


Fig. 8.7. Price changes in the 20th century
Annual rates of change in the Consumer Price Index.

An alternative to plotting the rate of change in the price level is to use a *semi-logarithmic* or "ratio" chart. Thus Figure 8.8 has the same data as were plotted on Figure 1.3, but the $\log_e p$ rather than p itself is plotted on the ordinate against time (t) on the abscissa. The rationale behind this procedure will be clear if we return to equation (4). This is a nonlinear relationship, which is difficult to plot on a graph. But if we take logs of both sides we have:

$$\ln p_t = \ln p_0 + t[\ln(1 + \dot{p})]\,. \qquad (6)$$

Thus the law of compound interest yields a straight line when $\ln p_t$ rather than p_t is plotted on the ordinate. The intercept is $\ln p_0$ and the slope is

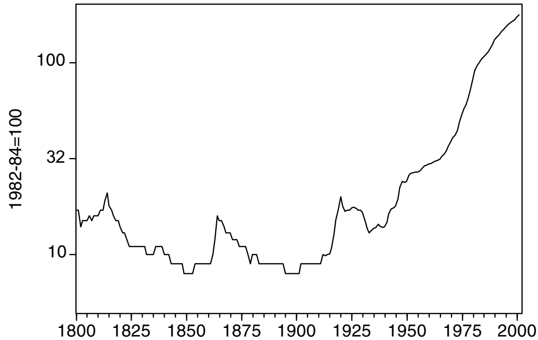


Fig. 8.8. Two centuries of inflation, semi-log scale.

$\ln(1+\dot{p}) \cong \dot{p}$.[14] So the slope of the curve plotting the $\ln p$ against time is the rate of inflation.

A semi-log graph has two advantages:

1. One can judge the level of prices from this graph, which one cannot do with the percentage change plot.
2. The slope of the curve plotted on a semi-log graph accurately reveals the rate at which prices are changing, which is difficult to see when p_t is plotted on the ordinate.

The primary shortcoming of the semi-log graph is that it is difficult to explain to those who are not familiar with logarithms. To circumvent this problem, economists testifying before Congressional Committees and writing for the *Economic Report of the President* are careful to call a semi-log chart a *ratio chart.*

8.4.3 *Manipulating price indices*

Changing the base year

Consider the gasoline price information reported on the top row of Table 8.8. This row reports an index for the price of gasoline normalized with

[14]The approximation follows from neglecting all but the first term of the Taylor's series expansion of $\ln(1+\dot{p}) = \dot{p} - \dot{p}^2/2 + \dot{p}^3/3 - \dot{p}^4/4\ldots$, which converges if and only if $-1 < \dot{p} < 1$. The approximation is most accurate for small inflation rates; for example $\ln(1+2\%) = 0.01980262729618$, which is almost exactly 2%; $\ln(1+10\%) = 0.09531017980432$, which reasonably approximates 10%, and $\ln(1+50\%) = 0.4054651081082$, which is not as precise.

Table 8.8. OPEC, the price of gasoline, and the CPI.

	1960	1972	1980	1990	2000
1982–84 = 100					
1 Gasoline	24.4	28.4	97.4	100.9	128.5
2 CPI ∼ all items	29.6	41.8	82.4	130.7	172.2
3 real price of gasoline	82.5	67.9	118.2	77.2	74.6
2000 = 100					
4 gasoline	19.0	22.1	75.8	78.5	100.0
5 CPI ∼ all items	17.2	24.3	47.9	75.9	100.0
6 real price of gasoline	110.6	91.0	158.4	103.5	100.0

Percent increase	1960–2000	1972–1980	1972–2000
7 all items	481.95%	97.1%	311.96%
8 gasoline	426.33%	243.0%	352.46%
9 real price of gasoline	−9.56%	74.0%	9.83%

Note: The "real price of gasoline" is the price index for gasoline divided by the CPI (all items).
Source: *Economic Report of the President*; Bureau of Labor Statistics.

"1982–84 = 100." That expression, ridiculous from a mathematical viewpoint, simply means that the index has been scaled so that it averaged out to 100 over the years 1982 through 1984. This was achieved by expressing the price of a gallon of gasoline in each year as a percentage of the average price of a gallon in those three years.

The use of the 1982–84 average as the base for purposes of reporting the movement of gasoline prices may be arbitrary, but it has become rooted in custom, just as has the decision to measure temperature in Celsius or in Fahrenheit. Row 4 of Table 8.8 reports the gas price index re-scaled so that it equals 100 in year 2000. This re-scaling is no more complicated than converting a measurement reported in feet into yards by dividing by 3. To make 2000 the new base year (2000 = 100) it was only necessary to divide all the entries in row 1 by the level of the index in 2000. Taking 1960 as an example, we divide 24.4, the 1960 entry from row 1, by 128.5, the entry for 2000, or 24.4/128.5 = 19.0%. This is the first entry in row 4, but the percent sign is omitted, as is customary when working with index numbers.

Real prices

The Consumer Price Index (CPI) measure of the overall level of prices, also with the base year 1982–84, is reported on row 2 of Table 8.8. Comparing row 1 with row 2 reveals that after OPEC was formed the price of gasoline soared dramatically above the overall price level. But in the 1980s the CPI eventually caught up and surpassed the price of gasoline. Row 3 of the table reports the "real price of gasoline." This is the gasoline price index divided by the CPI. It shows how the price of gasoline changed relative to the average price of all commodities during the great OPEC inflation and its aftermath.

The Rule of 70

If prices are rising by 10% per annum it will not take a full ten years for prices to double: by the end of the first year of 10% inflation prices will have climbed from 100 to 110; but by the end of the second year of 10% inflation they will have reached $(1 + 10\%)^2 = 121\%$. In only about 7 years of 10% inflation prices will have approximately doubled, $(1 + 10\%)^7 = 194.9\%$.

The Rule of 70 is a rough approximation that says: *Dividing 70 by the rate of inflation yields the number of years required for prices to double.*

Examples:

- With 10% inflation prices will double in about $70/10 = 7$ years. Similarly, money in the bank earning 10% interest will double in value in approximately 7 years.
- With 5% inflation it will take 14 years for prices to double but with 20% inflation it would take $70/20 = 3.5$ years.

To see why the rule of 70 works, note that if you placed P dollars (the principal) in the bank at rate of interest i for t periods you will have the amount

$$A = P(1 + i)^t \,. \tag{7}$$

If t is to be long enough for the funds in the bank to double we must have $A = 2P$, which means that

$$\frac{A}{P} = 2 = (1 + i)^t \,. \tag{8}$$

Taking logs of this equation yields $\ln(2) = t \ln(1 + i)$. Hence,

$$\frac{\ln 2}{\ln(1 + i)} = t \,. \tag{9}$$

Now as a first approximation $\ln(1 + i) = i$; also, $\ln 2 = 0.69315 \approx 0.7$ and we have $70\%/i = t$.

8.4.4 *Measuring price movements*

In the early 1970s, when the rampant inflation sparked by OPEC oil price hikes dominated the news, television newscasters were confronted with the difficult task of explaining to their audience how inflation is measured. They illustrated how the general level of prices was changing from month to month by showing a housewife purchasing a typical shopping cart loaded with groceries at the checkout counter. The newscasters demonstrated the magnitude of inflation by comparing the cost this month of the goods in the shopping cart with what those same goods had cost a month earlier.

A similar but much more elaborate procedure is used by the BLS in calculating the level of the *Consumer Price Index* (CPI). Each month the BLS prices out a much more comprehensive set of commodities, including many items that are not sold at the supermarket but which are important items in the representative household's budget, such as auto repairs and gasoline, medical and dental bills, electricity and rents, telephone and TV cable, etc. The BLS collects prices for a carefully selected sample of about 80,000 purchases. The CPI estimate of inflation is watched by the more than 2 million workers who are covered by collective bargaining arrangements tying their wage increases to changes in the CPI. It is also of special interest to senior citizens, for their Social Security benefits are adjusted for inflation with the CPI.

Simplified example

The grossly simplified numerical example on Table 8.9 shows the basic principles involved in constructing a price index. To keep things simple, our hypothetical consumer purchased only two commodities. The prices of both commodities went up over the decade, but not by equal amounts. The strategy is to compute an average increase, but with more attention paid to the commodities on which the typical consumer spends the most — or as statisticians say, more *weight* is given to commodities which we purchase in greater number. The table reveals that it would have cost \$700 in year 2000 to buy the quantities of the two commodities that the consumer actually purchased in 1990. By this reckoning, prices have increased by a factor of \$700/\$300 = 233.3%. We should say: "in 2000 the price index was at 233.3 (1990 = 100)." Note that we write 1990=100 to indicate that the prices are being reported as a percent of prices prevailing in 1990, which was chosen as the base year. Also, we followed the custom of reporting the index in percentage terms, but without the percent sign. We refer to our

Table 8.9. Constructing price indices — a simplified example.

	1990 data			2000 data	
	quantity consumed	prices	expenditure	prices in 2000	cost of 1990 goods at 2000 prices
Hamburger	100	\$1.00	\$100	\$3.00	\$300
Movie ticket	50	\$4.00	\$200	\$8.00	\$400
Total expenditure			\$300		\$700

The Laspeyres consumer price index, 1990 = 100, is \$700/\$300 = 233.3.
(It is customary to report price indices in percentage terms but without the percentage sign.)

index as the Laspeyres Price Index after Etienne Laspeyres [1834–1913], the statistician who developed the procedure.

Observation

In comparing the temperature of two cities, it matters little whether you consistently use the Fahrenheit or the Celsius scale as the unit of measure; similarly, the choice of 1982–84, 1990 or 2000 for the base year in measuring the overall price level amounts only to the choice of the unit of measure.

We must try to state in a more precise way the principles guiding the calculations on Table 8.9. We used quantity weights from 1990 for our hypothetical example. That is to say, the significance attached to each commodity depended on how many units of that commodity were purchased in 1990. To construct the Laspeyres price index p^L we divided the \$700 that it would have cost to buy the 1990 purchases in year 2000 by the actual cost of \$300 that was spent on those goodies in year 1990:

$$p^L = \$700/\$300 = 233\% = 233(1990 = 100)\,.$$

We can sum up our procedure by saying we have calculated the following concept:

$$\text{Laspeyres Price Index} = \frac{\text{quantities at earliest time} \times \text{latest prices}}{\text{quantities at earliest time} \times \text{earliest prices}}\,.$$

In terms of the summation notation used by statisticians, with q_i^E denoting the quantity and p_i^E the price of good i from the earliest time period, and p_i^L its price in the latest time period:

$$p^L = \frac{\sum_{i=1}^{n} q_i^E p_i^L}{\sum_{i=1}^{n} q_i^E p_i^E}. \tag{10}$$

An alternative index, the Paasche index developed by Hermann Paasche [1851–1925], is constructed by using the current year quantity weights, q_i^L, instead of the initial year quantities, q_i^E. The calculations are shown on Table 8.10. Back in 1990 the commodities actually purchased in 2000 would have cost $80 \times \$1 + 100 \times \$4.00 = \$480$. Since the set of commodities that cost \$1,040 in 2000 would have cost only \$480 in 1980 prices, the Paasche price index stands at

$$p^p = \$1040/\$480 = 2.167 = 216.7(1990 = 100)\,.$$

To summarize the procedure, we write

$$\text{Paasche Price Index} = \frac{\text{quantities at latest time} \times \text{latest prices}}{\text{quantities at latest time} \times \text{earliest prices}}.$$

Or in terms of summation notation,

$$p^P = \frac{\sum_{i=1}^{n} q_i^L p_i^L}{\sum_{i=1}^{n} q_i^L p_i^E}. \tag{11}$$

Should we use the Laspeyres Price index (beginning quantity weights) or the Paasche Index (most recent quantity weights)? As a practical matter, the Laspeyres procedure is most commonly employed because it generates an estimate that can be published in a timely fashion in that it does not require information about quantities consumed at the end of the period. It is difficult to produce the Paasche index in a timely fashion because only fragmentary information is available on quantities consumed in the most recent period. The Laspeyres index makes do with base year quantity weights which can be measured with greater precision but may be outdated

Table 8.10. Constructing Price Indices — Laspeyres versus Paasche.

	1990 data quantity consumed	price	expenditure		
Hamburger	100	$1.00	$100		
Movie ticket	50	$4.00	$200		
Total expenditure			$300		
	2000 data quantity consumed	**price**	**expenditure**	**cost of 1990 goods at 2000 prices**	**cost of 2000 goods at 1990 prices**
Hamburger	80	$3.00	$240	$300	$80
Movie ticket	100	$8.00	$800	$400	$400
Total expenditure			$1,040	$700	$480

Notes: The Laspeyres consumer price index, 1990 = 100, is the cost of buying the initial bundle of goods at current prices divided by the cost of buying that same bundle of goods at initial year prices: $p \quad = \$700/300 = 233.3$ (beginning period quantity weights).

The Paasche consumer price index, 1990 = 100, is current period expenditure (year 2000) divided by cost of buying that same set of goods in the base year (1990): $p \quad = \$1,040/\$480 = 216.7$.

Marshall Index (Arithmetic Average): $p \quad = (233.3 + 216.7)/2 = 225.0$

Fisher Index (Geometric Average): $p \quad = (233.3^{*}216.7)^{1\ \ 2} = 224.8$

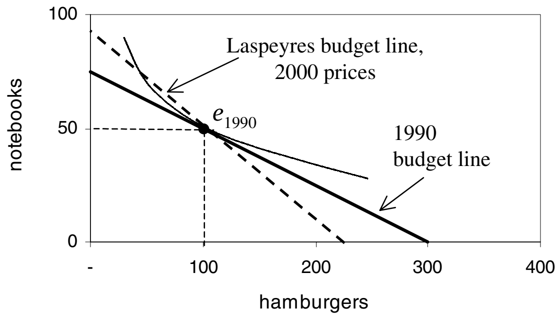


Fig. 8.9. Laspeyres price index exaggerates inflation
In 1990 our consumer purchased at point e; therefore, our consumer's indifference curve must be tangent to the 1990 budget line at this point.

The Laspeyres budget line goes through e_{1990}, allowing our consumer \$700 to buy the 1990 purchases at year 2000 prices. But the Laspeyres budget line has a different slope than the 1990 budget line, because relative prices are different. Therefore, the Laspeyres budget line is not tangent to the indifference curve at point e_{1990}. The Laspeyres budget line would enable our consumer to do better than at e_{1990} by substituting away from hamburgers, which have gone up most in price. Thus the Laspeyres price index overstates the increase in money income required to give our consumer the same utility as enjoyed in 1990.

and not reflect the current purchasing interests of consumers. But as we shall see, there are serious problems with both of these indices.

Substitution bias

Using the initial price weights, as with the Laspeyres price index, tends to overstate changes in the cost of living. To see why, consider Figure 8.9. Our consumer spent \$300 to maximize utility in 1990 by purchasing at tangency point e_{1990} on the 1990 budget line. In calculating the price index $P_L = \$700/\$300 = 233.3$, Laspeyres is implicitly allowing our consumer an increase in income to \$700 in order to enjoy the same utility by purchasing the same bundle of goods at the new prices. But if relative prices have changed, as in this example, the new hypothetical Laspeyres budget line can not be tangent to the indifference curve at point e. This means that the consumer could enjoy more utility than in 1990 with \$700 in 2000 by substituting away from the good that has had the biggest percentage increase in price, consuming more notebooks and fewer hamburgers by moving along the \$700 year 2000 budget line to a new tangency point. That is why the Laspeyres index *overstates* the increase in income required to keep pace with inflation. The distortion in the price index generated by the tendency of consumers to substitute away from goods that increase most in price is called substitution bias.[15]

Unfortunately, the Paasche price index is also subject to substitution bias but in the opposite direction. The Paasche index tends to understate changes in the cost of living because the consumer could have purchased for less at the earlier date a bundle of goods giving as much satisfaction as the goods purchased in the final year. Or to put it another way, while \$480 would have enabled the consumer to purchase in 1990 the same bundle of goods that was actually purchased in 2000, our utility maximizing consumer would have done better than this with the \$480 at the prices prevailing in 1990. Why? Because the indifference curve going through that bundle of goods was tangent to the year 2000 budget line rather than one reflecting prices in 1990 (See exercise 8.4). Less than \$480 would have enabled the consumer to buy a bundle of goods at 1990 prices yielding the same level of satisfaction as was enjoyed in year 2000. This means that the Paasche index, the ratio \$1040/\$480, tends to *understate* the amount of inflation — the denominator is too large.

Since the Laspeyres index tends to overstate the effect of rising prices on the cost of living of a utility maximizing consumer while the Paasche tends to understate the inflation, we can certainly say that the true increase is between the two estimates. English economist Alfred Marshall [1852–1924] suggested taking the arithmetic average of the two indices, which is an obvious compromise, $p^M = (p^L + p^P)/2$. Yale Professor Irving Fisher [1867–1947] used an elaborate argument to explain why the geometric average of the Laspeyres and Paasche indices provides an "Ideal" Index Number:[16] $p^F = (p^L \times p^p)^{1/2}$.

New products and product improvement

The introduction of new products and quality change may be a blessing to the consumer, but they create serious problems for the economic

---

[15]See Chapter 4.3.3: Compensating for inflation, for an indifference curve example illustrating why the use of the initial period weights tends to overstate inflation. If only we could estimate it, the expenditure function, $M(U, p\ \ , p\ \ )$, equation 39 of Chapter 4.3.3 would provide a precise way of determining the income required to achieve in year 2000 the same level of utility as enjoyed in year 1990. We would then calculate the true increase in the cost of living as $M(U, p^{2000}, p^{2000})$ divided by actual expenditures in 1990.

[16]It can be shown that the geometric average can never be greater than the arithmetic average, which means that Fisher's Ideal Index Number can never show more inflation than Marshall's arithmetic average.

statistician trying to measure changes in the general price level. Consider the ubiquitous pocket calculator: When first brought to market in the early 1970s, a model graced with a square-root key, considered advanced at the time, sold for as much as $75; today a much more powerful calculator can be purchased for four or five dollars but is regarded as inadequate for teaching high school math because it does not have a plotting function. Clearly the price of calculators has fallen dramatically, but it is not easy to determine how much of an additional allowance should be made for the fact that today's product is so much superior to that which was available when this product first came on the market.

Two examples will indicate why it is so difficult to make appropriate adjustments for quality change:

1. In the 1950s Yale Professor Richard Ruggles used to ask his economics majors to compare, as a mental exercise, the goods currently available in the marketplace with those advertised 20 years earlier in an old Sears Roebuck mail-order catalog.[17] If he were to give them $200 to spend, would they prefer to spend it on merchandise in the 20 year old catalog, buying goods of 20 years ago at the prices of 20 years ago, or would they prefer to spend the $200 buying goods currently available in the market place at today's prices. Students almost universally said they would prefer to buy today's goods at today's prices. Ruggles then pointed out that their answer implied that the increase in prices had been more than matched by product improvements. Or to put it another way, there had been a decline in prices, deflation rather than inflation, once due allowance is made for quality improvement.
2. In the 1990s it became apparent to a number of academic economists and the government statisticians that the reduction in the cost of computation brought about by the computer revolution had not been properly taken into account in calculating overall price increases. As a result, inflation had been exaggerated and hence the economy's real output (measured by dividing the market value of output by the price index) had been seriously understated.

[17] Sears Roebuck used to market its products primarily through the mail, pioneering catalog sales. Now it has given up catalog sales and shortened its name to Sears. One could perform Ruggles experiment to day using the catalog of L. L. Bean or any other well known mail-order firm.

Chain linked indexes

The statisticians make every effort to cope with the difficulties created by new products, product improvement and substitution bias.[18] One way is to *chain link* a sequence of one year estimates. To illustrate, suppose we wish to construct a chain link index from 1990 to 2000 with 1990 as the base year. The first step is to construct a sequence of one year Fisher geometric average indices. We calculate $P^F_{1991} = (P^L_{1991}P^P_{1991})^{1/2}$ with 1990 as the base year and the average of the quantities of the commodities consumed in 1990 and 1991. In general, $P^F_t = (P^L_t P^P_t)^{1/2}$ with year $t-1$ as the base year and using the quantities of the relevant commodities for year $t-1$ and t. Then we construct the chain linked index as follows: $P^c_{1990} = 100$ because 1990 is the base year. $P^C_{1991} = P^F_{1991}P^C_{1990}$ because P^F_{1991} is our measure of prices in year 1991 relative to prices in 1990. In general, $P^C_t = P^F_t P^C_{t-1}$. For example, if the first Fisher index had increased by 4% from 1990 to 1991 and the second by 6% from 1991 to 1992, we would have $P^C_{2002} = 1.04 \times 1.06 = 110.24$. The primary advantage of the chain linked index is that it controls for substitution bias and allows for the gradual introduction of new products because it uses the quantities for each pair of adjacent years. The obvious disadvantage is that it is expensive to estimate in a timely fashion because it requires much more information about quantities.

CPI-U and C-CPI-U

The Bureau of Labor Statistics has traditionally reported a Laspeyres index known as CPI-U. Its weights, which had been based on a detailed 1993–1995 survey of the typical urban resident's spending patterns, were not updated until January of 2002 when data on spending patterns for 1999–2000 became available. In 2002 the BLS began publishing alongside the CPI-U a new Chained Consumer Price Index for all urban consumers, C-CPI-U.[19] That substitution and new product bias were far from negligible problems with the Laspeyres index is suggested by the fact that the chained linked,

---

[18]The distinction between product improvement and new products is not always clear cut. It might be asserted that the electronic calculator was not really a new product! Rather, it was only an improvement on an existing product because there were slide rules and noisy mechanical calculators (some powered by electric motors) long before the pocket calculator, introduced in the 1970s, relegated these old mechanical calculating devices to the museum.

[19]The procedure used to calculate the C-CPI-U index is similar to that described in the preceding paragraph except that a sophisticated Tornqvist formula is used instead of the Fisher index.

C-CPI-U, was about 0.8% lower than the Laspeyres CPI-U in 2002. The primary disadvantage of the new index is that it is not possible to produce it in a timely fashion because it requires current year quantity information. Frequent revisions of the estimates are necessary. For example, in January 2003 the BLS published the preliminary estimate for that month, but announced that a revision of that estimate would be released a year later and a final estimate for January of 2003 would not be provided until February of 2005. All this makes it difficult for evening news reporters to summarize what is happening to inflation in a short sound bite. And it means that policy officials in Washington must make judgment calls on the basis of preliminary estimates of inflation that may turn out, once the revisions are in, to have been far off the mark.

8.5 How to live with inflation

Here is how one observer described the hectic experience of living in Israel during the early 1980s when prices were rising at a 140% average annual rate and, at its peak, hit a monthly inflation rate of 27.5%.[20]

> "Like all countries in which inflation is rampant, the name of the game is to part with the local currency as quickly as possible. Invest or spend! Don't dawdle. It was similar to the children's game "hot potato." Don't be left with shekels in your hand or bank account."
>
> "Rumors of price rises of 20% or more (sometimes a lot more) on food items would spread among the public and supermarkets would become total chaos, as the public would hoard goods from supermarket shelves. After all, what you don't buy now will cost you at least 20% more tomorrow. The same applied to gasoline and other consumer goods. So, why not buy now?"
>
> "When my wife and I purchased a condominium in Rishon LeZion in 1981, interest rates on mortgages were already at a 39% annual rate, a very high rate one might think. No one could foresee that the annual rate of inflation would soon exceed 800% a year. Fixed rate mortgage payments became laughably easy to make as inflation shot up to 800% a year. That is because salaries more or less kept pace with inflation, in a wage-price inflationary spiral."

The Israeli experience with high inflation was far from unique.

[20] "The Israel Experience," by Michael Rosenbloom, downloaded on July 30, 2002 from http://www.ohav.org/travel/experience/inflation.html.

Our experience in the United States, like that in much of the rest of the world, is that there are both winners and losers in times of inflation. Those with fixed incomes, such as retirees living on fixed (unindexed) pensions, suffer from inflation. Workers lose out to the extent that their wages are not adjusted to keep pace with rising prices. Those in the United States who took out 35-year home mortgages at 6% interest in the 1960s benefited from the unanticipated inflation of the 1970s! Why? Because they were able to pay back their mortgages with dollars of greatly reduced purchasing power. Those who borrowed gained while those who had loaned the funds were losers.

Experience suggests that no matter what country one lives in, one should not count on living a lifetime without experiencing inflation. Here are three major lessons about living with inflation:

1. When inflation strikes, it pays to minimize the amount of currency in your pocket and the deposit in your checking account because your money will buy less tomorrow than it will today.
2. One strategy is to hold foreign currency, and the dollar has often been the currency of choice.
3. Those who happened to borrow at low rates of interest before the inflation was felt or even anticipated are blessed with an unexpected surprise. Homeowners who borrowed before the inflation get to pay off their mortgage with currency of vanishing purchasing power. Those who lent money at the low rates of interest that prevailed before inflation was anticipated are hurt.

It is also clear that how much workers are hurt depends on how rapidly their wages are adjusted in response to inflation. Inflation is disruptive, in part because of the time and effort put in to adapting to it and in part because of the uncertainties that it generates.

8.5.1 *The real wage*

Price indices are not only useful for measuring the rate of inflation. They are used to judge whether money wages have been increased by enough to offset the effects of inflation on the purchasing power of workers' wages. Table 8.11 presents data on the minimum wage. As can be seen from the first column of data on the table, a college student today holding a job paying the minimum wage is earning many more dollars per hour than the student's parents may have been earning a generation back when they were in college.

Table 8.11. The minimum wage versus inflation.

Year	Minimum wage	Consumer Price Index (1982–84 = 100)	Real Minimum Wage ($ of 1982–84 purchasing power)	Real Minimum Wage ($ of year 2000 purchasing power)
1938	0.25	14.1	1.77	3.05
1946	0.40	19.5	2.05	3.53
1950	0.75	24.1	3.11	5.36
1960	1.00	29.6	3.38	5.82
1968	1.60	34.8	4.60	7.92
1980	3.10	82.4	3.76	6.48
1990	3.69	130.7	2.82	4.86
2000	5.15	172.2	2.99	5.15

Source: Bureau of Labor Statistics.

But prices have increased as well, as is indicated by the CPI column. As a result, the money wage that a worker receives (aka the nominal wage) does not provide an accurate measure for making comparisons over time. The problem is that the dollar bill is a "rubber yardstick" providing an imprecise measure for making comparisons of market values over time.

To net out the effects of price changes, an estimate of the *real wage* is calculated by dividing the nominal wage by the consumer price index (keeping in mind that the % sign is suppressed in reporting price indices):

$$\text{Real Wage} = \text{Nominal Wage}/\text{Price Index}. \tag{12}$$

By the *nominal wage* we mean the wage that the worker receives *unadjusted* for inflation. For example, in year 2000 the minimum wage was \$5.15 and the CPI (1982–84 = 100) stood at 172.2. Therefore, the real minimum wage in year 2000 was \$5.15/1.722 = \$2.99 (1982–84 prices); for 1968 we have \$1.60/0.348 = \$4.60.

The real minimum wage, reported in the third column of the table, reveals that over long periods of time increases in the minimum wage have not kept up with inflation. In terms of purchasing power, students and others with low wage jobs are being paid less today than those similarly situated a generation ago. The real wage figures may be more meaningful if they are expressed in terms of year 2000 prices, as in the last column, rather than in the archaic 1982–1984 prices corresponding to the base year of the CPI. To shift the base to year 2000 it is only necessary to multiply each figure by 1.72, which is the level of the CPI in that year. For example, measured in terms of year 2000 prices, the real minimum wage in year 1968 was 1.72 × (\$4.60/0.348) = \$7.92, which is much higher than the \$5.15

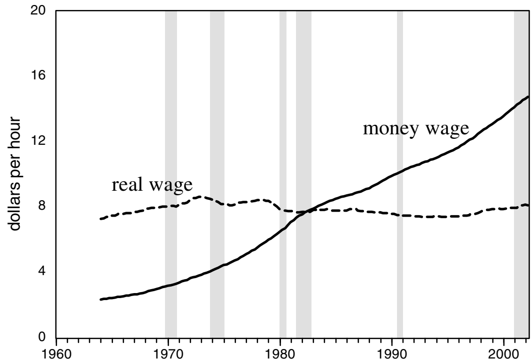


Fig. 8.10. Real wages
Although average money wages (which exclude health and other benefits) have increased substantially over the decades, so has the price level. As a result, the real hourly wage has remained quite stable.

prevailing in 2000. Note though, that the decline in the real minimum wage is exaggerated to the extent that the CPI overstates the extent of inflation because of product improvements and substitution bias.

The concept of the real wage will help us to interpret the evidence about the Great Depression on Table 8.2. Observe that money wages, reported in row 13, fell in the early 1930s in response to the high unemployment rate, declining for workers in manufacturing from a 1929 average of 56¢ per hour to 44¢ in 1933. But prices fell even faster, the consumer price index declining from 17.1 to 13.0 (1982–84 = 100). Thus the real wage increased from 56¢/17.1% = \$3.27 to 44¢/13.0% = \$3.38 in 1933, (1982–1984 prices). This means that those who managed to keep their jobs may have benefited from the depression, *provided* they were lucky enough to put in the same number of hours on the job.

Figure 8.10 shows that money wages increased many times over during the last half of the 20th century. But prices increased about as much, which means that the real wage, adjusted for inflation by dividing by the consumer price index, was quite stable. However, employers found that real hourly worker compensation was up because the price of fringe benefits increased substantially, particularly for workers receiving costly health insurance.

Indexing Wages

In times of rapid inflation, labor unions are likely to insist that labor contracts contain a wage indexing clause specifying that the money wage will be adjusted in response to inflation. Such a clause protects the worker from inflation during the life of the employment contract. And the employer may be willing to acquiesce to the request for an indexing clause because, if inflation does materialize, it will be possible to sell their product for more. Indexing of wages automates the process of adjusting to inflation, but it may also exacerbate the inflation. In the absence of indexing clauses, worker purchasing power might decline with inflation. This might reduce inflationary pressure by leading to a reduction in demand for the nation's output.

8.5.2 *The real rate of interest*

The real interest rate is similar to the real wage, but it is the nominal interest rate that is adjusted for the distortions of inflation.

Definitions

> The *nominal rate of interest* is the annual rate paid on a loan, without adjusting the figures for inflation.[21]
>
> The *real rate of interest* (i^r) is defined as the rate of return paid on a loan *net* of adjustments for inflation.

An example clarifies the distinction:

> Suppose that at the beginning of year 2000 you borrow \$1,000 from your bank at 10% interest for one year. Your contact specifies that you must pay back \$1,100 at the end of the year. If in the meantime prices rise by 5%, the \$1100 you pay back will buy only as much as $\$1{,}100/1.05 = \$1{,}047.61$ would have purchased at the beginning of the year. Looking behind the veil of inflation, you will observe that the loan has cost you in terms of actual purchasing power the same amount as a 4.761% rate loan would have cost in a year of stable prices because $(1 + 4.761\%) \times \$1{,}000 = \$1{,}047.61$.
>
> A borrower who is not deceived by inflation would conclude that the "real rate of interest" is only 4.76%, once proper adjustments are made for the distorting effect of inflation on the contracted interest rate.

[21] Lenders to consumers in the United States are required by the federal Truth in Lending Law to specify clearly in the contract the annual percentage rate of interest that will be charged on the loan.

Similar calculations reveal that a 40% loan in a time of 35% inflation is equivalent to a 3.7% loan in times of stable prices: $1.4/1.35 - 1 = 3.7\%$; i.e., the real interest rate is 3.7%. More precisely, the real rate of interest i^r earned on a loan satisfies the relationship

$$A_t^r = (1 + i^r)A_{t-1}^r \,, \tag{13}$$

where A_{t-1}^r represents the real value of a sum borrowed at date $t - 1$ and A_t^r the real value that sum has grown to at date t.

To find the formula for calculating the real rate of interest, suppose that last year you had placed the sum A_{t-1} in a bank earning interest at nominal rate i. After one year your deposit will equal $A_t = (1 + i)A_{t-1}$. Dividing by the price level p_t we find that in real terms you will have

$$A_t^r = A_t/p_t = (1 + i)A_{t-1}/p_t = (1 + i)A_{t-1}^r p_{t-1}/p_t \,. \tag{14}$$

Since $\dot{p}_t = (p_t - p_{t-1})/p_{t-1}$, $1 + \dot{p}_t = p_t/p_{t-1}$ and on substituting into (14) we have

$$A_t^r = \left(\frac{1+i}{1+\dot{p}}\right) A_{t-1}^r \,. \tag{15}$$

Comparing (15) with (13) reveals

$$i^r = \frac{1+i}{1+\dot{p}} - 1 = \frac{1+i}{1+\dot{p}} - \frac{1+\dot{p}}{1+\dot{p}} = \frac{i-\dot{p}}{1+\dot{p}} \,. \tag{16}$$

Therefore, the real rate of interest is

$$i^r = \frac{i-\dot{p}}{1+\dot{p}} \,. \tag{17}$$

As a close approximation, it is often said that the real rate of interest is the nominal rate of interest minus the rate of inflation; i.e.,

$$i^r \doteq i - \dot{p} \,. \tag{18}$$

(The symbol $\doteq$ indicates approximation.)

The approximation is valid when $\dot{p}$ is small so that the denominator $1 + \dot{p}$ of (17) is close to 1.

The real rate of interest tends to be much more stable than the nominal rate of interest. To see why, consider the demand supply analysis of 8% inflation on Figure 8.11. Point e marks the initial equilibrium in the market for loans in the absence of inflation. If inflation is generally anticipated, consumers will be willing to pay a higher nominal rate of interest to borrow

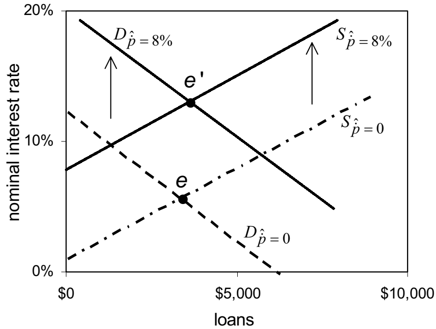


Fig. 8.11. Real versus the nominal interest rate
Anticipated inflation causes both the demand and supply functions for loans to shift upward by the amount of expected inflation. As a result, the nominal rate of interest rises by the expected rate of inflation, but neither the volume of loans nor the real interest rate changes.

the same amount because everyone thinks they know that they will be able to pay back their debt with dollars of deflated purchasing power; thus the demand curve for loans shifts upward. But lenders, also anticipating inflation, will only be willing to loan the same amount if the nominal interest rate is high enough to compensate them for the declining purchasing power they will receive when the loan matures. That is why the supply schedule for loans also shifts upwards by the amount of the anticipated inflation. At the new equilibrium, point e', the nominal rate of interest is up 8%, but the market clearing volume of loans is the same as before. Neither borrowers nor lenders are hurt by inflation when it is generally anticipated, and in the absence of other disturbances the real rate of interest remains stable.

Ex ante versus ex post real interest rate

The distinction between the ex post and ex ante real rate of interest is of fundamental importance:

> The *ex ante* real rate of interest is calculated with the anticipated rate of inflation; it is the relevant variable for explaining the behavior of economic agents, but it is hard to observe.
> The *ex post* real rate of interest is calculated after the fact, when the rate of inflation is known.

The ex ante real rate is relevant in considering how the public responds to anticipated inflation. The ex post rate is relevant in determining, after the fact, who was hurt by the inflation. When inflation is not fully anticipated, nominal interest rates may not rise by enough to offset the inflation and lenders suffer. The opposite situation may arise after inflation has prevailed for a time. When everyone expects inflation but it does not materialize, the borrowers may be hurt if they are stuck paying off loans negotiated at high nominal interest rates.[22]

No one will be eager to make a loan if they anticipate that they will receive back less in purchasing power than they loaned out. But borrowers will be willing to pay a higher nominal interest rate if they anticipate paying back the loan with dollars of reduced purchasing power. That is why high nominal interest rates are the natural response of money markets to anticipated inflation. This linkage between the nominal rate of interest and anticipated inflation is captured by rewriting equation (18) as

$$i = i^r + \hat{\dot{p}}\,. \tag{19}$$

This relationship says that the nominal rate of interest is the sum of the ex ante real rate of interest and the anticipated rate of inflation. Because this is in accordance with the analysis of Irving Fisher [1867–1947], one of the greatest economists of his day, equation (19) is often called the *Fisher equation*. Now it is not unreasonable to assume that i^r tends to be stable, apart from short-term disturbances and errors in anticipating inflation. The real interest rate will be stable if it closely approximates the natural rate of interest. As explained in Chapter 4.4, the natural rate is that rate of interest that balances the availability of profitable investment opportunities with the public's choice between current and future consumption. Given i^r, the Fisher equation says that the nominal rate of interest will rise if the public starts to fear, rightly or wrongly, that inflation will prevail. Thus the mere anticipation of inflation may push up nominal interest rates. The natural response captured by the Fisher equation may be reinforced if the monetary authorities decide to fight the inflation by following a restrictive monetary policy and push for higher interest rates, as will be explained in Chapter 10.

---

[22] If in retrospect a borrower finds that interest rates have gone down, it may be advantageous to "refinance" the loan at a lower interest rate. Some bonds provide a pre-payment clause, which allows the issuing party to pay off the debt early. As explained in footnote 24, when interest rates dropped to unexpected lows in 2000, 2001, and 2002, many homeowners found that they could refinance their outstanding mortgages at drastically reduced nominal interest rates.

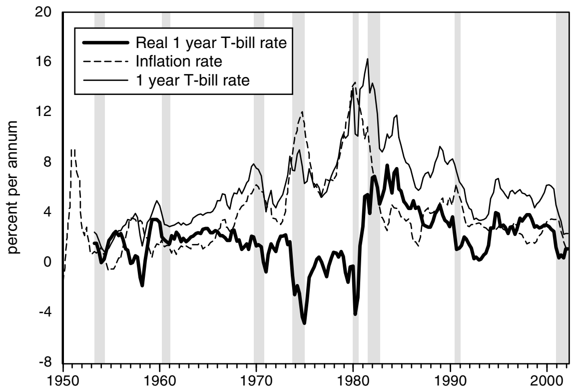


Fig. 8.12. Real rate of interest, ex post
The *ex post* real rate of interest is estimated by subtracting the rate of change in the consumer price index from the nominal one year Treasury Bill Rate.

The *ex post* real rate of interest reported on Figure 8.12 was calculated by subtracting the rate of inflation, as measured by the annual rate of change in the CPI, from the rate of interest paid by the government of the United States on 1 year Treasury bills, which is a major way in which the government borrows to cover the national debt. Because of inflation the real rate is below the nominal rate, but the gap is far from uniform. The fact that this ex post rate of interest is somewhat more stable than the nominal rate suggests that the public was not fooled all that much by unanticipated inflation. But lenders were hurt during the OPEC inflations of the 1970s and 80s because the ex post real interest rate was negative; i.e., lenders received back less in purchasing power than they had lent out because the inflation was not anticipated when the loans were made. And later, when the inflation was being brought under control, the ex post real rate soared because the fall in prices was not fully anticipated.

Tax complications

Taxes complicate real interest rate calculations. Suppose that you loan me \$1,000 for one year at 15% interest. Assuming that I do indeed pay off the loan on schedule, you will receive back the principle plus \$150 of interest income. But if you are in the 30% tax bracket, you will have to pay \$45 in taxes on your interest income, leaving you with \$1,105 after taxes. If there has been 10% inflation, you will have only \$1,105/1.1 = \$1,004.55. Thus, your after tax real rate of interest, i^r_{at} is only 0.455%. More generally, if

t is your marginal tax rate, your after tax nominal rate of interest is

$$i^t = (1 - t)i\,. \tag{20}$$

and your *after-tax real interest rate* is

$$i^{rt} = \frac{(1 - t)i - \dot{p}}{1 + \dot{p}}\,, \tag{21}$$

or as the approximation valid for small $\dot{p}$,

$$i^{rt} \doteq (1 - t)i - \dot{p}\,. \tag{22}$$

To illustrate, observe that in a time of stable prices a lender in the 30% tax bracket would net 3.5% after taxes on a 5% loan. In a period of 7% inflation, a loan at 12% would appear to provide a 5% inflation adjusted interest rate, but it would actually yield only a $3.4\% = (1-0.3)\times 12\% - 5\%$ real return after taxes according to equation (22), or more precisely $3.18\% = [(1-0.3)\times 12\% - 5\%]/(1+7\%)$ by equation (21). A 15.35% nominal interest rate would be required to reward the lender with a 3.5% after tax real rate of return when prices are rising at 7% per annum. The higher interest rate is less painful for the borrower to pay if, as with mortgage payments in the United States, the interest payment can be deducted in computing income for tax purposes.

Tax considerations suggest that Fisher equation (19) should be modified to read

$$i \doteq \frac{(i^r + \hat{\dot{p}})}{(1 - t)}\,, \tag{23}$$

which makes the nominal interest rates more responsive to anticipated inflation than the original formulation, equation (19), implied.

8.5.3 *The mortgage twist*

The market treats some generations less kindly than others. Those who entered the job market after graduating from college in 1933, when the unemployment rate was 25%, obviously faced a type of world very different from that which welcomed those who graduated in 1929, just before the worst stock market crash of all time. Inflation and the accompanying high interest rates have at times disrupted family life. Those planning to buy a new home in 1983 were disappointed to find when they approached their banker for a conventional mortgage that they would be charged an interest

Table 8.12. $200,000 conventional mortgage.

Interest rate	Length (years)	Monthly payment	Qualifying income
7.5%	25	1,478	70,943
7.5%	30	1,398	67,125
5.5%	30	1,136	54,508
10.0%	30	1,755	84,247
12.0%	30	2,057	98,747
15.0%	30	2,529	121,387

rate of at least 15%. This was double the rate that an earlier generation of home purchasers had paid a decade earlier; and as it was to turn out, by 1993 interest rates on new mortgages would have dropped back down to about 7%.

The high nominal interest rates were generated by double-digit inflation (i.e., $\dot{p} \geq 10\%$) and the efforts of the monetary authorities to counter it. The unfortunate couple taking out a 15% mortgage of $200,000 will be told by their banker that they will have to commit themselves to a payment of $2,529 to be made every month for the 30 year life of the mortgage (see Appendix 8.1, equation (26) for the calculation details). As can be seen from Table 8.12, if the 7.5% rate had been available, the monthly payment on that $200,000 mortgage would have been only $1,398. The high interest rate not only meant a high monthly payment; it also meant that it was much more difficult for a potential borrower to qualify for a loan. Bankers often require, as a rule of thumb, that potential borrowers can qualify for a mortgage only if their total housing costs, including taxes and insurance as well as the mortgage payment, will be less than 29% of their before tax income.[23] With the higher interest rates prevailing in much of the 1980s, an annual income of about $121,400 would be required to qualify for the $200,000 mortgage rather than the $67,000 or so that might suffice when interest rates were at 7.5%. Thus the high interest rates meant that many potential homebuyers could no longer afford their dream home. The resulting fall in demand for new houses was hard on the construction and lumber industries as well as on a generation of young homemakers interested in moving out of apartments into their own homes.

[23] The rules of thumb used by mortgage lenders have varied over time and may be less stringent for particularly credit worthy borrowers. The qualifying incomes on Table 8.12 are approximations calculated by taking 4 times the annual mortgage payment without adding in taxes and insurance.

A graphical comparison of the essential features of the 7.5% and 15% mortgages will provide assistance in understanding how inflation affects the mortgage options available to potential homeowners. Figure 8.13 reports on a 30-year $200,000 mortgage at 7.5%. The required $1,398 nominal monthly mortgage payment is represented by the horizontal line on the top panel. During the early years of the mortgage, the interest costs will eat up almost all of the mortgage payment, with little left for paying off the mortgage. As years go by the principal of the mortgage is gradually paid off, and as a result, the interest payments take a smaller and smaller portion of the monthly payment. Specifically, if P_{t-1} was the amount outstanding at the end of last period, interest payments of iP_{t-1} will be due this period; if M is the monthly mortgage payment, then $M - iP_{t-1}$ will be available to pay off the loan. Therefore, the outstanding balance on the mortgage at the end of period t will be $P_t = (1 + i)P_{t-1} - M$. The bottom panel of Figure 8.13 shows how the outstanding balance declines over the life of the mortgage.

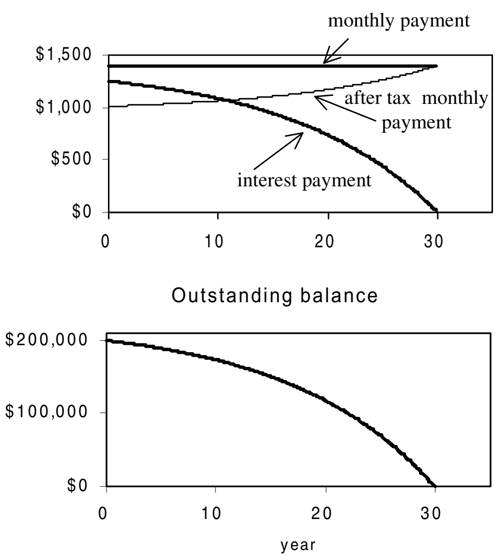


Fig. 8.13. A 30 year $200,000 conventional 7.5% mortgage
During the early years of the mortgage, most of the monthly payment is used to pay interest on the outstanding balance. As more and more of the mortgage is paid off, the portion of the monthly payment used to cover the interest on the outstanding balance declines. This means that the interest deduction on the federal income tax return gradually declines, causing the after-tax mortgage payment to increase.

In the United States, a major tax advantage to owning your own home arises because the interest payments are deductible in computing one's personal income tax. This means that the net after-tax cost of the monthly payment is $(A_t - tiP_{t-1})$, where t is the household's marginal tax rate, say 30%, and P_{t-1} is the outstanding principal. The monthly interest payment and hence this tax advantage decline over the years as the debt is gradually paid off. That is why the after-tax monthly payment curve is upward sloping on the upper panel of the graph.

The conventional mortgage contract works out quite differently in times of inflation, as can be seen by comparing Figure 8.14 with Figure 8.13. Assuming the 8% per annum inflation is anticipated, the nominal interest rate will rise to about 15%. With a 15% interest rate, the monthly mortgage payment is initially a whopping \$2,529. But because of inflation the real value of the monthly mortgage payment falls through time, as is illustrated by the downward sloping curve on the upper panel of Figure 8.14. After about seven years, the real value of the monthly payment is the same as it would have been if there had been no inflation. The real value

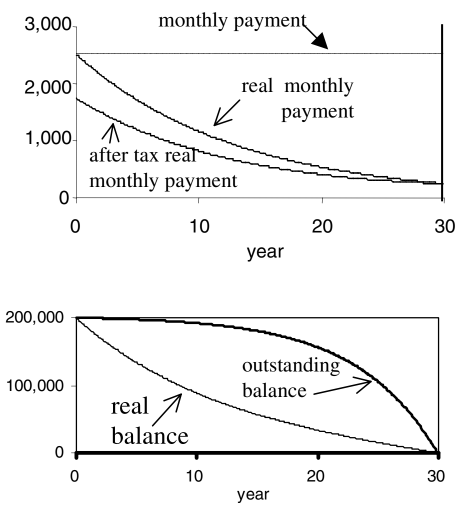


Fig. 8.14. A 30-year conventional mortgage with 8% anticipated inflation
When inflation is anticipated, rising interest rates make monthly mortgage payments much more difficult to meet during the early years of the mortgage. If the inflation persists, the real after tax burden of the mortgage declines and eventually falls below the mortgage payment that would have been imposed if prices had been stable. Thus anticipated inflation front-loads the mortgage.

Table 8.13. 30-year 15% conventional mortgage at 8% anticipated inflation.

A $200,000 mortgage with nominal monthly payments of $2,529;
real interest rate = 6.481%

Month	CPI	real monthly payment	outstanding balance nominal	outstanding balance real	after tax payment nominal	after tax payment real
12	108.0	2342	199,628	184,841	1,755	1,625
60	146.9	1721	197,441	134,375	1,764	1,200
120	215.9	1171	192,050	88,956	1,784	826
240	466.1	543	156,748	33,630	1,919	412
360	1,006.3	251	0	0	2,519	250

of the monthly mortgage payments continue to fall and are trivialized by inflation long before the 30-year mortgage is fully paid off. All this means that the conventional mortgage, which involved a level monthly payment throughout the life of the mortgage when prices were stable, becomes heavily front-loaded in times of anticipated inflation. Thus the anticipation of inflation tilts toward the present the cost of buying a house in terms of actual purchasing power. While everything may balance out over the 30-year life of the mortgage, this offers little solace to the household that cannot afford to buy that dream house because of the prohibitively expensive upfront payments.[24]

Markets do adapt to inflation, but not always completely. In the inflation of the 1970s and 80s, nominal interest rates soared. The real rate of interest did not fluctuate nearly as much as nominal rates. But the adaptation of markets to inflation was incomplete, and despite its inadequacies the conventional mortgage instrument survived the great inflation. Banks did introduce *variable rate mortgages* — with this type of mortgage instrument your monthly payment is adjusted periodically to reflect changes in nominal interest rates. This protected the banks from the loss imposed when unanticipated inflation pushed short term interest rates above the rates they were being paid on mortgages that they had issued when prices were stable, but the variable rate mortgage exposed the home owner to the

[24]It is interesting to note that those households that did succeed in financing their home when interest rates were high in the 1980s were able to escape the burden once inflation was brought under control and the mortgage rate returned to normal levels. The borrowers could refinance their mortgage at a lower rate by negotiating a new mortgage, perhaps with another bank, at the new competitive market rates, using the proceeds to pay off the remaining balance on their original mortgage.

prospect of substantial increases in nominal monthly payments if and when inflation returned.

8.5.4 *How harmful is inflation?*

How much pain inflation generates depends in part on whether it is anticipated. If inflation is a surprise, borrowers gain and lenders lose because loans are repaid with dollars of reduced purchasing power. If inflation is anticipated, nominal interest rates adjust in accordance with the Fisher equation, keeping the real interest rate stable.

There are ways in which the market can lessen the dislocations of unanticipated inflation. For example, when workers negotiate cost-of-living adjustment clauses in their labor contracts, they are protecting themselves from inflation surprises. Similarly, lenders and borrowers can reduce the risk of unanticipated inflation or deflation by specifying the real rather than the nominal rate of interest in the loan contract. The U.S. Treasury does issue "indexed bonds" which are designed to insulate the lender from inflation by, in effect, guaranteeing a specific before tax real rate of interest over the life of the loan. But in practice the market does not offer all that much protection. As but one example, we have seen that potential homebuyers find that when inflation threatens, their initial real monthly mortgage payments increase dramatically

While inflation, particularly when unanticipated, has marked effects on the distribution of real income and wealth, this is not its only cost. Inflation adds to uncertainty. It can lead to a crisis in the foreign exchange markets. More than this, it can cause confusion and misjudgments about appropriate prices that can contribute to inefficiencies in the allocation of resources. When inflation is extremely fast, everyone's attention is diverted to the task of coping with inflation. Workers may take time off from work in a rush to cash their latest paycheck. In countless ways, inflation works to undermine productivity.

8.6 Business cycles

Markets fluctuate. The economy has both good and bad years. With some regularity, cyclical reversals interrupt the growth of capitalistic economies. The bursting of the speculative South Sea Bubble in 1720 was only one of a series of financial panics that frequently harassed the money markets throughout the 18th century until the present day. The history of the business cycle in the United States since the Civil War has been meticulously recorded by scholars at the National Bureau of Economic Research (NBER). The NBER defines a *recession* as a period of significant decline in total output, income, employment, and trade, usually lasting from six months to a year, and marked by widespread contractions in many sectors of the economy. A *depression* is a recession that is particularly severe in terms of both depth and duration. The NBER recessions are traditionally marked on economic time-series graphs by shaded vertical stripes corresponding to periods of recession, such as those on Figure 8.1.

The upper turning point or peak of the business cycle is the date at which the economy moves from boom into recession. The lower turning point or trough is the date at which the economy moves from the recession into the recovery phrase. Scholars at the NBER have identified the chronology of business cycle turning points, recorded on Table 8.14. As is clear from the table, the word "cycle" is a misnomer if it conveys the impression that business fluctuations occur with a steady and uniform rhythm. The

Table 8.14. NBER business cycle reference dates.

Reference dates		Length in months of cycle phase			
Trough	Peak	Contraction Trough From Previous Peak	Expansion Trough to Peak	Cycle Trough from Previous Trough	Peak from Previous Peak
Jun-38	Feb-45	13	80	63	93
Oct-45	Nov-48	8	37	88	45
Oct-49	Jul-53	11	45	48	56
May-54	Aug-57	10	39	55	49
Apr-58	Apr-60	8	24	47	32
Feb-61	Dec-69	10	106	34	116
Nov-70	Nov-73	11	36	117	47
Mar-75	Jan-80	16	58	52	74
Jul-80	Jul-81	6	12	64	18
Nov-82	Jul-90	16	92	28	108
Mar-91	Mar-01	8	120	100	128
Nov-01		8			
Average, all cycles:					
1854–1991 (32 cycles)		17	38	55	56
1854–1919 (16 cycles)		22	27	48	49
1919–1945 (6 cycles)		18	35	53	53
1945–2001 (10 cycles)		10	57	67	67

Source: National Bureau of Economic Research: http://www.nber.org/cycles.html.

expansionary phases of the business cycle since World War II have averaged 57 months; the average recession has lasted 10 months. But there is considerable variation. Thus the contraction from July 1981 to November 1982 lasted 16 months, but the next recession lasted only 8 months. The ten-year expansion enjoyed in the glorious 1990s is the longest expansion in recorded history. And there is tremendous variation in the amount of unemployment experienced in different recessions. The level of unemployment increased from 3.2% to 24.9% from 1929 to 1933 — clearly this was a depression of the first order. From 1960–2 to 1961–1 the unemployment rate climbed by a modest 1.1%, from 5.6 to 6.7%. Clearly, no two recessions are alike.

8.6.1 *Identifying business cycle peaks and troughs*

The *NBER Business Cycle Dating Committee*, a small group of distinguished academic macro economists, is generally recognized as providing the authoritative dating of business cycle turning points. Throughout the summer and fall of 2001 the financial press devoted much space to the question of whether the U.S. economy was entering a recession, but the NBER's Business Cycle Dating Committee waited until the accumulating evidence on economic conditions became clear.[25] Finally, on November 26th the NBER announced that the economy had been contracting since March, 2001. Thus there was an eight month *recognition lag* before it was clearly established that the economy was indeed suffering from recession. Eight months may seem like an inordinately long time before recognizing an economic illness, and the uncertainty obviously create serious problems for government policy makers concerned with stabilizing the economy. But often the recognition lags are even longer, and difficulty in promptly diagnosing current economic conditions may affect election outcomes. For example, the fact that March 1991 marked the end of a recession was not identified by the NBER Business Cycle Dating Committee until December 1992, well after George H. W. Bush had lost his battle for reelection to William W. Clinton, in part because of voter concerns about the state of

[25] Economic journalists sometimes say that a recession begins when GDP declines for two quarters in a row. One problem with this definition is that the GDP data is subject to frequent revision, which means that the question of whether GDP actually declined for two quarters may not be resolved for quite some time. The NBER relies on a host of monthly time series in order to more promptly appraise the condition of the economic patient.

the economy. The committee did not recognize until July 2003 that the George W. Bush recession had ended in November, 2001.

8.6.2 *Forecasting with leading economic indicators*

From the NBER Business Cycle Chronology on Table 8.14 it is clear that the task of forecasting economic turning points cannot be easily addressed by any naïve projection technique, such as simply predicting that the current boom will be of the same duration as the last. The business cycle is much too variable in both length and intensity for that. An alternative technique is to employ "barometric" procedures. This approach tries to identify one or more economic variables that will serve as leading indicators by consistently turning in advance of the change in overall business conditions, just as changes in barometric pressure are used to forecast the weather.

Arthur F. Burns and Wesley C. Mitchell published in 1938 the results of a major research effort at the National Bureau of Economic Research that was based on the barometric approach. They had examined several hundred economic time series, some stretching as far back as the Civil War. Most of the series did not have precise cyclical patterns. A major conclusion of the investigation was that there is no single infallible barometer of cyclical movements in economic activity. Nevertheless, the authors concluded that a skillful practitioner would find certain indicators particularly helpful in analyzing business cycle developments.

The most useful of the leading economic time series have been consolidated into a single index, called the *Index of Leading Economic Indicators*, which is now compiled by The Conference Board, a non-partisan not-for-profit association of business leaders. The list of leading economic indicators, which has changed over the years, includes such series as the following: average weekly hours of work in manufacturing, average weekly initial claims for unemployment insurance, manufacturers' new orders for consumer goods and materials, manufacturers' new orders for capital goods, building permits for new private housing units, prices of 500 common stocks, the money supply and interest rates.

Critics have repeatedly objected that barometric indictors can make two types of errors:

1. Missed turns: the indicator may fail to signal a business cycle turning point
2. False signals: the indicator may predict a turning point that does not materialize.

When proper deductions were made for both types of errors, said the critics, the indicators were of no direct forecasting value. Maybe so, but defenders can argue that the Index of Leading Economic Indicators is the best one-eyed monster in the country of the blind.

8.6.3 *Econometric models*

Forecasting the business cycle is said to be more art than science. At times it has been a profitable business. In the 1960s Harvard University Professor Otto Eckstein founded Data Resources Incorporated (DRI), University of Pennsylvania Professor Lawrence Klein established Wharton Econometrics, and Michael Evans entrepreneured Chase Econometrics. These three enterprises competed in the sale of consulting services and economic forecasts. These were profitable businesses that appeared to be operating on the cutting edge of economic science. Each enterprise developed large scale empirically based simultaneous equation econometric models estimated with historical data. The DRI econometric model involved more than 800 equations! In the 1970s and early 1980s their forecasts were regularly cited on the evening TV news. But in practice, it turned out that they were not all that successful at predicting economic developments and they have faded from the public's eye. Nowadays econometric forecasts are not taken nearly as seriously as they were a quarter of a century ago.

8.6.4 *Consumer Sentiment*

Economic journalists often focus on the Index of Consumer Sentiment when discussing what the future holds in store for the economy. The *Index of Consumer Sentiment*, compiled by the University of Michigan's Survey Research Center since the mid 1950s, is based on the qualitative answers provided by telephone respondents to five questions:[26]

1. "Would you say you (and your family) are better off or worse off financially than you were a year ago?"

[26] These attitudinal questions were *not* originally intended to elicit useful information. Rather, when Michigan Professor George Katona was developing the Survey of Income and Wealth for the Federal Reserve Board, he added the attitudinal questions in order to loosen up the respondents so that they would be more willing to answer the questions about their income and other personal financial details. When he analyzed the results he concluded that they had useful forecasting value.

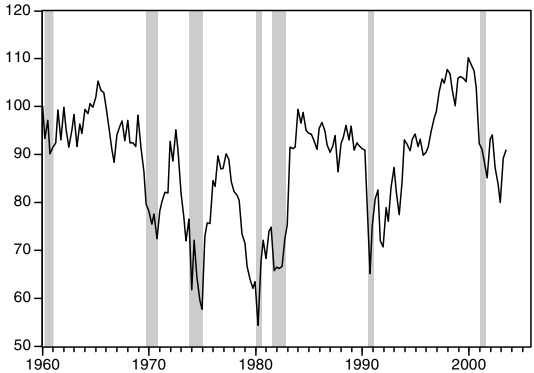


Fig. 8.15. University of Michigan Index of Consumer Sentiment

2. "Now looking ahead — do you think that a year from now you will be better off financially, or worse off, or just about the same as now?"
3. "Now turning to business conditions in the country as a whole — do you think that during the next 12 months we'll have good times financially, or bad times, or what?"
4. "... which would you say is more likely — that in the country as a whole we will have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?"
5. "... do you think now is a good or bad time for people to buy major household items?"

Responses to these five questions are averaged in constructing the Consumer Sentiment Index.

The Michigan Index of Consumer Sentiment and an alternative Consumer Confidence Index compiled by The Conference Board are frequently cited by economic journalists. Some economists have found these indices helpful in explaining fluctuations in consumption and other economic variables, but consumer sentiment is a far from precise guide as to what the future holds in store for the economy.

Summary

1. Examination of an Economic Report Card summarizing the performance of the economy on a decade by decade basis since the great depression revealed that each decade contained surprises. The Great

Depression decade of the 1930s was a disaster, the unemployment rate reaching 25%. The 1970s and early 1980s were stagnation years, the economy suffering simultaneously from high inflation and high unemployment. The performance of the American economy was spectacular during World War II. In the 1990s the United States enjoyed low unemployment without inflation, growing for an unprecedented ten years straight without a recession.

2. The unemployed are those who want work but cannot find jobs. The labor force is the sum of those who are working plus the unemployed. The unemployment rate is the percentage of the labor force that is unemployed.
3. The seasonally adjusted unemployment rate has been corrected for the systematic movements that characterize particular seasons of the year, such as the tendency for the unemployment rate to rise in June when students getting out of school are looking for jobs. Because the initial unemployment rates are subject to extensive revision as new data become available, month-to-month fluctuations in the current estimates of the unemployment rate may not accurately reflect changes in underlying economic conditions.
4. Unemployment discriminates. In both boom and recession the black unemployment rate is approximately twice the white unemployment rate.
5. Okun's Law is the proposition that a 1% increase in the unemployment rate may cause a 2.5% shortfall in the nation's output, GDP.
6. Inflation obeys the law of compound interest. The "rule of 70" says that if inflation takes place at rate $\dot{p}$ per annum then the price level will double in approximately $70/\dot{p}$ years.
7. The Bureau of Labor Statistics measures the overall price level for the United States each month by taking a weighted average of the prices of the goods and services purchased by the typical consumer. In estimating price movements the statisticians try to allow properly for the development of new products, quality changes and for the changing tastes of consumers. The price index is normalized so as to equal 100 in the base year (or years), such as 1982–84 = 100.
8. "Substitution bias" is the difficulty created in the construction of price indices by the fact that consumers tend to substitute away from commodities that have the largest price increases and tilt toward items that have become relatively cheaper. The Laspeyres price index uses the earlier quantity weights, which tends to overstate inflation, while

the Paasche uses the most recent quantity weights, which tends to understate inflation. Irving Fisher advocated the use of the geometric average of these two indices.

9. Because of price changes, a comparison of changes in money wages over the years does not reveal what has happened to actual purchasing power. The real wage, obtained by dividing the money wage by the price index, facilitates such comparisons. Wages are said to be "indexed" if the worker's contract calls for the money wage to be adjusted in response to changes in the price index.
10. The nominal rate of interest quoted in loan contracts is misleading in times of fluctuating prices. The real rate of interest, which is usually approximated by computing $r^r = r^n - \dot{p}$, takes account of the fact that rising prices mean that the money paid back at the end of the loan may have less purchasing power than the money original lent. The *ex post* real rate of interest is calculated using the actual change in prices. The *ex ante* real rate is the rate calculated with anticipated price changes, which are hard to measure.
11. The Fisher equation, named for Irving Fisher, states that the nominal rate of interest is the real rate plus the anticipated rate of inflation, $i^n = i^r + \dot{p}$. Given the real rate, which may be determined by the supply of savings and the demand for investment funds, any factor that causes the public to anticipate more inflation will cause the nominal rate to climb.
12. Borrowers benefit from unexpected inflation because they are able to obtain funds at low rates of interest but pay back the loans with dollars of reduced purchasing power. Lenders lose. After a period of protracted inflation the nominal rate of interest rises because the public learns to anticipate that inflation will continue. If the rate of inflation falls below the anticipated rate, it is the borrowers who lose because they will find themselves paying a high real rate of interest while the lenders gain because they receive back more in purchasing power than the market anticipated. Adaptation to anticipated inflation may be incomplete, as was the case with house mortgage instruments in the 1980s.
13. The business cycle refers to the tendency for the economy to fluctuate between good and bad times, boom and recession. The difficult task of diagnosing when the economy has entered a recession has been assumed by the Business Cycle Dating Committee of the NBER. Identifying precisely when the economy has entered a recession is far from easy and can involve a recognition lag of more than a year.

14. The Index of Leading Economic Indicators can be used as a barometer to forecast when the economy has entered into a recession or moved on to recovery. It is constructed by averaging a number of economic time series that usually turn down in advance of recession. Unfortunately, the index is not particularly reliable, sometimes missing turns and sometimes giving a false signal of recession when none occurs. The Index of Consumer Sentiment, based on a survey of consumers about their current economic condition and hopes for the future, is not a particularly reliable forecasting device. Complex multi-equation econometric forecasting models have not lived up to their promise. Thus, recessions are often a surprise not only to the public and the stock market but also to policy makers attempting to stabilize the economy.

Appendix 8.1. Compound interest

Money in the bank

If you leave a principle of P dollars on deposit for 2 years at rate of interest i you will have at the end of the second year $A_2 = (1+i)^2 P$. More generally, for any number of periods n you will have

$$A_n = (1+i)^n P\,. \tag{24}$$

For example, if you leave $P = \$1{,}000$ in the bank for 20 years at interest rate $i = 10\%$ you will have $1.1^{20} \times 1000 = \$6{,}727.50$.

Suppose the Second National Bank also offers you an interest rate of 10% but will compound the interest (compute the interest due and credit it to your account) every six months.[27] At the end of six months you will have earned \$50 of interest and have \$1050 in your account. At the end of the year you will have earned an additional \$52.50 on your \$1050 deposit and have \$1,102.50 in your account, which is \$2.50 more than if interest were compounded only annually. We may calculate for any horizon of n years:

$$A_{n,2} = (1+i/2)^{2n} P\,. \tag{25}$$

[27] Up until 1978 the Federal Reserve Board's Regulation Q prohibited savings and loan institutions and commercial banks from paying an annual rate of interest higher than 5.5% on savings deposits. Many banks tried to compete for additional deposits by compounding quarterly, monthly or even continuously.

In particular, for $i = 10\%$ and $n = 20$ you will have \$7,039.99. If the bank compounded daily, 365 days per year, one would have \$7,387.03. In general, to find the value for a horizon of n years with frequency of compounding k we calculate

$$A_{n,k} = (1 + i/k)^{kn} P \tag{26}$$

Finally, suppose that we find a bank that compounds continuously. Now

$$A_{n,\infty} = \lim_{k\to\infty} (1 + i/k)^{kn} P = e^{in} P\,, \tag{27}$$

where the second equality is a standard calculus result. Thus \$1,000 earning interest at $i = 10\%$ compounded continuously for $n = 20$ years will yield \$7,389.06.

Mortgage

You have purchased a house, signing a mortgage for the borrowed amount $P = \$100{,}000$ at 8% interest payable in equal monthly installments over 30 years. You are committing yourself to $n = 30\times 12 = 360$ equal payments, but what will be your monthly mortgage payment, call it M?

The first payment, due in one-month's time, will have a present value of $M/(1 + i/k)$, where $k = 12$. The second payment will have a present value of $M/(1 + i/k)^2$, an ith payment $M/(1 + i/k)^i$, etc. The present value P of those n payments must equal the \$100,000 that the bank is advancing to you:

$$P = M/(1+i/k)+M/(1+i/k)^2+\cdots+M/(1+i/k)^i+\cdots+M/(1+i/k)^n \tag{28}$$

It will be easier to solve for M if we substitute $\beta = 1/(1+i/k)$ into equation (25), obtaining

$$P = M\beta(1 + \beta + \beta + \cdots + \beta^{n-1})$$

Since the expression in parentheses is a truncated geometric series with $0 < \beta = 1/(1 + i/k) < 1$, it sums to $(1 - \beta^n)/(1 - \beta)$ and so $P = M\beta(1 - \beta^n)/(1 - \beta)$;[28] therefore, the monthly payment will be

---

[28]The expression for the sum of a geometric series with $n + 1$ terms was discussed in footnote 7 of Chapter 6. In the present application there is one less term in the series.

$$M = P\frac{1-\beta}{\beta - \beta^{n+1}} \,. \tag{29}$$

For our example we have $P = \$100{,}000$, $n = 12 \times 30 = 360$, and $i = 8\%$ per annum or a monthly rate of $i_m = i/12 = 0.6667\%$.[29] Thus $\beta = 0.99338$, which yields a monthly payment of \$733.76.

Bonds

Bonds are certificates constituting a legally enforceable promise to pay a specified amount of money on the maturity date plus periodic payments (e.g., every 6 months) until the bond matures. For example, suppose you inherit from your late aunt an 8% \$10,000 corporate bond. The certificate says that the bond pays an 8% annual rate in biannual installments and that it matures in year 2020. This is a promise to pay the holder of the bond \$10,000 in year 2020 plus payments of \$400 (8% of \$10,000/2) every six months until maturity. At the bottom of the bond there are a series of dated coupons with a face value of \$400, one maturing every six months until the bond itself matures. Twice each year you clip the appropriate coupon off the bottom of your bond and ask your banker to redeem it for you.

Bonds are usually readily transferable because they are "issued to bearer," meaning that the payments are made to whomever holds the certificate when a payment is due. This not only means that bonds should be kept in a safe deposit box. It also means that as an alternative to clipping coupons until maturity in year 2020, you may sell your bond through a broker. How much you will receive if you sell before maturity depends on supply and demand for such securities, which is influenced by general money market conditions and monetary policy.

---

[29] By convention, established long before pocket calculators and computers made involved computations cheap, interest rates for a fraction k of a year are calculated using simple interest: i/k. The correct compound interest formula for fractions of a year is $(1+i)^{1} \quad -$ 1. With an 8% annual rate of interest, the monthly compound interest rate is 0.6434% rather than 0.6667%. This is a small difference, but the compounding procedure yields a monthly payment of \$714.40 rather than \$733.76 on a 30 year \$100,000 loan. The discrepancy is larger when rapid inflation has pushed up interest rates. This appears to be to the advantage of bankers, but if competition among bankers determines the monthly payment, then the only effect of using simple rather than compound interest would be to change the interest rate quoted in the loan document.

Let us consider the problem of determining the value today of a bond with maturity value M in n years that yields annual payments of rM, where r is the "coupon rate" quoted on the instrument.[30] We need to consider the present value of the stream of future interest payments plus the present value of the principle. Suppose that the current market yield on bonds of this quality and maturity is $i = 9\%$, what will be the current value of the security? The value of the bond today is

$$V = \left[\frac{\beta - \beta^{n+1}}{1-\beta}\right] rM + M\beta^n \tag{30}$$

where $\beta = 1/(1+i)$ is the discount factor, the first term on the right of the equality is the value today of the stream of future coupon payments, and the second term is the present value of the bond's face amount of M that will be received on maturity. In the case of our bond, $r = 9\%$, $i = 8\%$, $M = \$10{,}000$ and $n = 20$, we have $\beta = 0.917431$ and

$$\begin{aligned} V &= \left[\frac{0.917431 - 0.917431^{21}}{1 - 0.917431}\right] 0.08 \times 10{,}000 + 10{,}000 \times 0.917431^{20} \\ &= \$9079.92\,. \end{aligned}$$

Because the 8% bond must compete with other assets currently yielding 9%, its value falls below its par value of \$10,000 to \$9079.92.

Key Concepts

base year, 360
Bureau of Labor Statistics (BLS), 342
Census-X12 (ARIMA), 348
chain linked index, 369
civilian non-institutional population, 342
Consumer Price Index, 362
 CPI-U, 369
Consumer Sentiment index, 388
depression, 385
discouraged workers, 342
employed, 342
employment population ratio, 342

[30] Bonds are printed with a series of coupons at the bottom of the certificate. For a 20 year \$10,000 8% bond with semi-annual interest payments there will be 40 coupons, each worth \$400 and one maturing every six months over the life of the bond. The holder of the bond clips off the coupon and turns them in to a banker for collection.

Exercises

1. A preliminary statistical analysis of data for the United States suggests[31]

$$(Y_t - Y_{t-1})/Y_{t-1} = 0.0081 - 1.83(U_t - U_{t-1}) + e\,.$$

where t denotes the t-th quarter (3 month period), Y_t is GDP and U% is the percentage unemployment rate. Because the equation does not fit the data exactly, an e is added on the extreme right to represent the error of fit. The average value of e is zero. For this exercise we can pretend that $e = 0$ always.

[31] The equation was estimated by the method of "least squares."

a. How rapidly does the equation suggest output will have to grow each quarter, because of population growth and increased worker productivity, just to keep unemployment constant?
b. How much would GDP have to grow next quarter in order to reduce the unemployment rate by 1%.
c. In 1982, when the civilian unemployment rate averaged 9.7%, GDP was $4,624 billion (1992 dollars). How much more output would be produced in the next period, according to this equation, if the unemployment rate were miraculously reduced to 5.5%?

2. In one month in the 1978–1986 inflation, Israeli prices climbed 27.5%.

   a. What was the annual rate of inflation? Was this a "hyper-inflation"?
   b. If you had a deposit in the bank earning 25% interest, what was the real rate of interest?
   c. If you were subject to a 30% marginal tax rate, what was your after tax real rate of interest?
   d. What rate of interest would the bank have to pay in order for you to receive a 5% annual return after taxes?

Table 8.15. Army pay.

Year	Monthly pay Private E1	Consumer Price Index (1982–84 = 100)	Real Monthly Wage of Private E1 (1982–84 dollars)	(year 2001 dollars)
1963	$78	30.6	______	______
1973	______	44.4	734	______
1978	397	65.2	______	______
1987	658	113.6	______	1,026
1989	699	______	564	______
2001	964	177.1	______	964

Source: *Statistical Abstract of the United States.*

3. Consider the data on pay for new army recruits on Table 8.15.

   a. Fill in the blanks on the table.
   b. In what year did privates receive the highest real pay?
   c. At what rate per annum was the general price level increasing.
   d. If prices continue to increase in the future at the same rate as from 1963 to 2001, how long will it take for the CPI to reach twice the level of 2001?

e. If you arranged a loan from your bank at 10% interest today, and you expected the same rate of inflation to prevail in the future as from 1963 to 2001, what would be your *ex ante* real rate of interest? Explain.

4.* Construct a graph similar to Figure 8.9 but explaining why the Paasche index tends to understate inflation.

9

GDP Accounting and the Multiplier

9.1 Overview

What determines the pace of economic activity? How rapidly will output grow? Why is our economy subject to abrupt fluctuations, with exhilarating booms and inflation followed by periods of recession and excessive unemployment? These are important but difficult questions that have worried generations of economists involved in the study of macroeconomic policy. And the answers given by economists have varied, not only over the decades but also among economists belonging to the same generation but associated with different schools of economic thought. Macroeconomics is the study of these economy wide issues. To discuss macroeconomic policy is to open up for public viewing a host of controversial issues.

When President George W. Bush took office in 2001, one of his very first orders of business was to push for a tax cut. Bush had campaigned on the promise that he would cut taxes. But he was particularly eager to get the tax cut in place because he wanted to stimulate an economy that was thought to be faltering on the edge of recession. Far from being a historic first, forty years earlier President John F. Kennedy had pushed for a tax cut upon entering the White House in 1961. In both instances, the tax cut policy initiative was motivated, at least in part, by a desire to stimulate the economy. Cutting taxes in order to influence the pace of economic activity is one type of *fiscal policy.* Fiscal policy is the strategy of adjusting the federal budget by changing tax rates or the level of government spending in order to contribute toward the achievement of macroeconomic objectives, such as full employment, sustained long-term economic growth, and price stability.

Stimulating the economy by increasing government spending or cutting taxes is controversial economic medicine. When President Franklin D. Roosevelt ran for office in 1932, he promised to balance the government's budget. He failed to deliver on that promise, in part because tax revenues were down as a result of the Great Depression. Republicans complained when Roosevelt ran for reelection in 1936 and again in 1940 that he was spending the country into bankruptcy. But when World War II forced a massive increase in government spending that was only partially financed by higher tax revenues, the unprecedented government deficits were associated with record low unemployment and a massive expansion of Gross Domestic Product. At the time the contrast between the wartime boom and the Great Depression was interpreted by many as evidence of the potency of fiscal policy.

This chapter will develop the multiplier, a concept clearly articulated in 1931 by R. F. Kahn, a lecturer at Cambridge University, England. Kahn developed a model explaining how changes in government spending and tax policy might help to stimulate economies suffering from depression. The multiplier was adopted by John Maynard Keynes as a key component to his *General Theory of Employment, Interest and Money*, published in 1936. This book sparked the Keynesian revolution, the fundamental rethinking of economic thought required in order to explain the Great Depression.

The multiplier model is recommended as a starting point because it is fairly simple. But simplicity is achieved at a cost. The multiplier model is a bare-bones model that is not capable, by itself, of providing an adequate explanation of inflation or a guide to monetary policy. In later chapters we will augment the initial model in order to arrive at a more realistic model capable of explaining a wider range of phenomena.

While fiscal policy has been advocated by both Democratic and Republican presidents, many macroeconomists argue that its efficacy has been greatly overrated. The multiplier model that will be developed in the first part of this chapter purports to provide an explanation of how fiscal policy works. Later in the chapter we will review a variety of arguments that have been advanced by economists to explain the limitations of the multiplier model before proceeding in subsequent chapters to develop a more elaborate framework for examining how fiscal and monetary policy work.

9.2 GDP, disposable income and consumption spending

9.2.1 *Defining GDP*

Gross Domestic Product is easy to define:

> GDP is the market value of all the goods and services produced by the economy during a specified time period, such as a year.

The concept is easy, but there are some subtle points requiring clarification.

Goods and services

GDP includes not just physical *goods* but *services* as well. It includes the services of the barber, the college professor, and the taxi driver as well as the value of the automobiles, bread and other goods produced during the year. Services are just as much a part of the output of the economy as books and cars and other physical objects. Services, like commodities, produce satisfaction. Services amount to about 45% of all consumer spending.

Market value

You cannot add apples and pears, so how can the national income accountant add together the output of all the different goods produced by a modern economy? The trick is to add up the *values* of all the different kinds of output. A nation's currency, e.g., the dollar bill, provides the common denominator. For example, if red delicious apples cost $8.00

per bushel and one million bushels are produced, then these apples add \$8 million to this year's GDP. If farmers also produce half a million bushels of pears, valued at \$9.00 per bushel, this will add \$4.5 million to this year's GDP. We cannot add apples and pears, but we can add the *values* of the apples and pears together with the market value of all the other goods and services to arrive at GDP.

Final not intermediate product

Some goods are used up during the year in the production of other goods. Thus the wheat produced by the farmer is used to make flour, and the flour in turn is used by the baker in making bread. In this example the wheat and the flour are intermediate goods while the bread is the final product. Similarly, the steel used in the production of a car is an intermediate good rather than final product. Only the final product is counted in computing the Gross Domestic Product. To count both the final product (e.g., the bread) and the intermediate goods used up in its production (the wheat and the flour) as part of the GDP would lead to a grossly exaggerated statement of what the economy has produced in the way of goods and services. It would constitute "double counting," because only the bread is available to satisfy wants.

Also, the purchase of secondhand goods does not count as part of the GDP — goods and services are counted only in the year in which they are produced.

Accounting period

It takes time to produce output. In measuring the rate at which an oil well is operating, it is customary to speak in terms of barrels-per-day. Similarly, in measuring the output of the economy it is necessary to specify the time dimension. National income accountants customarily use one year as the basic accounting period. But *quarterly* as well as annual data can be extremely useful and is available for the United States since World War II. A quarter is a three-month period, a quarter of a year. Thus, 2000:1 refers to the first quarter of year 2000 (January, February and March) while 2000:2 is the second quarter (April, May and June), etc.

Annual rates

The GDP for 1999 was \$9,248.4 billion. The GDP reported for the first quarter of that year, denoted 1999:1, was \$9,072.7 billion. That does not mean that the output of the economy in those three months was $9,072.7 billion, because the data are reported at annual rates; i.e., the actual output of $2,268.2 was multiplied by four so that the numbers reported for each quarter would be immediately comparable to the annual figures.[1]

Flow not a stock

The GDP is not a measure of an economy's wealth. It measure the flow of current output rather than the resources available for the production of output. It includes the output of new cars during the production period but not the stock of cars existing at a particular point of time. GDP includes as investment the construction of new factories and new houses, but not the value of existing factories and houses. Thus a nation's GDP accounts are analogous to the income statement of a business enterprise, discussed in Chapter 5.3.2, rather than the firm's balance sheet.

Gross Domestic Product (GDP) versus
Gross National Product (GNP)

Gross Domestic Product refers to the goods and services produced *within* a country. Gross National Product is the market value of the goods and services produced with resources supplied by residents of the country. For example, the contribution to output in Canada made as the result of an investment by an American citizen in Canada is counted as part of America's Gross National Product. It is included in the Gross Domestic Product of Canada because that is where the output is produced.

9.2.2 *What's not in GDP*

GDP is not everything. Although the national income accountants make heroic efforts to measure many facets of the nation's output, they do not claim to be measuring economic welfare. In order to prepare estimates of output in a timely and precise manner, the national income accountants neglect a number of complicating items that are difficult and expensive to estimate. GDP does not include the value of do-it-yourself activities in the home, such as child care, preparing meals, household cleaning and

[1] If you travel 16 miles in 15 minutes, you will have been driving at an average speed of 64 miles per hour. Similarly, if an economy produces $2,250 billion of goods and services in April, May and June of 2000 (the second quarter) we will say that GDP in 2000–2002 was $9,000 billion at annual rates.

lawn care. GDP does not include the value of leisure time or of the food grown in a home garden. Television programs may be fun to watch, they may even at times be educational, but only cable services are counted as consumption spending. The GDP accountants regard the costs of producing broadcast TV as part of the marketing costs of the advertised commodity; it is an intermediate good rather than a component of GDP, and the cost of bringing that program to you shows up only indirectly in the consumption component of GDP, to the extent that it affects the price of Budweiser beer, brand name clothing, etc. On the other hand, GDP does include certain activities that do not so much produce satisfaction or contribute directly to welfare as limit or contain evil or prevent disaster. For example, such activities as crime control, fire protection, homeland security and national defense spending are sometimes referred to as "regrettable necessities," but they are part of the GDP.[2]

To adjust the official GDP estimates to take all such complications into account would at best provide a much less accurate estimate of a very interesting concept. When such factors are roughly proportional to GDP, the omissions would not be much of a problem. They may lead to more serious misinterpretations in times of major shifts in economic activity, such as World War II. When the economy was mobilized for all out war, the increase in measured GDP exaggerated the increase in output and economic welfare because it neglected the decline in unmeasured household production when women joined the workforce in unprecedented numbers.

From time to time attempts have been made to construct adjusted GDP estimates taking such complications into account. The most comprehensive of these studies was undertaken by Robert Eisner of Northwestern University.[3] Eisner reports that adding the value of child rearing services, meal preparation, and so forth results in an output estimate about 33% larger than official GDP. Including investment in human capital as measured by the value of the opportunity cost of student time might add almost 10% to GDP. Including the value of uncompensated jury time and volunteer work would add still more to GDP. Eisner's computations suggest that if all such items were taken into account, one might find that the output of the economy would be as much as 50% larger than the reported GDP.

[2]Cf. William D. Nordhaus and James Tobin, "Is Growth Obsolete?" in *The Measurement of Economic and Social Performance*, New York, National Bureau of Economic Research, 1973.

[3]Robert Eisner, "The Total Incomes System of Accounts," *Survey of Current Business*, January 1985. The percentages reported in the text were derived from his detailed accounts.

9.2.3 *Real GDP*

Care is required in the interpretation of the national income accounts because of the distorting effects of inflation. Unfortunately, the dollar provides a rubber yardstick for measuring the value of the nation's output because its value varies over time. For example, the United States GDP of \$7,813.2 billion for 1996 was measured with the prices prevailing in 1996 while the GDP of \$8,318.4 billion for 1997 was measured with prices prevailing in 1997; therefore, we cannot tell how much of the change of \$8,318.4 – \$7,813.2 = \$505.2 billion reflects an actual increase in the output of goods and services and to what extent the increases result from changing prices from one year to the next. This is similar to the problem of comparing wage rates over time that we discussed in explaining how the real wage is calculated in Chapter 8.5.1.

The *real* Gross Domestic Product data resolve the problem by reporting output measured in prices of a particular base year (e.g., 1996). The real value of GDP in 1996 is $\mathrm{GDP}^r_{1996} = \$7{,}813.2$ billion while for 1997 it is $\mathrm{GDP}^r_{1997} = \$8{,}159.5$ billion, where output in each case is measured with the prices that had prevailed in 1996. Thus the increase of $\mathrm{GDP}^r_{1997} - \mathrm{GDP}^r_{1996} = \$8159.5 - \$7{,}813.2 = \$346.3$ billion is the increase in actual output measured in 1996 prices. The data on the output of the American economy on Figure 1.1 were reported in terms of the prices that prevailed in 1999, which we call "GDP (billions of 1999 dollars)."

The National Income Accountants follow an involved two step procedure in calculating these estimates that is similar to the chain linked procedure described in Chapter 8.4.4. First they use Irving Fisher's geometric average formula to estimate the percentage change in GDP from one year to the next using price weights for each pair of adjacent years. Then these annual quantity changes are chained (multiplied) together to form the real GDP time series. This involved procedure is also used by the GDP accountants in estimating the real value of the components of GDP.

GDP price indices

The same procedure yields estimates of price movements for GDP, but with quantities instead of prices used as the weights. The most striking difference between the GDP price deflator and the Consumer Price Index that we discussed in Chapter 8.4 is that the consumer price index uses as its quantity weights estimates of the mix of goods and services purchased by the average consumer. The GDP price index includes all the goods and services in the GDP, including investment spending and government purchases — it includes factory buildings and jet fighters as well as consumer goods.

The national income accountants also compute chained price indices for GDP components, including consumption. The Personal Consumption Expenditure Price index (PCEPI) does not show precisely the same rate of inflation as the Bureau of Labor Statistics Consumer Price Index. For one thing, the PCEPI covers the spending of non-profit groups as well as individual consumers. Also, the PCEPI includes employer paid health insurance and Medicare expenditures while the CPI only includes out of pocket medical spending. And while the weights for the CPI are based on what household report in consumer expenditure surveys, the PCEPI is based on weights reported by business.

9.2.4 *The composition of GDP*

Every item in the Gross Domestic Product is carefully classified by the national income accountants into one of the following components: consumption (C), private investment (I), government spending on goods and services (G), exports (X) and imports (M). Letting Y denote GDP we observe that the total supply of goods available in the economy is $Y + M$. Since these goods must be used as C, I, G or X, we have:[4]

$$Y + M \equiv C + I + G + X\,. \tag{1}$$

Subtracting M from both sides yields the fundamental *national income accounting identity*:

$$Y \equiv C + I + G + X - M\,. \tag{2}$$

This identity will prove of fundamental importance when the time comes to explain how changes in government spending and tax policy may affect the level of GDP and unemployment. The breakdown of the United States GDP into these components for year 2000 is presented on Table 9.1. Now we must focus more closely on the definitions of each component

C ~ Consumption spending

This category of GDP spending includes goods and services purchased by the public. Spending on food, fuel and clothing are part of consumption.

[4]The identity symbol $\equiv$ indicates that the equality holds by definition.

Table 9.1. The composition of GDP, year 2000.

		billions $	% of GDP
Personal Consumption Expenditure (C)		6,728	68%
Gross Private Domestic Investment (G)		1,768	18%
Net Exports of Goods and Services			
Exports (X)	1,103		11%
less Imports (M)	1,467		15%
Foreign Trade Balance ($X - M$)		(−364)	−4%
Government Expenditures on Goods and Services			
Federal	590		6%
State and Local	1,151		12%
Total Government Expenditures (G)		1,741	18%
Gross Domestic Product (GDP)		9,873	100%

Note: Numbers in parentheses are subtracted.

Movies and baseball game admissions also count as consumption. So are services when purchased by a household, including those of a physician, a lawyer and a carpenter.

Some subtle points confront the GDP accountants in deciding what types of spending to define as consumption. Spending on education may be a wise investment, but the GDP accountants include your tuition and books as consumption. If you go to a state college or university, the cost of your education, to the extent it is paid for by taxes rather that by tuition and fees, is included in government spending instead of consumption.

$I \sim$ *Investment Spending*

Production for the future rather than for current satisfaction is involved in investment spending. Included in investment is the output of new machinery and production equipment. The construction of shopping centers and factory buildings, new houses and college dormitories are all forms of investment. These are all investment activities because they all involve the allocation of current output of goods and services to future uses.

The investment spending component of GDP does *not* include buying stock in a corporation or purchasing a New York City bond. The stockbroker may tell you that he is sure that your decision to buy shares in Microsoft is a wise investment, but the GDP accountant regards it as a paper transaction involving the exchange of existing assets and does not count it as part of GDP. However, when Microsoft constructs a new office building, this construction project does count as investment spending because people are being put to work and resources are being used to provide for the future.[5]

Gross versus Net investment

The national income accountants find that measuring investment spending is particularly difficult because so many different accounting procedures are used to compute depreciation, as was explained in Chapter 5.3.4. Therefore, they report two different figures for investment. Gross investment is the output of new capital goods without any deduction for depreciation. It is an *overstatement* of how much of current output is being set aside for future needs because it does not deduct for the wear and tear on existing equipment. Net investment is derived from gross investment by subtracting out an allowance for depreciation that is estimated by the GDP accountants. Because depreciation is difficult to measure precisely, net investment is a rather imprecise estimate of the addition that is being made to the nation's productive capacity.

The symbol I is ambiguous in that it is often used interchangeably to refer to both gross and net investment. When necessary to avoid confusion, I^n can be used to represent net investment and I^g for gross investment. Thus we have for net investment:

$$I_t^n = I_t^g - D_t \,, \tag{3}$$

where D_t is depreciation. The same distinction holds between Gross Domestic Product (GDP) and Net Domestic Product (NDP); i.e., $\text{NDP}_t = \text{GDP}_t - D_t$. The National Income Identity, equation (2), holds either way: Gross investment yields Gross Domestic Product (GDP):

$$Y^g \equiv C + I^g + G + X - M \,. \tag{4}$$

Net investment yields Net Domestic Product (NDP):

$$Y^n \equiv C + I^n + G + X - M \,. \tag{5}$$

A nation's capital stock, K_t, is the total quantity of productive resources that it possesses at point of time t, including factories and machinery and also houses and apartment buildings. With this definition we can say that

[5]Neither usage of the word "investment" is incorrect — they are homonyms. The two quite distinct concepts must be carefully distinguished, just as a goat that is a butter must not be confused with the butter that the gentle cow provides.

the investment component of NDP is the addition to the country's capital stock. If we let K_t denote the capital stock at the end of period t and I_t^n net investment in period t then

$$K_t = K_{t-1} + I_t^n = K_{t-1} + I_t^g - D_t\,, \tag{6}$$

where D_t is an estimate of depreciation. It rarely happens, but in some years in the Great Depression of the 1930s net investment was actually negative in the United States because the output of new capital goods was less than depreciation. When net investment is negative, the country's capital stock is shrinking, the productive capacity of the nation is in decline, and the welfare of future generations may be in jeopardy.

Inventory investment is part of the I component of GDP. Inventories include logs waiting to be cut at the lumber mill, new cars in transit from factory to the consumer, and bread on the shelves at the supermarket. Inventories serve as a buffer between producers and consumers, falling when sales exceed output and increasing when output exceeds sales. Inventory investment refers to changes in these stocks, and may be positive or negative. When the grocer adds new supplies to inventory, as when purchasing more bread than is sold, part of the economy's current output is being reserved for future use; this is positive inventory investment and is counted in GDP. When the grocer allows stocks to run down rather than replenishing all the goods on the shelf when they are sold, customers are obtaining more than is currently being produced, and the difference is called *negative* inventory investment.

G ∼ *Government spending on goods and services*

This category of GDP includes tanks and aircraft and munitions. In principle G should include the value of public education, but the task of placing a dollar price tag on this commodity is insurmountable. Instead, public education and other government services are valued at cost of production. Thus G includes the salaries of public school teachers. G also includes the salaries of policemen, but the payment to a private detective hired by a worried spouse to investigate marital infidelity is private consumption. When merchants hire the services of a Pinkerton detective to provide supplemental police protection (or to bust strikers over the head in the bad old days), this is part of the cost of production rather than a component of GDP.

The government provides veterans' benefits, social security payments and welfare payments, but they are *not* included in either G or GDP. They are examples of *transfer payments* in that they are *not* made in return for the provision of current services.

$X - M \sim$ *The foreign trade balance*

Exports, X, includes goods and services that are provided to foreigners. M, represents imports, the goods and services we receive from abroad. The difference between exports and imports is the foreign trade balance. The United States traditionally has a foreign trade deficit because we import more than we export.

9.2.5 *From GDP to disposable income, consumption and saving*

Figure 9.1 shows the time path of GDP, disposable income and consumption. Because we will find these concepts fundamental to an understanding of the multiplier process, we must study how they are related. In 2000 the GDP of the United States was \$9,873 billion dollars, but only \$7,031 billion filtered down to consumers as disposable personal (after tax) income, denoted Y_d. The \$2,842 billion gap between GDP and disposable income, summarized on Table 9.2, requires careful explanation.

- *Capital Consumption Allowances (Depreciation)*: This is an expense item from the point of view of the owners of the capital equipment that is depreciating, so it is not income to anyone.
- *Net Domestic Product (NDP)*: Subtracting the capital consumption allowances from GDP gets us to Net Domestic Product (NDP).

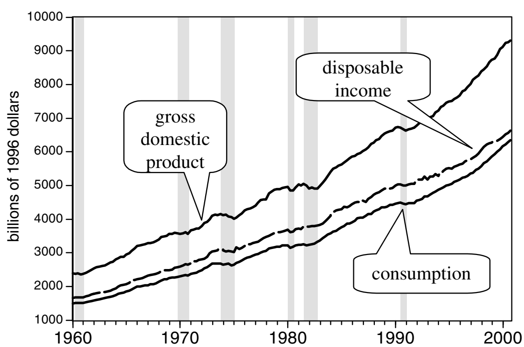


Fig. 9.1. GDP, disposable income and consumption

Table 9.2. From GDP to disposable Income and consumption in 2000.

	billions of $		% of GDP
1 Gross Domestic Product (GDP)		9,873	100%
2 Less Capital Consumption Allowances (Depreciation) (D)		(1,241)	13%
3 Net Domestic National Product ($= 1 - 2$)		8,632	87%
Taxes: (T)			
4 Personal Income Taxes (T)	1,288		13%
5 Indirect Business Taxes (T)	763		8%
6 Corporate Profits Taxes (T)	272		3%
7 Social Security Contributions (T)	702		7%
8 Less Total Taxes (T) ($= 4 + 5 + 6 + 7$)		(3,024)	−31%
Transfer Payments			
9 Government Transfer payments	1,036		10%
10 Business transfer payments	33		0%
11 Plus Total Transfer Payments (T) ($9 + 10$)		1,069	11%
Net Interest Payments, etc.			
12 Government Interest Payments	262		3%
13 Consumer Interest Payments	205		2%
14 Total Interest Payments (Int) ($12 + 13$)		468	5%
15 Less Undistributed Corporate profits (π)		(220)	−2%
16 Plus Miscellaneous		107	1%
17 Disposable Personal Income (Y)		7,031	71%
Personal Outlays			
18 Personal Consumption Expenditures (C)	6,728		68%
19 Interest paid by persons	205		2%
20 Personal transfer payments to rest of world	30		0%
21 Less Total Personal Outlays ($18 + 19 + 20$)		(6,963)	−71%
22 Personal Savings ($= Y$ − Total Personal Outlays)		68	1%

- *Taxes*: All taxes must be subtracted in moving from GDP to disposable personal income. In addition to the income tax, it is also necessary to subtract out sales taxes, excise taxes (e.g., the gasoline tax and the beer tax), corporate profit taxes and social security taxes.
- *Transfer Payments*: Added back in are social security payments, unemployment insurance, veterans' benefits and other payments made by the government and business firms to individuals that are not included in GDP because they are *not* payments for current productive services. These accrue as income to their recipients but they do not stem from the production of the nation's current output.
- *Net Interest Payments*: Interest payments by the government and consumer are not counted in the GDP but they are income to their recipients and belong in the lender's disposable income. Like transfer payments,

these interest payment must be added back in when moving from GDP to disposable income.

- *Undistributed Corporate Profits*: Unlike dividends, the sizable fraction of after tax corporate profits that are retained by corporations rather than distributed to stockholders obviously do not wind up as part of disposable income.[6]

These are the major items accounting for the sizable gap between GDP and disposable income. As can be seen from Figure 9.1, when GDP dipped during recessions, disposable income also declined. On average over the years, about 70% of GDP has filtered down to the consumer as disposable income.

What do consumer's do with their after tax income? As can be seen from the table, the great bulk is consumed. The purchasing power that has filtered down from GDP to the consumer is largely recycled back into the GDP when consumers make personal consumption expenditures — this is the C component of GDP in equation (2). Personal interest payments are listed as a separate component of personal outlays because such payments are not counted either in GDP or personal consumption expenditures (C). Personal saving is what remains of disposable income after personal consumption spending and interest payments. In year 2000 savings amounted to less than one percent of disposable income.[7]

9.2.6 *Two empirical relationships*

The scatter plot on Figure 9.2 presents the disposable income and consumption data of Figure 9.1 in a different way.[8] Each of the more than 40 data points reports for a particular year the level of disposable income, measured on the abscissa, and the level of consumption, plotted on the ordinate. The scatter plot suggests that a straight line would approximate the data points very well. That is to say, there is a quite tight linear relationship between consumption and disposable income. This relationship is

[6] Corporate retained earnings were discussed in Chapter 5.3.2.

[7] The national income accountants exclude both realized and unrealized capital gains from disposable income and hence saving. This means that personal saving did not include the substantial increase in wealth that consumers enjoyed in 1999 because of the soaring stock market. Or to put it another way, the saving ratio is much larger when capital gains are counted in income and savings.

[8] The scatter plot technique was used in Chapter 8.3.1 to show the relationship between black and white unemployment rates.

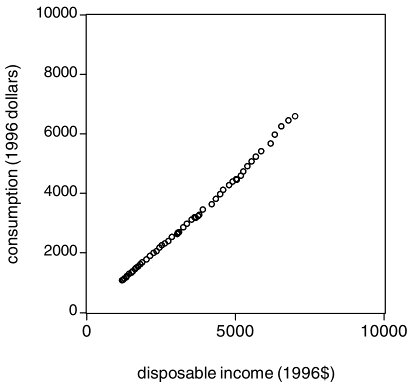


Fig. 9.2. Consumption on disposable income
This chart replots the data from Figure 9.1. Each dot represents the level of disposable income and consumption in a particular year. The scatter plot suggests a tight linear relationship between C and Y .

known as the consumption function:

$$C = c_0 + c_1 Y_d \,. \tag{7}$$

One could estimate the two parameters of this equation, the intercept c_0 and the slope c_1, by drawing a straight line to approximate the data points on the graph. When the computer was asked to fit the data as closely as possible with a straight line it reported[9]

$$C = -1.64 + 0.96 Y_d \,. \tag{7}^*$$

Thus the estimated value of the intercept is $c_0 = -1.64$ billion dollars and the slope is $c_1 = 0.96$. The parameter c_1 is called the *marginal propensity to consume* — the estimate of 0.96 suggests that consumers are likely to spend 96¢ out of every extra dollar of income they receive. While these estimates should not be taken too seriously, we shall use them in

[9]The parameters were estimated using a statistical procedure know as simple-least-squares with 162 quarterly observations covering the period from 1959:1 to 1999:2. The intercept suggests that if income were low enough consumption would be negative, but this should not be taken seriously. Such low ranges of income are completely outside the realm of past experience and even a profound pessimist would never expect such low levels in the future. The negative value helps to yield the best fit over the range of our observations. The linearity assumption might well break down at extremely low levels of income.

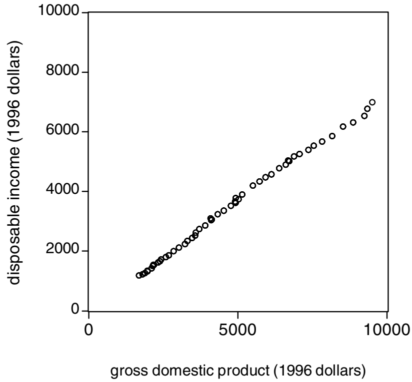


Fig. 9.3. Disposable income on GDP

developing illustrative numerical estimates of the multiplier effects of changes in government spending and tax policy on the level of GDP.

A second linear relationship, called the disposable income function, is suggested by the plot of disposable income on GDP displayed on Figure 9.3:

$$Y_d = d_o + d_1 Y \tag{8}$$

The computer calculated as the line of best fit:

$$Y_d = -58.8 + 0.74Y \tag*{(8)*}$$

So we have as the estimated intercept $d_o = -58.8$ billion and the slope $d_1 = 0.74$.

We shall find that the consumption and disposable income functions are essential for understanding the multiplier explanation of how an increase in government spending and tax cuts are said to stimulate the economy.

9.3 The simple multiplier: Depression economics

The multiplier mechanism provides an explanation, perhaps oversimplified, of how fiscal policy affects the pace of economic activity. The theory provides a useful first step toward the development of more complicated models capable of explaining much more about how the macro economy works.

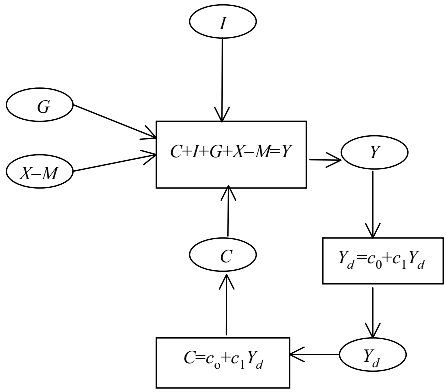


Fig. 9.4. The multiplier flowchart

9.3.1 *Rounds of induced expenditure — The Kahn multiplier*

The flowchart presented on Figure 9.4 shows how the various components of the multiplier model interact. The variables of the model are indicated by circles. The equations are contained in boxes. The arrows indicate the direction of causation. We classify the variables into *endogenous variables* determined by the model (Y, Y_d and C) and the *exogenous variables* that are taken as given for purposes of analysis (I, G, X, M). Note on the graph that each of the endogenous variables has an arrow pointing toward it from the equation that determines it. The exogenous variables, because they are not determined by the model, only have arrows flowing out of them.

Suppose that there is an increase of \$10 billion in government expenditure on goods and services. We write the change as $\Delta G = 10$, using the symbol Δ to indicate change. How will this policy affect the pace of economic activity? The immediate effect is to add \$10 billion to GDP because government spending is a component of GDP. But this is far from the end of the story, equation (8) suggests, because the \$10 billion increase in GDP will lead to an increase in disposable income: $\Delta Y_d = d_1 \times \$10$ billion $= \$7.4$ billion. This increase of $\Delta Y_d = \$7.4$ billion will, according to equation (7) lead to an increase in consumption spending: $\Delta C = 0.96 \times \$7.4$ billion $=$ \$7.1 billion. These changes, are recorded in the first column of Table 9.3. This process has involved going around the $Y \to Y_d \to C \to Y$ "feedback" loop on the flowchart. But we are still far from being at the end of the story, for this is only the first round of spending generated by the increase

Table 9.3. Rounds of induced expenditure induced by $\Delta G = 10$.

Round:	1	2	3	4	5	$\sum_{=1}^{\infty}$
ΔY	10.0	7.1	5.0	3.6	2.5 ... =	34.5
ΔY	7.4	5.3	3.7	2.7	1.9 ... =	25.5
ΔC	7.1	5.0	3.6	2.5	1.8 ... =	24.6

in G. Since C is a component of GDP, the $\Delta C = \$7.1$ billion increase from Round 1 is carried over to the top of the Round 2 column of the table to make a further contribution to GDP of $\Delta Y = \$7.1$ billion. And this second round increase in GDP will lead to added disposable income of \$5.3 billion, which will in turn generate another \$5 billion more consumption, as explained in the Round 2 column of the table. The \$5 billion of extra consumption at the end of round 2 of the feedback loop must be carried over as an addition to GDP to spark a third round of induced expenditure, and so on forever more. Each expenditure round is initiated by the rise in consumption that is carried over from the preceding period.

Only the first five of an infinite number of feedback rounds are presented on Table 9.3. That the infinite rounds of induced spending add up to a finite sum was recognized by British economist Richard F. Kahn, who published a paper on the multiplier process in 1931.[10] Kahn recognized that the effects of successive rounds was to generate a geometric series of increases in GDP. The last entry in the ΔY row of Table 9.3 claims that the sum of the infinite rounds of higher GDP adds up to \$34.5 billion. To see why this makes sense, note that for our numerical example,

$$\Delta Y = 10 + 7.1 + 5.0 + 3.6 + 2.5 + \cdots + (0.71)^n 10 + \ldots \tag{9}$$

While the series continues into the indefinite future, as long as there is no change in policy, its sum is finite. To see this, note that the sum can be written:

$$\Delta Y = [1 + 0.74 + (0.74)^2 + (0.74)^2 + \cdots + (0.74)^n + \ldots]\$10 \text{ billion}.$$

Since the expression in brackets is a geometric series, it sums to $1/(1 - 0.74) = 3.45$.[11] Therefore, the estimated change in GDP resulting

[10] Richard F. Kahn, "The relation of home investment to unemployment," *Economic Journal*, June 1931.

[11] The equation for the sum of a geometric series was derived in footnote 6 of Chapter 6. We encountered geometric series again when we discussing interest rates in Chapter 8.

from the \$10 billion increase in government spending (i.e., $\Delta G = \$10$ billion) is $\Delta Y = \$34.5$ billion. This is the number that appears at the end of the ΔY row on Table 9.3. That last column is identified by Σ, the capital Greek sigma, which is a common symbol for sum.

In general, for any value of the parameters c_1 and d_1, we have,

$$\begin{aligned}\Delta Y &= \Delta G + c_1 d_1 \Delta G + (c_1 d_1)^2 \Delta G + (c_1 d_1)^2 \Delta G + \cdots + (c_1 d_1)^n \Delta G + \ldots \\ &= [1 + c_1 d_1 + (c_1 d_1)^2 + (c_1 d_1)^2 + \cdots + (c_1 d_1)^n + \ldots] \Delta G\,. \qquad (10)\end{aligned}$$

Because the expression in brackets is a geometric series with $0 \leq c_1 d_1 < 1$, it sums to $1/(1 - c_1 d_1)$. Therefore, we have

$$\Delta Y = \left[\frac{1}{(1 - c_1 d_1)} \right] \Delta G\,. \qquad (11)$$

Since the model is linear, the increase in GDP is proportional to the increase in government spending. The expression in brackets, the factor of proportionality, is the government spending multiplier. It shows how big a multiplicative increase in GDP will be caused by an increase in government spending. That is to say, our government spending multiplier is

$$\frac{\Delta Y}{\Delta G} = \frac{1}{(1 - c_1 d_1)} \doteq 3.45\,. \qquad (12)$$

For example, a \$20 billion increase in government spending will, according to our multiplier model, generate a $3.45 \times \$20 = \$6.9$ billion increase in GDP.[12]

Tax-cut Multiplier

As an alternative to increasing government spending, how about a tax cut? Consider the simplest possible case, a lump sum tax cut given to all citizens, such as the \$300 tax cut (\$600 for joint-filers) given to all income tax payers in 2001. Or as another example, the veterans' bonus, such as that passed in 1934 to reduce the hardship of the Great Depression on those who had served their country in World War I. A tax cut taking this particularly simple form will increase d_0, the intercept of the disposable income relationship. If the total tax change is $\Delta T = -\$10$ billion, then the change in d_0 will be $\Delta d_0 = +\$10$. How this will affect the economy,

[12]The multiplier would be 25 if, as is sometimes assumed in the interest of simplification, $d_1 = 1$! That is one reason why the factors explaining the gap between GDP and disposable income are so important.

Table 9.4. Tax cut multiplier (lump sum tax cut of $10 billion).

Round:	1	2	3	4	5	$\sum_{=1}^{\infty}$
ΔYd	10.0	7.1	5.0	3.6	2.5 ... =	34.5
ΔC	9.6	6.8	4.8	3.4	2.4 ... =	33.1
ΔY	9.6	6.8	4.8	3.4	2.4 ... =	33.1

according to the multiplier model, may be visualized by returning to the multiplier flowchart on Figure 9.4. Note that the initial effect of the tax cut will be an increase in disposable income of $-\Delta T$, which will lead to a rise in consumption spending of $-c_1\Delta T$, which gets added into GDP. This rise in GDP induces further rounds of increased disposable income and consumption as summarized on Table 9.4. The key point to note is that each round of the tax-cut multiplier process is only a fraction c_1 of the corresponding round of the government spending multiplier on Table 9.3. Hence the total multiplier effect of an income tax cut is only the faction c_1 of the multiplier effect of an increase in government spending. The tax-cut multiplier is

$$\frac{\Delta Y}{\Delta d_0} = \frac{c_1}{(1 - c_1 d_1)} \doteq 3.31\,. \tag{13}$$

Tax cuts stimulate the economy, but tax hikes are a downer, serving to slow down the pace of economic activity. Because the system is linear, a decrease in d_0 generated by a lump-sum increase in personal income taxes will cause a decrease in GDP of the same magnitude but of opposite sign; i.e., the tax hike multiplier is -3.31, which is smaller than the investment multiplier.

Balanced budget multiplier

Both the government spending and the tax cut multipliers will lead to a deficit in the government's budget. While candidates for president often campaign on the promise of a balanced budget, more often than not, once elected, presidents find themselves generating budget deficits. But it is possible to put the multiplier to work without generating a deficit! Suppose that the government simultaneously increases government spending and income taxes by $10 billion. The budget deficit will not be directly affected by this change because the tax hike will serve to finance the extra government spending. But what will happen to GDP? The government spending

Table 9.5. Balanced budget multiplier (\$10 billon tax hike + \$10 billion G increase).

Round:	1	2	3	4	5		$\sum_{=1}^{\infty}$
ΔY	10.0	−2.5	−1.8	−1.3	−0.9	... =	1.4
ΔYd	−2.6	−1.8	−1.3	−0.9	−0.7	... =	−9.0
ΔC	−2.4	−1.9	−1.3	−0.9	−0.6	... =	−8.6

increase is an economic upper but the tax hike, because it reduces d_0, is a downer. Now we saw that changes in personal income taxes have a smaller multiplier effect than changes in government spending because a fraction $(1 - c_1)$ of the change in the income tax is absorbed in saving rather than contributing directly to GDP. Because our model is linear, the effect of the two medicines is additive. Thus we have

$$\Delta Y = \frac{1}{1 - c_1 d_1}\Delta G + \frac{c_1}{(1 - c_1 d_1)}\Delta d_0 \,. \tag{14}$$

But we are considering the balanced budget case where the tax hike equals the government spending increase, $-\Delta d_0 = \Delta G$. Therefore, the balanced budget multiplier is

$$\left.\frac{\Delta Y}{\Delta G}\right|_{\Delta d_0 = -\Delta G} = \frac{1 - c_1}{(1 - c_1 d_1)} \doteq 0.14 \,. \tag{15}$$

A \$10 billion increase in government spending coupled with an equal hike in taxes will increase GDP by $\Delta Y = \$1.4$. The resulting rounds are presented on Table 9.5. Since the balanced budget multiplier is so small, a president constrained by a campaign promise to balance the government's budget would have to apply a much bigger fiscal dose to obtain a given degree of economic stimulus.[13]

9.3.2 *Deriving a family of multipliers*

While calculating the rounds of induced expenditure generated by a tax cut or an increase in government spending provides insight into how the multiplier process works, a more precise understanding will be obtained by solving the system explicitly. The first step is to substitute consumption

[13]Observe that the balanced budget multiplier would be one if d_1 were equal to one. But this would be the case only if an increase did not generate additional depreciation, additional tax revenue, corporate retained earnings, etc.

function equation (7) into disposable income function (8) in order to obtain

$$C = c_o + c_1 d_o + c_1 d_1 Y \,. \tag{16}$$

Next we substitute this equation into the national income accounting identity, equation (2), to obtain

$$Y = c_o + c_1 d_o + c_1 d_1 Y + I + G + X - M \,. \tag{17}$$

Subtracting $c_1 d_1 Y$ from both sides of this equation yields:

$$Y - c_1 d_1 Y = (1 - c_1 d_1) Y = c_o + c_1 d_o + I + G + X - M \,, \tag{18}$$

or

$$Y = \left[\frac{1}{(1 - c_1 d_1)}\right] (c_o + c_1 d_o + I + G + X - M) \,. \tag{19}$$

This equation allows us to determine the endogenous variable Y on the basis of the exogenous variables of the system and the parameters of the consumption function and disposable income equation. It is an example of a *reduced form equation*, which is defined as an equation explaining a model's endogenous variable in terms of the exogenous variables and the model's parameters.

In general, a complete model will have one reduced form equation for each endogenous variable. Since our multiplier model has three endogenous variables, there are two more reduced form equations. The reduced form equation for disposable income is obtained by substituting equation (19) into disposable income equation (8):

$$Y_d = d_0 + \left[\frac{d_1}{(1 - c_1 d_1)}\right] (c_o + c_1 d_o + I + G + X - M) \,. \tag{20}$$

A reduced form equation for consumption is obtained by substituting equation (19) into (16):

$$C = c_o + c_1 d_o + \left[\frac{c_1 d_1}{(1 - c_1 d_1)}\right] (c_o + c_1 d_o + I + G + X - M) \,. \tag{21}$$

The reduced form equations allow us to determine directly a whole family of multipliers. For example, the government spending multiplier is obtained by differentiating reduced form equation (19) with respect to G:

$$\frac{\partial Y}{\partial G} = \left[\frac{1}{(1 - c_1 d_1)}\right] , \tag{22}$$

which is identical to the multiplier derived using the rounds of induced expenditure approach. Differentiating equation (19) with respect to d_0 reveals that the lump-sum tax cut multiplier is $\partial Y/\partial d_0 = [c_1/(1-c_1 d_1)]$. Similarly, the multiplier effects on disposable income and consumption can be found by differentiating reduced form equations (20) and (21). Further, an increase in private investment spending will also have a multiplier effect on the pace of economic activity, for differentiating (19) with respect to I yields the investment spending multiplier:

$$\frac{\partial Y}{\partial I} = \left[\frac{1}{(1-c_1 d_1)}\right]. \tag{23}$$

Thus an increase in private investment spending would provide the same stimulus as an increase in government spending of the same dollar magnitude.

Foreign trade

As an alternative to resolving unemployment problems by stepping up government spending or cutting taxes, governments are often tempted to provide stimulus by subsidizing exports and restricting imports. When labor and business enterprises appeal to their representatives in Washington to protect jobs by countering "unfair" foreign competition, a hike in tariffs (taxes on imports) or restrictive quotas may result. When exporters face intense competition when trying to sell their goods in foreign markets, they often appeal for subsidies from Uncle Sam. To see how such measures influence the pace of economic activity, we differentiate (19) with respect to the foreign trade balance, $X - M$, obtaining the foreign trade multiplier:

$$\frac{\partial Y}{\partial (X-M)} = \left[\frac{1}{(1-c_1 d_1)}\right]. \tag{24}$$

Clearly, increasing exports through subsidies or discouraging imports with tariffs and quotas may indeed serve to stimulate the pace of economic activity, although it will be offset if other countries, worried about their own unemployment, retaliate by restricting imports from us. Unlike government spending, which has the equivalent multiplier effect, the foreign trade remedies do not contribute to the government's budgetary problems. But as was emphasized in Chapter 2, economists have worried ever since the days of Adam Smith and David Ricardo that such measures contribute to an inefficient allocation of resources.

Tax alternatives

There are many different ways to cut taxes. Instead of a simple lump sum tax rebate or Veterans Bonus one can cut tax rates. For example, the Tax Reduction Act of 2001 reduced tax rates by 0.5% for most taxpayers for tax year 2001 and provided for further reductions in future years. The multiplier analysis of changes in the personal income tax rates is slightly more complicated because the slope parameter, d_1, rather than the intercept changes as a result of the reduction in tax rates. Once the effect on d_1 is determined, we can differentiate equation (19) to find the effect on GDP:

$$\frac{\partial Y}{\partial d_1} = \left(\frac{c_1}{(1-c_1 d_1)^2}\right)(c_0+c_1 d_0+I+G+X-M) = \left(\frac{c_1}{(1-c_1 d_1)}\right)Y\,. \quad (25)$$

Tax cuts may also focus on stimulating private investment spending. For example, in 1962 the Kennedy Administration enacted an investment tax credit allowing business firms to deduct from their tax bill a sum equal to 7% of any investment spending on plant and equipment that they had made during the year. This meant that a firm buying a \$10,000 machine saved \$700 on taxes, which served to reduce the net cost of the machine to only \$9,300. The hope was that the tax credit, by lowering the cost of investment, would induce firms to invest more. The difficult step in analyzing the effects of this measure on the national economy is to estimate the effect of this tax-subsidy on investment. Once the effect on I of the credit is determined, the change in investment must be multiplied by the investment spending multiplier, $\Delta Y/\Delta I$, to find the change in GDP. By the chain rule of differentiation,

$$\frac{\partial Y}{\partial t_{\text{credit}}} = \frac{\partial Y}{\partial I}\frac{\partial I}{\partial t_{\text{credit}}} = \left(\frac{1}{1-c_1 d_1}\right)\frac{\partial I}{\partial t_{\text{credit}}}\,, \quad (26)$$

where $\partial I/\partial t_{\text{credit}}$ is the effect of the investment tax credit on investment spending. While the investment tax credit was repealed long ago, there remains a special tax credit of 20% for investment in the preservation and reuse of certified historic buildings.

In times of recession politicians championing cuts in the capital gains tax, the dividend tax, and the corporate profits tax often argue that they will contribute to economic expansion by providing a stimulus to private investment. Of course, cuts in the personal income tax and increases in government spending are alternative techniques for stimulating the economy. The case for cutting taxes on corporations and investors rely on the assumption that that this will stimulate investment spending, that the extra investment will generate increases in worker productivity, and that greater worker productivity will contribute in the longer run to more rapid economic expansion and rising living standards.

9.3.3 *The multiplier, a graphical approach*

Although Richard Kahn focused on the geometric series generated by an increase in government spending in his 1931 article, the same analysis can be developed graphically, as shown on Figure 9.5. We have GDP (i.e., total output) plotted on the abscissa and various categories of spending on the ordinate. There are three curves on the graph. The $C(Y)$ line shows consumption as a function of GDP, as given by equation (16). The total spending curve $(C + I + G + X - M)$, the plot of (17), is obtained by adding $I + G + X - M$ onto the level of consumption as given by $C(Y)$ line. We seek the level of output at which GDP will precisely equal total spending; i.e., supply equals demand. To find it we have drawn the dashed 45^0 line emanating from the origin, which shows all points for which GDP $= C + I + G + X - M$, as dictated by the GDP identity (2). At any point to the right of the dashed 45^0 line, GDP is greater than total spending; at any

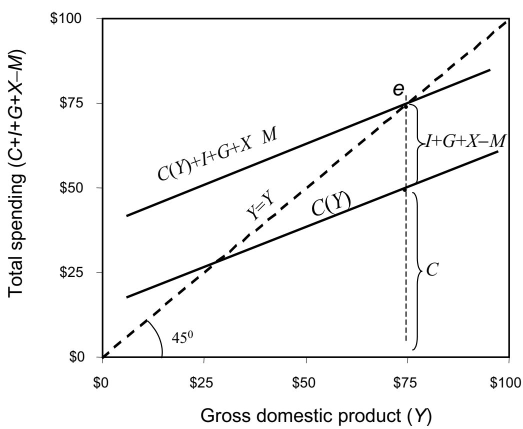


Fig. 9.5. Graphical analysis of GDP determination
The $C(Y)$ line on the graph is the plot of equation (16) relating total consumption spending to GDP. Parallel to it is the plot of equation (17) showing $C(Y)+I+G+X-M$ as a function of GDP. The dotted 45^0 line shows all points at which total spending is equal to GDP, $Y = Y$. Output is determined by equilibrium point e at the intersection of the total spending line and the $Y = Y$ line. Only at point e is total spending equal to GDP.

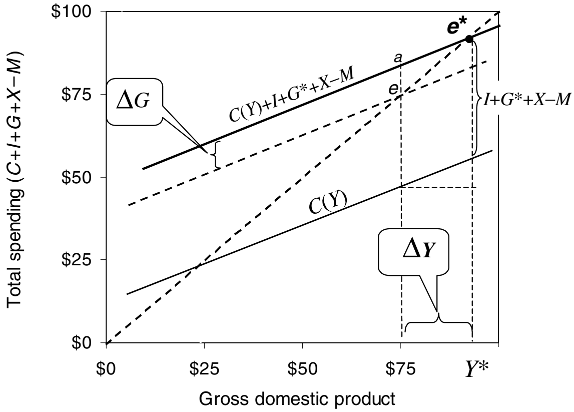


Fig. 9.6. Graphical analysis of the multiplier
When government spending on goods and services increases by ΔG to G^*, GDP increases by ΔY to Y^*. GDP increases by a multiple of the increase in investment because of the increase in consumption spending induced by the expansion in output. Self-test: where on the graph can you read off the increase in consumption spending?

point to the left of the line GDP is less than total spending. Equilibrium output is determined by point e where the total spending line intersects the 45^0 line. At this intersection, total spending, $C + G + I + X - M$, is equal to output. Consumption is read off the $C(Y)$ line.

The multiplier effect of an increase in government spending is shown on Figure 9.6. The increase in government spending has pushed up the total spending line. If GDP were to remain constant, total spending would increase only to point a. But the increase in government spending is added on to Y, which causes an increase in disposable income and hence consumption spending. The expansionary process continues as the economy slides up the new total spending line to point e^* where it crosses the 45^0 line. At this new equilibrium total spending is once again equal to GDP and, since consumption is on the $C(Y)$ line, both equations (7) and (8) are satisfied.

9.3.4 *Saving and the paradox of thrift*

Benjamin Franklin advised that "a penny saved is a penny earned." Multiplier analysis warns that just the opposite may well be the case. Suppose consumers, on average, decide to save more. Now by personal savings we mean disposable income less consumption spending, $Y_d - C$. Let us suppose that preferences between current and future consumption change so that

Table 9.6. The paradox of thrift: $\Delta c = -\$10$.

Round:	1	2	3	4	5		$\sum_{=1}^{\infty}$
ΔY	0.0	−10.0	−7.1	−5.0	−3.6	… =	−34.5
ΔYd	0.0	−7.4	−5.3	−3.7	−2.7	… =	−25.6
ΔC	−10.0	−7.1	−5.0	−3.6	−2.5	… =	−34.5
$\Delta S = \Delta Yd - \Delta C$	10.0	−0.3	−0.2	−0.1	−0.1	… =	9.0

consumers will save, collectively, \$10 billion more at whatever their disposable income turns out to be. But if saving increases, given disposable income, $C = Y_d - S_p$ must decrease. This \$10 billion increase in personal savings means that the intercept (c_o) of consumption equation (7), as plotted on Figure 9.2, is now \$10 billion smaller. Therefore, personal savings would be up by \$10 billion and consumption would be down by \$10 billion, *if* income were to remain unchanged.

What happens is shown on Table 9.6. In the first round saving is indeed up by \$10 billion. But that is not the end of the story, because the increase in saving is achieved by reducing consumption spending by \$10 billion. This means that GDP is down by \$10 billion for the next round. The result is a decline in disposable income and a further drop in consumption. As the table illustrates, the end result is a decline in GDP of \$34.5 billion. The decision to save more has had as strong a negative multiplier effect as a decline in investment of equal dollar magnitude. Disposable income has declined as a result of the drop-off in GDP, and as a result, consumption is down by \$34.5 billion and actual savings is up by only \$9 billion rather than the intended \$10 billion. Differentiating reduced form equation (19) confirms that the multiplier effect of an increase in c_o is comparable to a change in G: $dY/dc_o = 1/(1 - c_1 d_1)$. Differentiating equation (21) yields the consumption multiplier $dC/dc_0 = 1 + c_1 d_1/(1 - c_1 d_1) = 1/(1 - c_1 d_1)$; i.e., $dY/dc_o = dC/dc_0$, which makes sense since we must have $dY = dC$ because none of the other components of Y change.

All this means that if consumers decide to save more than in the past, the decline in c_0 means that the economy will slow down, unless there is an offsetting increase in investment or some other component of effective demand. Thus the primary effect of a decision to save more may be a decline in income, which is just the reverse of Ben Franklin's dictum — a penny saved is a penny *not* earned.

The surprising result that an attempt by the public to save more may lead to a decline in income, called the *Paradox of Thrift*, does not mean that saving is a bad thing. In inflationary times, when the economy is pushing hard against capacity, a decision by the public to save more may lead to a reduction in inflationary pressure rather than a reduction in output. For this reason, government leaders have at times implored the public to help fight inflation by saving more. The moral of the Paradox of Thrift is that in times of widespread unemployment, consuming a larger proportion of disposable income might help to get the economy rolling again, and conceivably the resulting rise in income might lead to more saving. In more normal times, it may be hoped that an increased desire to save will release resources that will allow greater investment spending and contribute to a more productive future.

9.4 A simple model of the business cycle

The multiplier model purports to explain what determines the level of economic activity at any point of time, but it does not explain why the economy fluctuates. We will consider what may well be the simplest macroeconomic model capable of generating a business cycle. This model shows that it is *not* necessary to advance separate explanations of the upper and lower turning points in order to account for cyclical phenomena. And it is not necessary to invoke a devil's theory explaining the occurrences of economic downturns as being the consequence of unfortunate errors made in Washington by the monetary or the fiscal authorities. The model, related to one developed near the end of the Great Depression by Alvin Hansen and Paul Samuelson,[14] is outdated but simple. It neglects depreciation, foreign trade, supply-side shocks, monetary policy and a host of other complications that can be considered with more elaborate explanations of cyclical fluctuations.

Assumptions

Consumption is determined by last period's income:

$$C_t = c_0 + c_1 Y_{t-1}\,. \tag{27}$$

The capital stock (machinery, equipment and factories) that entrepreneurs desire to process is proportional to output.

[14] Paul A. Samuelson, "Interactions Between the Multiplier Analysis and the Principle of Acceleration," *Review of Economic Statistics*, May 1939.

$$K_t^d = kY_{t-1}\,. \tag{28}$$

Net Investment, by definition, is

$$I_t = K_t - K_{t-1}\,. \tag{29}$$

Firms always undertake just enough net investment to keep their capital stock at the desired level:

$$I_t = K_t^d - K_{t-1}\,. \tag{30}$$

Therefore, $K_t^d = K_t = kY_{t-1}$ and $K_{t-1}^d = K_{t-1} = kY_{t-2}$, so we have a simple "accelerator" explanation of investment:

$$I_t = k(Y_{t-1} - Y_{t-2})\,. \tag{31}$$

This is how investment must behave in order that the capital stock will remain proportional to output. Finally, since $Y_t = C_t + I_t + G_t$, substitution yields a second order linear difference equation explaining the economy's law of motion:

$$Y_t = c_0 + (c_1 + k)Y_{t-1} - kY_{t-2} + G_t\,. \tag{32}$$

Dynamics

The simulation on Figure 9.7 illustrates the type of behavior generated by this system for parameter values $c_0 = 0$, $c_1 = 0.5$, and $k = 0.8$. Initially, the system is in equilibrium with $G = 50$, $Y = 100$, $C = 50$ and $I = 0$. But in period 5 government spending steps up to $G = 100$ and remains at this higher level forever more. The graph reveals the persistent cycle that results! The cycle is damped and Y approaches \$200 billion in the limit. What happened is best understood by considering the data on Table 9.7. In the simple multiplier model, investment spending was treated as exogenous. The multiplier does correctly predict the limiting value of GDP generated by our new endogenous investment spending model. The multiplier accelerator model generates a cycle because the expansion in government spending and subsequent growth in output sparks a temporary spurt of investment spending in order to adjust the capital stock to the new desired level. The temporary spurt in investment causes GDP to overshoot its equilibrium level. But once the capital stock catches up with the expanding GDP, investment drops off and so must GDP. Once GDP starts to decline we have a period of negative investment because business enterprises are liquidating

Table 9.7. Business cycle simulation #1: Convergent oscillations.

Period	Parameters: c_0, $c_1 = 0.5$ and $k = 0.8$ Government Gt	Consumption $C = c_0 + c_1 Y_{-1}$	Investment $I = K \quad - K_{-1}$	GDP $Y = C + I + G$	Capital $K \ = K_{-1} + I$	Desired capital $K \quad = kY_{-1}$
0	50.0	50.0	0.0	100.0	80.0	
1	50.0	50.0	0.0	100.0	80.0	80.0
2	50.0	50.0	0.0	100.0	80.0	80.0
3	50.0	50.0	0.0	100.0	80.0	80.0
4	50.0	50.0	0.0	100.0	80.0	80.0
5	100.0	50.0	0.0	150.0	80.0	80.0
6	100.0	75.0	40.0	215.0	120.0	120.0
7	100.0	107.5	52.0	259.5	172.0	172.0
8	100.0	129.8	35.6	265.4	207.6	207.6
9	100.0	132.7	4.7	237.4	212.3	212.3
10	100.0	118.7	−22.4	196.3	189.9	189.9
11	100.0	98.1	−32.9	165.3	157.0	157.0
12	100.0	82.6	−24.8	157.8	132.2	132.2
13	100.0	78.9	−6.0	173.0	126.3	126.3
14	100.0	86.5	12.1	198.6	138.4	138.4
15	100.0	99.3	20.5	219.8	158.9	158.9
16	100.0	109.9	17.0	226.9	175.8	175.8
17	100.0	113.4	5.7	219.1	181.5	181.5
18	100.0	109.5	−6.2	203.3	175.3	175.3
19	100.0	101.7	−12.6	189.1	162.7	162.7
20	100.0	94.5	−11.4	183.1	151.2	151.2

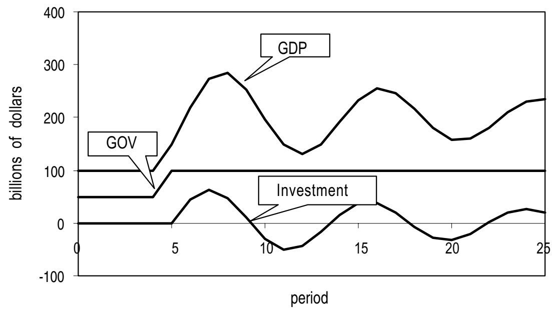


Fig. 9.7. Business cycle iterations: convergent cycles
A cycle is generated when the multiplier-accelerator model is disturbed by a step increase in government spending from 50 to 100. The system is gradually converging to the new equilibrium level predicted by the multiplier.

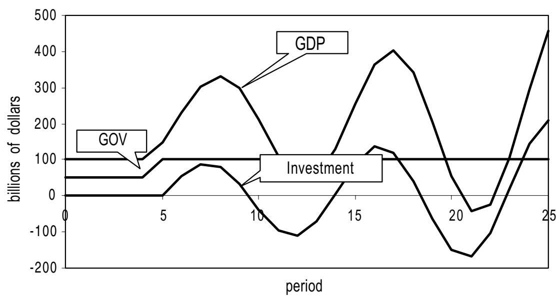


Fig. 9.8. Business cycle iterations — divergent cycles
An increase in the desired capital/output ratio from $k = 0.8$ to $k = 1.1$ unleashes a divergent cycle when the system is disturbed by an increase in government spending.

their excessive capital stock, but this negative investment leads to a further drop in GDP. Eventually the recession comes to an end and the cycle repeats.

The second simulation, reported on Figure 9.8, shows that the model can also generate divergent oscillations. Monotonic convergence to equilibrium is also possible, as with $c_1 = 0.9$ and $k = 1.1$.[15] As these simulations make clear, the type of business cycle developed by this model depends not

[15] The model's parameters determine the type of motion. It can be shown, by analyzing the roots of the characteristic equation of (32), that stability requires $k < 1$. For a more detailed discussion of this model consult Alpha Chiang, *Fundamental Methods of Mathematical Economics*, Ch 17.2.

on the nature of the shock but on the underlying parameters of the model. And the way to understand how the cycles come about is not to look at causes for each specific downturn and upturn when they are encountered. To understand the cycle we must look at the basic structure of the system.

9.5 Critique and qualifications — Shrinking multipliers

The multiplier theory that we have been discussing was developed just as the world was entering the Great Depression of the 1930s. In the intervening years the limitations of the theory have become more obvious. And the availability of much more accurate data and computers to process the evidence has led to substantial advances in what we know about how the economy functions. In subsequent chapters we will look at refinements of the theory and at alternative modes of analysis. Here we must look at some of the major limitations of the analysis.

9.5.1 *Modifications of the consumption function*

Wealth in the consumption function

A major problem with multiplier analysis is that the empirical evidence reveals that it is based on much too simple a theory of consumer behavior. One rather simple elaboration of the consumption function, equation (7), is to allow for the effect of wealth as well as income on consumption spending. The larger the stock of wealth that people have accumulated over the years in their saving accounts or by investing in real estate or the stock market, the more they are likely to consume, given their current income. Thus we might modify equation (7) to read:

$$C = c_0 + c_1 Y_d + c_2 W \,. \tag{33}$$

This more elaborate consumption function implies that a collapse of the stock market, by reducing the wealth of consumers, will lead them to consume less even if their income is unchanged. Indeed, Professor Albert Ando of the University of Pennsylvania argued that most of the Japanese recession of the 1990s may be due to the effect of declining stock market prices on household wealth and consumption.[16]

When empirical investigators include wealth in the consumption function, they generally find that the marginal propensity to consume, c_1, is

[16] "On the Japanese economy and Japanese national accounts," NBER working paper number w8033, December, 2000.

quite a bit smaller than the estimate of $c_1 = 0.96$ suggested by the consumption on disposable income scatter-plot of Figure 9.2. As can be seen from (22), a smaller c_1 means that the multiplier effect of an increase in government spending will also be much smaller.

Permanent versus transient income changes

In the 1950s Nobel Laureates Franco Modigliani and Milton Friedman independently developed much more elaborate theories of consumption. While there were major distinctions between their two theories, a common implication was that consumers would not adjust their consumption spending very much in response to changes in income that are thought to be temporary. For example, if the government cuts tax rates in an effort to combat recession, there will be very little stimulus if the public correctly perceives that the tax cuts are not likely to be permanent.

In their life cycle theory of the consumption function, Franco Modigliani and his coauthor Brumberg argued that in deciding how much to spend at any point of time the representative consumer considers the likely flow of future income and their changing financial needs over their remaining lifespan rather than just focusing on their current situation. When consumers receive an increase in disposable income that is perceived to be only temporary, they are inclined to put much of it aside to consume gradually in future years rather than splurging on consumer goods in the year in which they received the windfall gain in disposable income.

Milton Friedman reached much the same conclusion in developing his permanent income hypothesis, arguing that transitory income changes, such as that provided by a temporary tax cut, was not a decisive determinant of consumption. What matters are changes in permanent income — the average level of income expected over the years.

Both these theories imply that the marginal propensity to consume is quite small for changes in disposable income that are viewed as temporary, which means that the resulting multiplier effect of changes in government spending will be small as well. This suggests that a temporary tax cut designed to push an economy out of recession is unlikely to have much impact. However, there is an important qualification to this conclusion. It is possible that the recipients of the temporary tax cut may decide not to save their tax cut for the future by placing their funds in a savings account or purchasing financial assets. Instead, they may provide for future as well as present needs by investing the tax cut in durable consumer goods, such as a new car, or by using it to help finance the purchase of a new house. When consumers use their tax cut to finance investment in consumer durables it provides an obvious stimulus to the economy.

Ricardian equivalence and the government debt

The argument of Modigliani and Friedman was taken to its logical conclusion by Robert Barro, who advanced a concept known as Ricardian equivalence, which he traced back to David Ricardo's *Principles of Political Economy and Taxation*, 1817. Barro argued that if the government borrows more this year to make up for the loss of revenue from cutting taxes, the public will recognize that in future years they will have to shoulder higher taxes in order to pay off the resulting debt. In terms of the two-period model discussed in Chapter 4.4.5, a tax cut this period coupled with a repayment tax hike in the future has not shifted the budget line on Figure 4.22 at all. Therefore, according to Barro, the public will put aside as additional saving their increase in disposable income generated by the tax cut in order that they will be prepared to meet their increased future tax liability. By responding in this way, the public will have exactly the same consumption opportunities as would have been available if the government had not financed the tax cut by borrowing — it is this equivalence of opportunities, with or without the tax cut, which explains why this proposition is called Ricardian *equivalence.*

The argument still holds, argues Barro, even if the debt is not paid off during the taxpayer's lifespan. The taxpayers, recognizing that their heirs will have to pay additional taxes in the future as a result of the government's borrowing, will put aside the tax cut so that their heirs will be in a position to pay the taxes required to retire the government debt generated by the tax cut when it matures. By behaving in this way, the taxpayers leave both themselves and their heirs in exactly the same position as they were in before the tax cut was enacted. The result is to nullify the effect of the tax cut by allowing both the recipient of the tax cut and the heirs to enjoy precisely the same consumption spending that they would have in the absence of the tax cut. If Barro is right, the tax cut, because it is precisely offset by extra saving of the current generation, will not stimulate the economy at all!

Barro's argument has been hotly contested. One objection is that in reality many consumers are not nearly as far-sighted as Barro assumed, some enjoying the tax cut bonus without worrying about future taxes. A second objection is that credit restrictions limit the ability of the public to borrow as much as they would like to in order to shift more of their lifetime consumption to the present. The tax cut facilitates such a shift by enabling the credit constrained consumer to substitute more present for less future consumption.

9.5.2 *Capacity constraints and the inflationary gap*

The multiplier can give the economy a push, but what happens when a massive increase in government expenditure, as in a major war, pushes the economy beyond full employment into the capacity ceiling? Suppose that Y^c denotes the capacity level of output, possibly estimated by Okun's Law, and Y the level of GDP determined by the multiplier analysis as summarized by reduced form equation (19). Then if $Y > Y^c$ the full force of the multiplier obviously cannot be realized. Actual output, constrained by Y^c, will fall short of the level predicted by the multiplier. The result will be inflation. This particular type of inflation is sometimes called *demand-pull inflation.*

When the economy is straining its capacity, the excess of effective demand over capacity, $Y^e - Y^c$, is sometimes called the *inflationary gap.* Some economists have argued that the demand-pull inflation will be proportional to the inflationary gap. This mode of analysis suggests that when an economy is operating at full employment, increases in government spending or tax cuts will be translated into inflation rather than into additions to real output. But the theory of the inflationary gap is not a complete theory of inflation. It does not explain why inflation has at times taken place in periods of rather substantial unemployment. Gap theory does not take into account the role of the central bank in controlling inflationary pressures. These complications will be carefully investigated in subsequent chapters.

9.5.3 *Missing links*

One problem with multiplier analysis is that it is incomplete. The analysis assumes that private investment spending and the foreign trade balance will be unaffected by a change in government spending or tax policy — both I and $X - M$ are treated as exogenous. Taken at face value, the multiplier argument suggests that the Federal Reserve System — the central bank of the United States — cannot affect the pace of economic activity. And multiplier analysis, in and of itself, even when supplemented with the concept of the inflationary gap, provides only a very incomplete explanation of inflation. These complications will be explored in the next chapter.

Table 9.8. Multiplier estimates.

Simulation	Response at end of: 1st year	2nd year	3rd year	10th year
#1 Increase in Federal government spending on goods and services equal to 1% of GDP				
GDP	1.4	1.4	1.1	1.1
Unemployment rate	−0.5	−0.7	−0.7	−0.6
Consumer price inflation	0.1	0.5	0.7	
#2 Permanent increase in personal income tax equal to 1% of GDP				
GDP	−0.4	−0.8	−1	−1.5
Unemployment rate	0.1	0.3	0.4	0.7
Consumer price inflation	0	−0.1	−0.3	−1.2

Source: David Reifschneider *et al.*, "Aggregate disturbances, monetary policy, and the macro-economy: The FEB/US perspective," *Federal Reserve Bulletin*, January, 1999

9.5.4 *Multiplier Estimates*

Economists are not allowed to experiment on the economy, no more than astronomers are empowered to experiment by moving the planets. But we can perform experiments on artificial computer models of the economy. It is easy with a simple spreadsheet program, like Excel or Lotus, to generate a business cycle. That is how we generated the artificial business cycle data that was presented on Figure 9.7. But the modern computer allows macro-econometricians to do much more elaborate simulations on much more sophisticated models. Table 9.8 summarizes the results of one experiment conducted by researchers at the Federal Reserve Board in Washington using a 350 equation model of the U.S. economy.[17] Unlike the model used to generate the data for Figure 9.7, the parameters of their model have been carefully estimated from historical data rather than pulled out of the air. We will be looking in more detail at this model in Chapter 11.5.3. But now we shall take a quick look at the multiplier effects summarized on Table 9.8.

The simulation experiment looks at how the economy responds over time when disturbed by a permanent increase in government spending on goods and services equal to 1% of GDP, or about \$92.24 billion since GDP was \$9224 billion in 2000. The first year result would be a 1.4% increase in GDP, or a gain of \$129.1 billion dollars. Thus the multiplier is estimated

[17]David Reifshneider, Robert Tetlow and John Williams in "Aggregate disturbances, monetary policy, and the macroeconomy: the FRB/US Perspective," *Federal Reserve Bulletin*, January, 1999.

to be 1.4. Specifically, we have

$$\text{multiplier} = \frac{\Delta Y}{\Delta G} = \frac{1.4\% \text{ of GDP}}{1\% \text{ of GDP}} = 1.4\,. \tag{34}$$

The model also predicts that during the first year of the experiment unemployment would decline by 1/2 percent and consumer prices would rise by 1/10 of a percent. The response in the 2nd year 3rd year and 10th year of the experiment are reported in successive columns of the table.

The government spending multiplier reported on Table 9.8 is much smaller than was suggested by the simple multiplier model of Section 9.3. The estimated multipliers are smaller in part because of the use of the more elaborate consumption function relationship and in part because of a host of other complications built into the 350 equation model, some of which we will be exploring in Chapters 10 and 11. We shall find in Chapter 11 that the strength of the multiplier effect of an increase in government spending depends critically upon the response of the central bank. Further, the multiplier effect is likely to be much stronger when there is substantial excess productive capacity. In times of severe recession an increase in government spending can be expected to have a substantial multiplier effect on output because there is excess capacity to draw upon before inflationary forces will become manifest. In times of full or near full employment, the primary effect of expansionary fiscal policy is likely to be inflationary, as explained by the concept of the inflationary gap.

Summary

1. A nation's Gross Domestic Product (GDP) is the market value of all the goods and services produced by its economy during a year or quarter. The real GDP is corrected for price changes so as to facilitate comparisons over time.
2. The national income accounting identity partitions the GDP into its basic components: $Y \equiv C + I + G + X - M$. The investment ($I$) component of GDP does not refer to financial investment, such as the purchase of shares of stock in a corporation. I is that part of current output that is designed to provide for the future, such as the production of tools and machinery, factory buildings and houses, and highways and trucks. Government spending on goods and services (G) does not include Social Security benefits and other transfer payments.

3. Disposable income (Y_d), the income people have to spend after taxes, is usually about 70% of GDP. The gap between GDP and disposable income is explained by deductions for capital consumption allowances (depreciation), for taxes, and for corporate retained earnings. These deductions are partially offset by transfer payments and government interest payments. Personal saving equals disposable income less consumption and consumer interest payments.
4. Consumption spending is largely explained by disposable income: $C = c_0 + c_1 Y_d$. Disposable income in turn is a function of GDP: $Y = d_0 + d_1 Y$. Armed with these two behavioral equations and the national income accounting equation, we derived the reduced form equation for GDP:

$$Y = [1/(1 - c_1 d_1)](c_o + c_1 d_o + I + G + X - M)\,. \tag{19}$$

Differentiation yields the government spending multiplier $\partial Y/\partial G = 1/(1-c_1 d_1)$. With estimated parameter values $c_1 = 0.96$ and $d_1 = 0.74$, the multiplier is about 3.45, implying that a $10 billion dollar increase in government spending will lead to a $34.5 billion expansion in GDP. The multiplier can also be viewed as the sum of the rounds of consumption that are induced when the government increases spending on goods and services.
5. The multiplier effect may also be harnessed without increasing government spending. For example, a $10 billion lump-sum cut in personal income taxes provides a weaker stimulus to GDP because $\partial Y/\partial d_0 = c_1/(1 - c_1 d_1) = c_1 \partial Y/\partial G \doteq 3.31$. The balanced budget multiplier states that an increase in government spending financed by a lump sum personal income tax hike has a multiplier effect on GDP equal to the difference $\partial Y/\partial G - \partial Y/\partial d_0 \doteq 0.14$. If government subsidies stimulate exports or a protective tariff discourages imports, GDP will increase by the multiplier because $\partial Y/\partial X = -\partial Y/\partial M = \partial Y/\partial G$. Tax cuts may also be aimed at stimulating business investment spending.
6. When recession threatens, governments often try to harness the multiplier effect in an attempt to stimulate the economy, sometimes with tax cuts and sometimes with increases in government spending, and often with a mixture of the two. Both Republican and Democratic presidents have tried to put the multiplier process to work. Sometimes the expenditure is for national defense and sometimes for massive public works, such as the construction of major hydroelectric projects and interstate highways. But over the years academic economists have become more and more skeptical about the efficacy of fiscal policy.

7. Several refinements of the basic consumption function, equation (7), have been advanced. It seems reasonable to include wealth as well as income in the consumption function, which makes the estimated value of c_1 smaller and results in a substantially weaker multiplier. The Modigliani-Brumberg life-cycle hypothesis, based on the assumption that consumers plan their spending over their entire remaining life span on the basis of anticipated needs and expected future income, also makes consumption much less sensitive to changes in current income, implying a smaller multiplier. And Robert Barro's Ricardian Equivalence proposition holds that tax cuts fail to stimulate the economy because consumers are inclined to put aside the tax cut in order to be in a position in the future to pay the additional taxes that will be required when the government pays off its debt. While there are major differences between these various consumption theories, all imply that the multiplier is much smaller than the simple consumption function suggests.
8. The fiscal stimulus generated by the federal government will be weakened if budgetary constraints lead state and local governments to cut back on spending. And the multiplier effect will not be realized to the extent that the increased government spending is offset by a reduction in private investment spending. If a protective tariff designed to stimulate the consumption of domestically produced goods by discouraging imports prompts other countries to retaliate by enacting protective tariffs, a reduction in our exports may nullify the expansionary effects of our protective tariff.
9. The simplest possible model of the business cycle, developed by Professor Paul Samuelson in the 1930s, augments the multiplier model by making investment spending endogenous. Prior to the development of this model, business cycle theory had focused on trying to explain why the upper and lower turning points of the cycle occurred. In contrast, the multiplier model does not focus on the turning points in the cycle but instead shows how the structure of economy may have a natural tendency to generate cyclical fluctuations. The simple model may generate steady grown, stable cycles or explosive cyclical movements — everything depends upon the model's parameters.
10. A 350 equation computer model of the United States developed by economists at the Federal Reserve Board suggests that the government spending multiplier is about 1.4.

Key Concepts

Exercises

1. The foreign trade multiplier concept recognizes the fact that imports tend to climb when Y increases. Suppose that in Simple Land,

$$Y = C + I + G + X - M \tag{1}$$

$$C = c_0 + c_1 Y_d \tag{2}$$

$$M = m_0 + m_1 Y \tag{3}$$

$$Y_d = d_0 + d_1 Y\,. \tag{4}$$

   Here Y is GDP, C is consumption, I investment, G government spending on goods and services, X exports, M imports and Y_d is disposable income. Note that the addition of equation (3) has converted M into an endogenous variable.

a. Modify Figure 9.4 in order to take into account the addition of equation (3).
b. Derive the reduced form equation explaining Y in terms of the exogenous variables of the model (I, G, X).
c. Determine the equation for the government spending multiplier for this open economy.
d. Consider a simultaneous \$10 billion increase in government spending on goods and services coupled with a \$10 billion lump sum tax increase. Derive the equation showing the effect of this balanced budget multiplier policy on the level of Y.

2. *A Simple Macroeconomic Model*: Suppose that in Simple Land the GDP Identity holds:

$$Y \equiv C + I + G + X - M\,. \tag{5}$$

Also, disposable income is

$$Y_d = Y - T - R - D + T_r\,, \tag{6}$$

where T is tax revenue, D is depreciation, T_r is transfer payments and R is corporate retained earnings. Suppose corporate retained earnings are 5% of GDP, transfer payments are 5, depreciation is 10% of GDP and taxes are

$$T = -25 + .25Y\,. \tag{7}$$

Further, consumption is

$$C = 10 + 2/3\,Y_d\,. \tag{8}$$

a. Fill in the blanks on Table 9.9.

Table 9.9. Consumption and GDP in Simple Land.

Y	T	R	D	Yd	C
500	100	25	______	______	230
1,000	225	______	100	630	430
2500	______	______	______	______	______

Note: All figures are in billions of dollars, year 2000.

b. Draw a neat graph showing how disposable income and consumption spending are influenced by Y.
Hint: Plot Y on the abscissa and both Y_d and C on the ordinate.
c. Determine from the above information the parameters (coefficients c_0, c_1, d_0, and d_1) of the following equations:

$$C = c_0 + c_1 Y \qquad \text{(consumption as a function of GDP)}$$

$$Y_d = d_0 + d_1 Y \qquad \text{(disposable income function)}$$

Definition: dC/dY_d is the marginal propensity to consume. dC/dY is the "Marginal Propensity to Consume out of GDP".
e. What factors account for the difference between dC/dY_d, and dC/dY in this model?
f. Calculate dY/dG.

3.* Consider the artificial business cycle data on Table 9.7. Use a spreadsheet, such as Excel, on a personal computer to generate a new table with $k = 1.1$ instead of 0.8.

a. Does your spreadsheet generate a cycle similar to that displayed on Figure 9.8?
b. Now change the parameter k to 1.0. What type of cycle does your spreadsheet generate?

4.# Solve second order difference equation (32) for a function explaining GDP in terms of time t, the level of government spending G, and the initial level of GDP. Determine the values of the parameters for which the system will be stable. Are there parameter values that yield monotonic convergence to equilibrium?

10

Money, Prices and Output

10.1 Overview

The multiplier model we have been studying is only a single first step toward the objective of understanding how the macro economy works. It concentrates on the relationship between government tax and spending policies and the level of economic activity. But it does not provide an adequate explanation of inflation. It neglects the role of interest rates and the money supply. It does not tell us how a nation's central bank can influence the pace of economic activity, foreign exchange rates and inflation. This chapter will examine these important macroeconomic issues.

Fortunately, it will not be necessary to start afresh. Instead, we will be elaborating upon the multiplier model as a starting point in constructing a sequence of models of increasing complexity that will explain a wider range of economic phenomenon. Table 10.1 provides a schematic picture summarizing the models that will be developed in this chapter.

10.2 Explaining investment and the IS curve

Our first task is to develop the IS model, which elaborates on the multiplier mechanism by explaining investment behavior in terms of the rate of interest. After reviewing the essential features of the multiplier model inherited from Chapter 9, we will augment that model by adding an equation explaining how investment spending depends upon the rate of interest.

10.2.1 *Review of the multiplier (Model A)*

To review the analysis of Chapter 9.3.2, the simple multiplier model explains three endogenous variables — real consumption C, disposable income Y_d and output Y — with three familiar equations:

$$C = c_o + c_1 Y_d \qquad \text{(Consumption Function)} \qquad (1)$$

$$Y_d = d_o + d_1 Y \qquad \text{(Disposable Income Function)} \qquad (2)$$

$$Y \equiv C + I + G + X - M \qquad \text{(GDP identity)} \qquad (3)$$

The exogenous variables are private investment spending I, government spending G, and the excess of exports over imports $(X - M)$.

Substitution of equation (1) and (2) into (3) leads to

$$Y = c_o + c_1 d_o + c_1 d_1 Y + I + G + X - M\,, \qquad (4)$$

which simplifies, after subtracting $c_1 d_1 Y$ from both sides as in Chapter 9.3.2, to the following reduced form equation:

$$Y = \frac{1}{1 - c_1 d_1}(c_o + c_1 d_o + I + G + X - M) \qquad (5)$$

with government spending multiplier

$$\frac{\partial Y}{\partial G} = \frac{1}{1 - c_1 d_1}\,. \qquad (6)$$

Table 10.1. A nested sequence of models: Money, prices and output.

Model	Endogenous	Exogenous	Comments
Model A: The Multiplier $C = c_o + c_1 Y_d$ (Consumption function) $Y_d = d_o + d_1 Y$ (Disposable income function) $Y \equiv C + I + G + X - M$ (GDP identity)	C, Y_d, Y	G, I, X, M	$\partial Y/\partial G = \partial Y/\partial I = \partial Y/\partial X = 1/(1 - c_1 d_1)$ $\partial Y/\partial M = -\partial Y/\partial G$ $\partial Y/\partial d_0 = c_1 \partial Y/\partial G$ (tax cut multiplier)
Model B: The IS Curve Above + $I(i), \partial I/\partial i < 0$ (Investment function)	Above $+ I$	Above $- I + i$ $= G, X, M, i$	The rate of interest is a new policy variable.
Model C: The IS-LM Model Above + $M_1^r = L(i, Y)$ (Demand for money equation)	Above $+ i$	Above $- i$ $= G, X, M, M_1^r$	The interest rate is now endogenous. M_1^r, the (real) money supply, is a new exogenous variable.
Model D: Aggregate Demand Curve Above + $M_1^r = M_1/p$	Above $+ M_1^r$	Above $- M_1^r$ $+ p, M_1$	The central bank determines the nominal money supply (M_1) but not its real value (M_1^r)
Model E: Aggregate Demand & Supply (Short Run "Keynesian Case") Above + $w^r = w^m/p = \partial Q(L_d, K)/\partial L$ (Labor's marginal product = real wage) $L_d(w^r, K)$ (Labor demand; i.e., employment) $Y_s = Q(L_d, K)$ (Aggregate supply)	Above + p, w^r, Y_s	Above $- p$ $+ w^m, K$	Money wages are rigid downward, which may prevent full employment. Assuming profit maximization yields labor demand, given production function $Q(L, K)$, the real wage and the capital stock.
Model F: Aggregate Demand & Supply (Long Run "Classical Case") Above + $L_s(w^r)$ (Supply of labor depends on real wage) $L_s(w^r) = L_d(w^r)$ (w^r adjusts to preserve full-employment) $Q(L_s(w^r), K)$ (Full employment output) $Q(L_s(w^r), K) = Y_d(G, M_1/p)$ (Output always at capacity!)	Above + $w^n, L_s L_d$	Above $- w^m$ $= M_1, G, X, M$	Because money wages are assumed to be flexible, real wages adjust to achieve full employment. Effective demand equals full employment output. Shifts in aggregate demand curve due to monetary and fiscal policy determine the price level but not the level of output or the real money supply.

Addendum: Dynamic complications include $i^r = i^n - \dot{p}$ and $dK/dt = I$.
Note: M denotes imports, M^r the real money supply, and M_1 the nominal money supply.

10.2.2 *The IS relationship (Model B)*

The "IS Curve" elaborates on the multiplier by allowing interest rates to affect investment and output. For several reasons it is reasonable to suppose that investment expenditure is negatively related to the real rate of interest. Higher interest rates discourage potential homebuyers because an increase in the cost of borrowing means that a larger mortgage payment will have to be paid every month. Potential home buyers, no longer able to afford the new house of their dreams, must either scale back the size of their house or put off buying altogether — either way, investment spending is reduced. Also, manufacturing firms that had been planning to borrow funds to pay for the expansion of production capacity or to modernize existing buildings and equipment may decide not to invest if there is a rise in the cost of borrowing from their bankers and other sources. Those firms that had been planning to finance their investment internally with retained earnings will have second thoughts now that they can receive a higher interest rate by loaning out their surplus finds. For all these reasons, it seems reasonable to make investment depend inversely on the rate of interest.

What interest rate should we use in the investment function? The interest rate that has to be paid on a loan depends on a host of factors, including the credit worthiness of the borrower, the length of the loan and the general conditions of financial markets. Figure 10.1 reports on the rate of interest

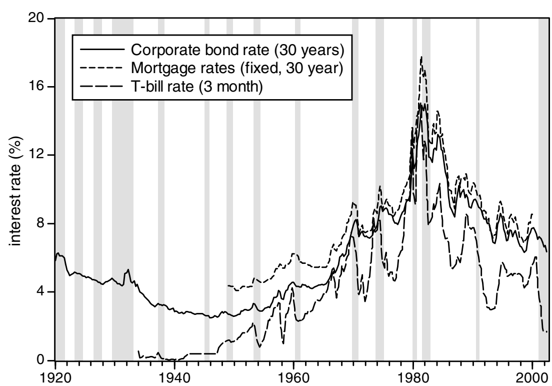


Fig. 10.1. Interest rate history
The 30 year Corporate bond rate is the rate of interest that corporations with excellent credit records pay on funds borrowed for 30 years. The mortgage rate is paid by new homeowners who borrow funds for 30 years. The United States Treasury pays the T-bill rate on funds that the government borrows for a 3 month period.

paid by the most credit worthy of corporations borrowing funds for thirty years, the rate of interest paid by homebuyers borrowing funds to finance the purchase of new homes, and the interest rate paid on funds borrowed for 90 days by the United States Treasury. The T-bill rate paid by the government is lower in part because government loans are more secure and in part because the loan is short term, which also reduces risk. The data on the graph are consistent with the general principle that short-term rates tend to be more volatile than long-term rates. Nevertheless, the rates do tend to move more or less together, suggesting that as a first approximation we may be permitted to talk about "the rate of interest" in discussing the determinants of investment spending

To transform the multiplier into the IS relationship, we add a new endogenous variable i to represent the rate of interest and convert investment I into an endogenous variable by adding an investment equation:

$$I(i)\,, \quad \frac{dI}{di} < 0\,. \tag{7}$$

We specify $dI/di < 0$ because investment spending becomes more costly when the interest rate that has to be paid to finance it is high. To keep things simple, let us assume that the relation is linear:

$$I(i) = k_o - k_1 i\,. \tag*{(7)*}$$

This new model is summarized by the flowchart on Figure 10.2, which was obtained by making some minor adjustments in the multiplier flowchart on Figure 9.4.

To solve our new model we substitute equation (7) into reduced form equation (5) for the multiplier model in order to obtain the IS function

$$\begin{aligned} Y(G,i) &= \frac{1}{1-c_1 d_1}[c_0 + c_1 d_o + I(i) + G + X - M] \\ &= \frac{1}{1-c_1 d_1}[c_0 + c_1 d_o + k_o - k_1 i + G + X - M]\,. \end{aligned} \tag{8}$$

Differentiation with respect to the interest rate yields

$$\frac{\partial Y}{\partial i} = \frac{1}{1-c_1 d_1}\frac{dI}{di} = \frac{-k_1}{1-c_1 d_1} < 0\,. \tag{9}$$

That is to say, $\partial Y/\partial i$ must be negative because $dI/di = -k_1 < 0$. Note

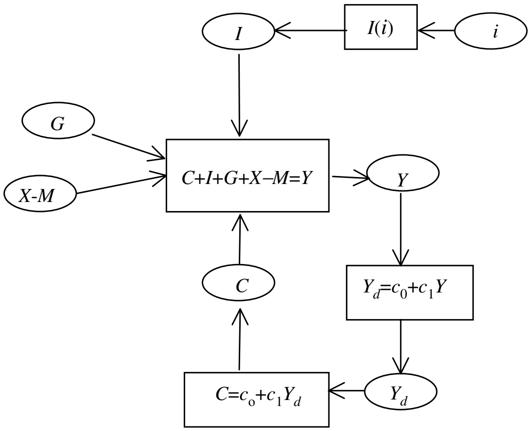


Fig. 10.2. The IS flowchart (Model B)
Investment is converted into an endogenous variable by adding a new equation, $I(i)$, where i represents the interest rate that investors in plant and equipment, housing and so forth must pay on borrowed funds.

that the government spending multiplier is

$$\frac{\partial Y}{\partial G} = \frac{1}{1 - c_1 d_1},$$

which is identical to equation (6) of the simple multiplier model.

10.2.3 *Plotting the IS curve*

The IS mechanism is displayed on Figure 10.3. On the left is the inverse relationship between investment and the rate of interest, $I(i)$. An increase in the rate of interest from 10% to 20% leads to a reduction in investment from $I(10\%)$ to $I(20\%)$. The IS curve on the right-hand panel shows how the reduction in investment, thanks to the multiplier, causes GDP to drop from 230 to 180, which gives us two points on the IS curve: $(230, 10\%)$ and $(180, 20\%)$. Thus the IS curve shows how changes in the rate of interest affect the pace of economic activity, given the sum of the exogenous spending variables: $G + X - M$.

The nature of the IS relationship is clarified by looking at the derivation of the specific numerical example underlying Figure 10.3. We have three structural equations plus the GDP accounting identity to explain four endogenous variables:

$$C = \left(\frac{3}{4}\right) Y_d \tag{10}$$

$$Y_d = \left(\frac{2}{3}\right) Y \tag{11}$$

$$I(i) = 100 - 250i \tag{12}$$

$$Y \equiv C + I(i) + G + X - M\,. \tag{13}$$

First we obtain $C = (1/2)Y$, the relationship between consumption spending and GDP, by substituting (10) into (11). Substituting this last equation and the investment equation into GDP identify (13) eventually yields the reduced form equation:

$$Y = \frac{1}{1 - 1/2}[100 - 250i + G + X - M]\,. \tag{14}$$

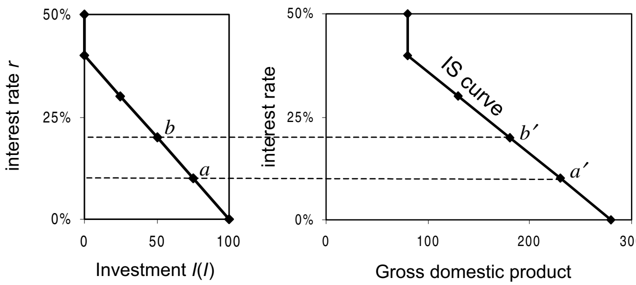


Fig. 10.3. The IS curve (Model B)
The rate of interest determines the level of investment. Given the interest rate, the resulting level of investment times the multiplier determines the level of GDP. If, for example, the rate of interest is $i = \$10\%$, then investment will be 75 (point a). The multiplier times $[I(i) + G + X - M]$ yields output at point a'. If the interest rate rises to 20%, output will drop by the multiplier times ΔI to b'.

Table 10.2. IS data.

Parameters: $c_1 = 3/4$, $d_1 = 2/3$; $k_0 = 100$; $k_1 = 250$
$G + X - M = 40$

i	$I(i)$	$Y(i, 40)$	Yd	C
0.5	0	80	53	40
0.4	0	80	53	40
0.3	25	130	87	65
0.2	50	180	120	90
0.1	75	230	153	115
0	100	280	187	140

With $i = 10\%$ we have investment $I(10\%) = 100 - 250 \times 10\% = 75$, which yields point a on the investment relationship plotted on the left-hand panel of the graph. Substituting into the multiplier relationship yields $Y = (75 + 40) = 2 \times 115 = 230$, which is point a' on the on the IS curve plotted on the right-hand panel of the graph. If i increases to 20%, we will have $I = 100 - 250 \times 20\% = 50$, which is point b on the left-hand panel and generates, thanks to the multiplier, $Y = 180$ at point b' on the IS curve.

10.2.4 *IS: saving = investment (goods market equilibrium)*

The "IS curve," was originally introduced by Sir J. R. Hicks in 1937.[1] He chose to call it the IS curve because at every point on this relationship investment is equal to saving. This is the case if we define gross saving as

$$S \equiv Y - (C + G + X - M)\,; \tag{15}$$

that is to say, gross savings is defined as that part of output that is not used up for current private consumption, purchased by the government, or used up in meeting the excess of export demand over imports. With this definition, we easily find from the national income accounting identity, equation (3), that saving is indeed equal to investment:

$$S = I\,. \tag{16}$$

Goods market equilibrium

There is a more significant way of looking at the IS relationship. At every point on the IS curve the *goods market* is said to be in equilibrium in the sense that the output of the economy just equals the amount that the public wants to purchase (I is determined by the investment equation, C is determined by the consumption function, and $C + I + G + X - M = Y$). Thus we say that the IS curve shows all combinations of the interest rate i and output Y at which the *goods market* is in equilibrium.

Crowding out

Newspaper editorial writers sometimes draw the wrong implication from equations (15) and (16). It is a mistake to argue that the equations say that any increase in government spending or private consumption must

[1] J. R. Hicks, "Mr. Keynes and the Classics, a Suggested Interpretation," *Econometrica*, 1937.

"crowd out" private investment by causing a reduction in gross saving, unless offset by a fall in the balance of trade $(X - M)$. This argument implicitly assumes that Y is fixed. Such crowding out need not occur if the increases in G or C are accommodated by an increase in Y, which may occur if the economy is not already operating at full capacity. Indeed, if there is plenty of slack in the economy, higher G will stimulate an expansion in C, Y, and S through the multiplier mechanism. A more precise examination of the extent to which government spending may be said to crowd out private investment will be presented later in this chapter.

10.2.5 *Government spending shifts the IS relationship*

Figure 10.4 shows how fiscal and monetary policy influence economic activity. This graph is similar to Figure 10.3, except that there is an additional IS curve corresponding to an increase in $G + X - M$ from 40 to 80. The IS curve shifts to the right by the increase in government spending times the multiplier. To see why, suppose that the interest rate stays at 20%. Then investment stays at 50, as before, but with the higher level of government spending we have find on substituting into equation (14) that Y has increased to 260, which is point a^* on the new IS curve plotted on Figure 10.4. This demonstrates the following proposition:

> *An increase in government spending shifts the IS curve horizontally to the right by the change in government spending times the multiplier.*

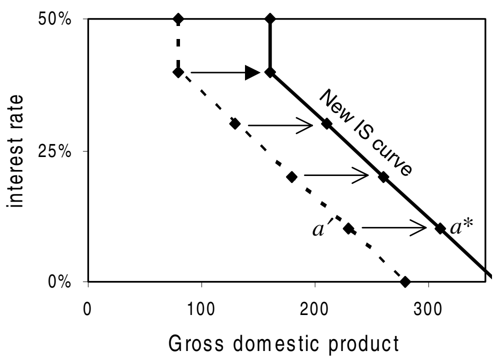


Fig. 10.4. $\Delta(G + X - M)$ shifts the IS curve
Y will increase by the multiplier times the change in $G + X - M$ *if* the rate of interest does not change.

That is why the IS curve shifts to the right by the multiplier times $\Delta(G + X - M)$.

Table 10.3. IS data on the multiplier.

Parameters: $c_1 = 3/4$, $d_1 = 2/3$; $k_0 = 100$; $k_1 = 250$		$G + X - M = 40$	$G + X - M = 80$
i	$I(i)$	$Y(i, 40)$	$Y(i, 80)$
0.5	0	80	160
0.4	0	80	160
0.3	25	130	210
0.2	50	180	260
0.1	75	230	310
0	100	280	360

Note: For any given interest rate, $\Delta G = 40$ yields $\Delta Y = 80$.

10.3 The money supply

While the IS curve is a useful construct explaining how changes in interest rates may affect the pace of economic activity, it leaves unanswered the question of what determines the rate of interest. This requires a look at the way in which the central bank determines the money supply and how changes in the money supply in turn influence the rate of interest.

Although the Bank of England, Britain's central bank, was founded in 1694, the United States central bank, our Federal Reserve System (or "Fed" for short), is less than a century old, having been established by act of Congress in 1913, just in time to help finance World War I. A *central bank* is a "bankers' bank" in that it is an institution where banks themselves can store funds, borrow, and obtain a variety of banking services. The central bank takes deposits from commercial banks just as the commercial banks in turn accept deposits from the public. The central bank stands by to make loans to commercial banks even when unstable financial markets threaten their solvency and no one else is willing to support the system. Thus the central bank is said to serve as the "lender of last resort." More than this, the central bank determines the nation's monetary policy and is responsible for regulating the availability of credit, the structure of interest rates, and the quantity of money in circulation.

The Fed is a powerful institution. *Newsweek* magazine once declared that Paul Volcker, who served as Chair of the Board of Governors of the Federal Reserve System from 1979 to 1987, was "the second most powerful man in America!" Precisely the same remark has been made about his successor, Alan Greenspan, who was initially appointed by Ronald Reagan and was reappointed by George H. W. Bush and Bill Clinton. Can the Chair of the Fed be more powerful than the Vice President, the Chief Justice of the Supreme Court or the Speaker of the House?

In order to understand how central bank policy influences economic developments it is necessary to learn how the quantity of money is defined and measured. It is also necessary to learn what determines the quantity of money that the public will desire to hold. This will lead us to the concept of the "LM Curve," which interacts with the IS curve to determine simultaneously the rate of interest and the level of output.

10.3.1 *Measuring the money supply*

In everyday language we use the word *money* to refer to a variety of related but not identical concepts. We may say that Bill Gates has a lot of money to mean that he is wealthy. But if he had left his wallet at home we might say that he was embarrassed to find himself without any "money," meaning that he was short of currency or coin. Economists use the word "money" in three basic senses:

#1. Money refers to our *medium of exchange*, meaning that it is an asset that we use to effect transactions, such as purchasing a loaf of bread, paying the rent, buying a car or purchasing government bonds.

#2. Money can also be used as a *store of value*, because we may choose to hold part of our wealth conveniently at hand in order to be prepared to effect purchases at a later date.

#3. Money is also said to be a *unit of account*; thus it may be useful not only to size up a person's wealth in terms of the almighty dollar but also to use dollars when recording debts in legal documents and to express the price of both a loaf of bread and a year's college tuition in terms of dollars.

While gold, silver and copper coins have often been used as money, economic anthropologists report that an amazingly wide variety of items, including sea shells and even cows, have at times constituted a society's accepted form of money. During World War II captured American soldiers in German prisoner of war camps commonly used cigarettes as the medium of exchange. Gold is particularly well suited for use as the medium of exchange because its scarcity means that substantial wealth can be represented by a rather small quantity; moreover, it is durable and convenient to carry.

Economists find it useful to measure the total quantity of money in circulation, in part because it is a useful indicator of economic conditions and in part because it is a fundamental concept helping us to understand how the monetary authorities attempt to influence economic developments. A confusing array of monetary concepts is recorded on Table 10.4. The first item on the table is currency in circulation, which includes coin as well as paper money. Not counted as currency in circulation are dollar bills that have been printed but have yet to be placed in circulation and currency that is in the vaults of banks rather than circulating. Currency, however, is not our primary medium of exchange or store of value; checking deposits are much more significant. So it is necessary to add to currency in circulation the value of demand deposits, which is the technical term for funds that the public has placed on deposit in checking accounts. The sum of these two items is M_1, the first measure of the quantity of money in circulation. A second "broader" measure used by many economists is M_2, which augments M_1 by adding funds in savings accounts and certain other assets. While there are many other definitions of the money stock, M_1 and M_2 are the two of primary importance.

Note: The symbols M_1 and M_2 refer to the two money supply concepts. This is entirely different from M, which stands for imports.

A serious problem with measuring the money stock that circulates in the United States is created by the fact that the $100 bill is a very attractive form of currency. A million dollars in $100 bills can be easily carried across national borders in a suitcase; and unlike checks and electronic cash machines, currency leaves no trail. In large part because the dollar is regarded as secure, the $100 bill is the international currency for money hoarders and the underworld. More than this, Ecuador, El Salvador, Guatemala, and Panama have all officially used U.S. dollar bills as their official currency, both as a means of payment and as a store of value. It is estimated that as much as one third of United States currency is held abroad. All this means that the quantity of currency in circulation in the United States is much less than the value of the bills that the United States Mint has placed in circulation.

Table 10.4. The money supply.

	Billions of dollars					
	1960	1970	1980	1990	1999	2000
Kinds of money						
Currency	28.8	48.6	115.3	247.0	516.9	530.1
+ Demand deposits	111.6	164.7	261.2	276.9	358.9	309.9
+ Other checkable deposits	–	0.1	28.1	293.7	241.4	240.9
+ Travelers checks	0.3	0.9	3.5	7.0	8.2	8.0
$= M_1$	140.7	214.3	408.1	824.6	1,125.4	1,088.9
+ Small denomination time deposits	12.5	151.2	728.5	1,173.4	952.4	1,043.6
+ Savings accounts	159.1	261.0	400.3	923.2	1,738.8	1,875.8
+ Money market mutual fund balances			63.5	358.0	846.1	934.0
$= M_2$	312.3	626.5	1,600.4	3,279.2	4,662.7	4,942.3
GDP (pY)	527.4	1,039.7	2,795.6	5,803.2	9,299.2	9,872.9
p (GDP price deflator, 1996 = 100)	22.4	29.3	57.4	86.8	104.4	107.0
Real value of the money supply						
$\underline{M}_1 = M_1/p$	628.1	731.4	711.0	950.0	1,078.0	1,017.3
$\underline{M}_2 = M_2/p$	1,394.2	2,138.2	2,788.2	3,777.9	4,466.2	4,617.5
Velocity of money						
$v_1 = pY/M_1$	3.7	4.8	6.9	7.0	8.2	9.1
$v_2 = pY/M_2$	1.7	1.7	1.7	1.8	2.0	2.0
$v = pY/\text{currency}$	18.3	21.4	24.2	23.5	17.9	18.6
currency/M_1	20.5%	22.7%	28.3%	30.0%	45.9%	48.7%
currency/M_2	9.2%	7.8%	7.2%	7.5%	11.1%	10.7%
M_1/M_2	45.0%	34.2%	25.5%	25.1%	24.1%	22.0%

Source: *Federal Reserve Bulletin*.

10.3.2 *How central banks control the supply of money and credit*

It is not easy to understand precisely how a country's central bank influences the quantity of money in circulation and the structure of interest rates. Indeed, the Federal Reserve Act was based on a misunderstanding of the process. It was thought that the primary mechanism by which the Fed would exercise control over the banking system would be the discount rates charged by the twelve Federal Reserve Banks when they made loans to commercial banks. In order to make sure that the lending activities of the Fed did not lead commercial banks to make loans that would be used for speculative purposes, loans made by the Fed to commercial banks were to be secured in turn by loans made by the commercial banks to their customers for "productive credit," such as loans to finance carrying inventories between the time the goods are manufactured and when they are paid for by the final customer. But it turned out that the entire process of discounting is of only secondary importance.

In the early 1920s the Fed discovered by accident that there was another much more potent channel by which it influenced economic conditions. In an effort to augment their profits, the Federal Reserve Banks made substantial purchases of government securities on the open market. That is to say, they purchased outstanding U.S. government bonds from brokers who were selling them on behalf of commercial banks, insurance companies, and other bondholders. As intended, the Federal Reserve banks did earn interest income on these bonds, but they were surprised to also observe that their purchases increased the reserves of the commercial banks and contributed to the increased availability of credit and lowered interest rates. At the time of this serendipitous discovery no one found it easy to explain precisely how an open market operation — the buying and selling of government securities by the central bank — influences banking conditions. We will work through the process step by step.

Reserve requirements

Banking institutions in the United States are legally required to keep reserves behind their deposit liabilities (i.e., funds the public has placed on deposit in checking and saving accounts). As a first approximation it is convenient to assume that these *required reserves* equal 10% of demand deposits.[2] The banks must meet this requirement by holding adequate reserves in the form of two types of assets: deposits at the Fed plus currency in the bank's vaults. Thus each bank must have

$$\begin{aligned}&\text{Deposits at Fed} + \text{currency at the bank} \geq \text{Required Reserves}\\&\quad \equiv 0.1\ \text{Demand Deposits} \qquad (17)\end{aligned}$$

Excess reserves refers to any excess of deposits at the Fed plus currency at the bank over Required Reserves.

A bank is said to have a *reserve deficiency* if its reserves fall short of required reserves. How banks address a reserve deficiency will be explained in a moment.

The commercial banks can put their excess reserves to work by making loans, and in doing so they not only earn interest income. We will find that their loan making activities also have a profound influence on the quantity of money in circulation.

An example will illustrate how the bank's excess reserves are reduced when it makes loans to the public. Suppose a bank with excess reserves makes a \$20,000 loan to a reliable customer to purchase a new car, placing the \$20,000 in the checking account of the borrower. The effect of this transaction on the customer's and the bank's financial positions are recorded on the abbreviated balance sheets presented on Table 10.5. In order to more clearly focus on the essentials the abbreviated balance sheets show only the changes resulting from the indicated transactions. The symbol Δ indicates the change from last year.

Step #1

The first step is for the lending bank to place the \$20,000 in the customer's checking account (i.e., the customer's demand deposit). This shows up as an asset on the customer's balance sheet but as an increase in liabilities on the bank's balance sheet — it owes the \$20,000 in the checking account to its customer. But there is an offsetting \$20,000 increase in the borrower's liabilities because he has borrowed that amount from the bank.

[2]Reserve requirements are imposed on commercial banks, savings banks, credit unions and other institutions accepting demand deposits. Technically, these institutions are referred to as "depository institutions," but it simplifies the discussion to refer to them as banks. The formula used by the Federal Reserve System in computing reserve requirements for banks and other depository institutions is more complicated than this simple 10% rule. A fuller explanation consumes almost half a page of small type in the *Federal Reserve Bulletin*, a monthly publication of the Fed.

Step #2

Next the borrower presents a \$20,000 check to the car dealer to pay for the car. The customer gets the car and the car dealer deposits the check in her bank.

Step #3

Finally, the car dealer's commercial bank may send the check for clearing to the Fed; the Fed will pay that bank for the check by placing \$20,000 in its deposit at the Fed. When the borrower's check clears, the borrower will have \$20,000 deducted from his checking account and the borrower's bank will have \$20,000 deducted from its deposit at the Fed.

Although these transactions mean that \$20,000 of deposits at the Fed have been redistributed from the lending bank to the car dealer's bank, there has been no change in total bank reserves. But the money supply, because it includes demand deposits, is up by the \$20,000 increase in demand deposits (0 net change for the car buyer and a \$20,000 increase in the car dealer's deposits). Further, required reserves are up by $1/10 \times \$20{,}000 = \$2{,}000$.

The essential point is that the \$20,000 loan has used up \$2,000 of excess reserves and generated a \$20,000 increase in the money supply.

Banks do not earn interest on their vault cash or their deposits at the Fed. They do earn interest by making loans. That is why each bank usually finds it profitable to continue to extend credit (i.e., make loans) to its customers until it is *loaned up*; i.e., they no longer have excess reserves.[3] Thus if there were excess reserves of \$1 billion dollars in the banking system, the banks would be able to make new loans of \$10 billion. Each dollar of excess reserves can support a \$10 increase in credit and generate a \$10 increase in the money supply.

Assuming that banks remain loaned up, we must have:

$$\begin{aligned}&\text{Demand Deposits}\\ &\qquad = 10 \times (\text{Deposits at the Fed} + \text{Currency at the banks})\,. \qquad (18)\end{aligned}$$

[3]Banks monitor their reserve position much more tightly when interest rates (the opportunity cost to the bank of holding excess reserves) are high. The reverse situation prevailed through most of the Great Depression of the 1930s. In 1939 the banks held \$11.5 billion of reserves when only about \$6.5 were required; they were holding excess reserves of \$5 billion because interest rates were extraordinarily low — the rate on Treasury bills was only 0.023%; it was 0.59% on government securities maturing in three to five years. With interest rates so low it did not pay banks to bother putting their excess reserves to work.

Table 10.5. Effects of a $20,000 bank loan to a car-buying customer.

Step #1: The bank lends a reliable customer $20,000, placing the funds in the customer's checking account.
Step #2: The customer writes a check for $20,000 to the car dealer; the car dealer deposits the check in her bank.
Step #3: The car dealer's bank sends the check for clearing through the Federal Reserve Bank.

Changes in Customer's Balance Sheet

Assets		Liabilities	
(#1) ΔDemand deposit	+ $20,000	(#1) ΔLoan from bank	+ $20,000
(#2) ΔDemand deposit	− $20,000		
(#2) ΔCar!	+ $20,000		

Changes in Lending Bank's Balance Sheet

Assets		Liabilities	
(#1) ΔLoans Outstanding	+ $20,000	(#1) ΔDemand Deposits	+ $20,000
(#3) ΔDeposit at the Fed	− $20,000	(#3) ΔDemand deposits	− $20,000

Changes in the Car Dealer' Balance Sheet

Assets		Liabilities	
(#2) ΔCar inventory	− $20,000		
(#2) ΔDemand Deposits	+ $20,000		

Changes in Car Dealer Bank's Balance Sheet

Assets		Liabilities	
(#2) ΔChecks in process	+ $20,000	(#2) ΔDemand Deposits	+ $20,000
(#3) ΔChecks in process	− $20,000		
(#3) ΔDeposit at Fed	+ $20,000		

Notes:
1. Each transaction must be offset by another transaction or transactions in order that the balance sheet will still balance.
2. Each transaction generally involves entries on the balance sheets of both parties to the transaction.
3. The result of the loan is an increase in demand deposits of $20,000 and an increase in reserve requirements of $2,000.

To see why this equation holds as an equilibrium condition, suppose instead that the banks had a reserve deficiency, demand deposits being more than ten times the reserves of the banking system. The banks could address the reserve deficiency by cutting back on their lending activity, which would reduce their demand deposits and hence their required reserves. With a reserve requirement of 10%, $10 in loans must be held back for every dollar by which reserves fall short of requirements. Contrariwise, if demand deposits were less than ten times reserves, the banks would find it financially profitable to put their excess reserves to work by making additional loans, which would increase demand deposits, drive up required reserves and eliminate the excess reserves.

Banks with excess reserves may lend to their own customers, but they may also adjust their reserve posture by using reserves to purchase government securities. Or they may lend funds to other banks suffering from a reserve deficiency or located in a region of the country where it is particularly profitable to expand lending operations. Banks may also borrow from the central bank. These mechanisms deserve careful examination.

Discount operations

Banks short of reserves may borrow from the Fed. The rate of interest charged the commercial banks by the Fed is known as the *discount rate.* While it was thought when the Fed was first established that discounting would be the major channel by which the Fed would influence the behavior of commercial banks, open market operations turned out to be the basic technique by which the Fed influences interest rates and the quantity of money in circulation.

Open market operations

The Fed influences the banking community primarily by buying and/or selling government securities on the open market. When the government borrows to help finance its spending, it does so in small part by selling saving bonds to the public, but the vast bulk of the government's debt is in the form of large denomination bonds and other securities that are freely transferable but cannot be redeemed until they mature, which may be as long as 30 years after the date of issue. When the government needs funds in excess of what it can finance with taxes, the treasury issues these securities by auction. Banks, insurance companies and private individuals submit sealed bids to the treasury. Because the securities are transferable, they do not have to be held by the original purchaser until maturity but can be sold to banks or to insurance companies or to the public through brokers at a price determined by supply and demand. The Fed influences money market conditions by buying or selling government securities in this open market. Usually the Fed focuses on buying or selling Treasury bills (T-bills). These are short-term securities maturing in 30 days, 90 days or 6 months.

What happens when the Fed sells \$1 billion of government securities on the open market? To keep things simple, suppose that the securities are purchased by commercial banks. As shown on Table 10.6, the banks pay

Table 10.6. The Fed sells government securities on the open market.

When the Fed sells $20 million of government securities on the open market, commercial banks purchasing the securities pay for them by running down their deposits at the Fed.

Changes in Fed's Balance Sheet			
Assets		Liabilities	
ΔT Bills	− $20 million	Member bank deposits	− $20 million
Changes in Commercial Banks' Balance Sheet			
Assets		Liabilities	
ΔT Bills	+ $20 million		
ΔDeposits at the Fed	− $20 million		

Note: The open market sale has reduced the reserves of the banking system.

for the securities by running down their deposits at the Fed, which means that the reserves of the banks are reduced by $1 billion and their ability to make loans to private borrowers is reduced by $10 billion.[4] That is how the open market sale of $1 billion results in a $1 billion decrease in the reserves of the banking system, which will lead to a $10 billion contraction in private credit and the money supply. Conversely, a Fed purchase of $1 billion of government securities in the open market will be expansionary because the selling banks are paid for the bonds by an increase in their deposit at the Fed.

Evolution

Banking structure continually evolves, and sometimes major changes can occur without legislative act that have a fundamental affect on the profitability of the banks and the reported magnitudes of the monetary aggregates. While the Federal Reserve Act still requires that banking institutions keep 10% reserves behind their transaction deposits, since 1994 the Federal Reserve Board has allowed commercial banks to reduce their required reserves by using computer programs that dynamically "sweep" their demand deposit liabilities into a type of personal savings account known as money market deposit accounts (MMDA). When customer demand deposits are

[4]The open market sale has essentially the same effect even if the government securities are purchased by an insurance company rather than by a commercial bank. The essential difference is that the insurance company pays for the securities by writing a check. Thus its demand deposits are reduced, and when the check clears, the commercial bank will have a corresponding reduction in its reserves at the Fed. The Fed can not dictate who will buy the security, but it does not have to. The Fed simply sells the securities in the market place at the going price.

swept into money market deposit accounts they are no longer subject to reserve requirements, they no longer count in M_1, but they are still included in M_2. Since the funds are moved back as required when the customers write checks on their accounts, the depositor is unaware of the practice. But it contributes substantially to bank profitability because they reduce the funds they have to tie up in reserves, which earn no interest. The banks still keep cash in the vault on a voluntary basis to meet the day-to-day needs of their customers; and they still have to have a balance at the Fed for check clearing and other purposes. It is estimated that by the beginning of year 2001 about 40% of M_1 was being swept out. Because the effective reserve ratio is now much lower than 10%, the impact of an open market operation of given size is considerably larger than before, but the essential features of the process are the same.[5]

The federal funds market

Rather than loaning out excess reserves to its own customers, a bank with excess reserves can lend them to other banks wishing to acquire reserves, either to meet a reserve deficiency or to expand their lending operations. Table 10.7 shows how the balance sheets are affected. Observe that the reduction in the reserves of the lending bank is precisely offset by the increase in the reserves at the borrowing bank. The funds are lent for a single day in amounts of $1 million or more, but banks with a continuing need or surplus of funds can and do enter the market every day for a protracted period. The loan is negotiated by one of a number of brokers who specialize in such transactions — this is known as the federal funds market. The interest rate paid on funds loaned from one bank to another in this way is called the *federal funds rate*, which is quoted daily in the financial press.

The federal funds rate, determined by market forces, fluctuates so as to equate the demand and supply for federal funds. When, as in recent times, the Fed chooses to focus its policy on stabilizing the federal funds rate at a targeted level, it does so by adjusting the reserves of the banking system through open market operations. Pumping reserves into the banking system through an open market purchase will reduce the need for the banks to borrow on the federal funds market, which will lead to a lower federal funds rate. Conversely, when the Fed sells government securities on the

[5]For further analysis see Richard G. Anderson and Robert H. Rasche, "Retail Sweep programs and Bank Reserves, 1994–1999", *Review of the Federal Reserve Bank of St. Louis*, January/Feburary, 2001, and the following website:
http://www.stls.frb.org/research/swdata.html.

Table 10.7. The Middletown Bank lends $500,000 to the New York Bank.

When the Middletown bank lends $500,000 to the New York Bank on the open market, the funds are transferred by adjusting their deposits at the Fed.

Changes in Middletown Banks Balance Sheet

Assets		Liabilities	
Deposits at the Fed	− $500,000		
IOU from New York Bank	$500,000		

Changes in New York Bank's Balance Sheet

Assets		Liabilities	
Deposits at the Fed	+ $500,000	IOU owed the Middletown Bank	+ $500,000

Note: There is a shift of reserves from the Middletown Bank to the New York Bank, but no change in the total supply of reserves for the whole banking system. Since there is no change in demand deposits, there is no change in required reserves at either bank or for the banking system as a whole.

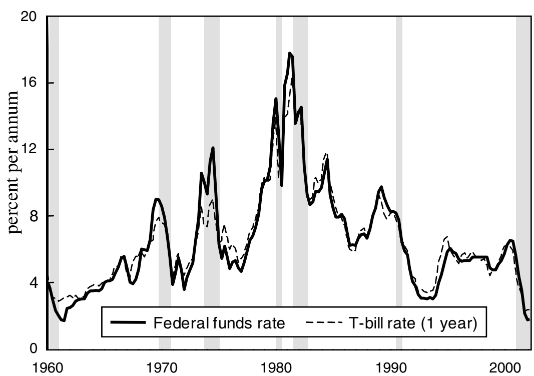


Fig. 10.5. The federal funds rate
The federal funds rate (the rate of interest banks pay on funds borrowed overnight from other banks) is closely tracked by the T-bill rate because, as an alternative to entering the federal funds market, banks can buy or sell T-bills.

open market they are paid for through a reduction in bank deposits at the Fed, which means that the banks reserves are reduced. The reduction in the supply of reserves tends to push up the federal funds rate. The time path of the Federal funds rate is reported on Figure 10.5.

Decision making at the Fed

The Federal Reserved System consists of the *Board of Governors*, 12 District Reserve Banks and the Open Market Committee. The seven governors of the Federal Reserve Board are appointed for fourteen-year terms by the President, subject to confirmation by the Senate. The President appoints one of the governors to serve as Chair for a four-year term. The key decision making body is the *Open Market Committee*, which consists of the Federal Reserve Board plus the President of the Federal Reserve Bank of New York and four of the other Federal Reserve Bank presidents, serving on a rotating basis. The Open Market Committee usually meets eight times a year in Washington to review the current economic condition and to determine appropriate policy. Between meetings they rely on conference calls to resolve matters requiring a prompt decision.

Populists sometimes object that this is a most undemocratic system. It is fair to say that the Fed is no less democratic than the Supreme Court, whose judges have lifetime appointments. And like the Supreme Court, the Fed must be above politics. Indeed it should be independent of the executive branch — otherwise a president might be tempted to pressure the Fed to stimulate the economy in advance of the next election.

Financing war

King Henry VIII of England [1491–1547] financed his wars with France by debasing the British coinage. All the old coins had to be exchanged for new, where the new coins were obtained by melting down the old. But Henry had the Royal Mint blend lead with the reclaimed gold, which meant that for each old coin turned in for replacement there was enough extra gold to mint another coin for Henry! This *seigniorage profit* was used to finance the King's expenditures, but the result was a substantial increase in the nominal money supply and inflation.

Governments no longer finance wars by debasing the coinage. When paper currency replaced gold as the primary medium of exchange, governments would resort to the printing press to finance war. George Washington's troops were paid with paper currency, the "Continentals." Nowadays, governments usually finance their wars by borrowing from the banking system, but the end results are much the same.

During World War II the Fed's independence was compromised. Instead of fighting inflation by hiking interest rates, the Fed was called upon to help the Treasury control the cost of servicing (paying interest) on the soaring national debt by keeping interest rates extraordinarily low, as was reported on lines 20 and 21 of Table 8.3. The rate on 3 month T-bills remained at only 0.375% throughout the war. The yield on long-term government bonds was kept below 2.5%. The Fed kept interest rates low by purchasing government securities on the open market, which gave commercial banks the reserves they needed to expand their lending activities and purchase government securities. The result was an excessive expansion of the money supply, which contributed to inflationary pressure.

An example will illustrate how financing a war by government borrowing can lead to an expansion in the money supply. Suppose the commercial banks purchase $50 billion of government bonds issued by the Treasury in order to help finance a War, "Bonds for Peace." What will be the effect of financing government spending in this way? The answer may be found by

Table 10.8. Borrowing $50 billion to finance a war.

Step #1: Suppose, for simplicity, that $50 billion of bonds are purchased by commercial banks.
Step #2: The Treasury spends the funds for the war, and the defense contractors (or their employees and suppliers) place the funds received from the government in their accounts at commercial banks.
Step #3: The Fed (Central Bank) purchases $5 billions of government securities sold by the commercial banks, paying for them by adding $5 billion to the deposits of the banks at the Fed.

Changes in Balance Sheet of Commercial Banks (consolidated)

Assets		Liabilities	
#1 ΔT Bills	+ $50 billion		
#1 Deposits at the Fed	– $50 billion		
#2 Deposits at the Fed	+ $50 billion	#2 Demand deposits	+ $50 billion
#3 Deposits at the Fed	+ $5 billion		
#3 ΔT Bills	– $5 billion		

Changes in the Fed's Balance Sheet

Assets		Liabilities	
		#1 Member bank deposits	– $50 billion
		#1 Treasury deposits	+ $50 billion
		#2 Treasury deposits	– $50 billion
		#2 Member Bank deposits	+ $50 billion
#3 ΔT Bills	+ $5 billion	#3 Member Bank deposits	+ $5 billion

Notes:
1. At the end of Step #2 demand deposits are up by $50 billion, which means that required reserves are up by $5 billion. If the commercial banks were loaned up before the war, they now suffer a reserve deficiency. This will push the Federal Funds rate higher, discourage bank lending and inhibit the purchase of additional government securities.
2. As a result of Step #3, the deposits of commercial banks at the Fed have increased by just enough to provide the reserves required by the $50 billion addition to the money supply of step #2.
3. There has been an increase in the money supply of $50 billion because demand deposits count as part of the money supply.

working through the balance sheet changes recorded on Table 10.8. The borrowing and subsequent spending of the funds by the government leads to an increase in demand deposits equal to the amount borrowed. Since the reserves of the commercial banks have not increased, the banks are under pressure to curtail their lending. They are in no position to buy more government securities. But as Step #3 shows, the Fed can support the Treasury's deficit financing of the war by purchasing enough government securities to replenish the reserves of the banking system. As long as the Fed is willing to bail out the Treasury in this way, it will be possible for the government to borrow more and more funds without putting upward pressure on interest rates, but at the expense of an increase in the money supply that is likely to contribute to inflation.

Expansionary or Contractionary Monetary Policy

How can one judge at any point of time whether the Fed is stimulating the economy, slowing it down, or taking a neutral course? This is a vital question for "Fed Watchers" trying to profit from predicting what the Fed will do and how its actions are likely to affect financial markets. It is also an issue that has contributed to prolonged controversy among academic economists.

Members of the "monetarist" school have argued that the Fed should focus on stabilizing the rate of growth of the money supply. At times the rate of growth of the money supply, sometimes M_1 and sometimes M_2, has been used by most economists and the financial press as the primary gauge for judging whether the Fed is adopting a tight or loose monetary policy. It is said that the Fed's primary focus should be on controlling the quantity of money in circulation, or its rate of growth, by buying or selling government securities on the open market. The Fed should target a rate of growth of the money supply equal to the long run sustainable rate of growth of the economy.

Currently, the primary focus of monetary policy is on the federal funds rate. When the Fed announces that it will lower the federal funds rate by pumping reserves into the banking system, the financial community concludes that the Fed wishes to stimulate the economy, perhaps because it is concerned that the economy may be sliding into a recession. In the reverse situation, when the Fed fears inflation, it pushes the federal funds rate up in an effort to slow down an over-heated economy that is subject to excessive expansionary pressure. When it wishes to pursue a "neutral policy" it keeps the federal funds rate stable.

10.4 Money and interest rates

Before we can understand how monetary policy works, we must learn what determines how much of their assets the public will want to hold in the form of money. The quantity of money (e.g., M_1 = currency + demand deposits) that the public will want to hold may reasonably be assumed to depend in part on interest rates and in part on the level of income. This is reasonable because the interest rate is the opportunity cost of holding money balances. When interest rates are high, the cost to the public of holding money rather than lending out the funds is increased. That is why a hike in interest rates, other things being equal, is likely to lead to a decrease in the quantity of money the public will desire to hold. But as with most commodities, demand is also influenced by changes in income. At a higher level of income the public will be spending more, which means that people will find it useful to hold larger money balances for executing transactions, at least if interest rates hold steady. Let $L(i, Y)$ denote the demand for money balances.[6] It is reasonable to suppose that $\partial L/\partial i < 0$, because higher interest rates make it more expensive to hold money balances. And $\partial L/\partial Y > 0$ because at a higher income level the public will want to hold larger money balances on hand to execute transactions. The demand and supply of money will be in balance only when

$$\frac{M_1}{p} = M_1^r = L(i, Y)\,, \tag{19}$$

where p is the price index and Y is real GDP. Here is a functional form for $L(i, Y)$ that captures the essential features of this relation:

$$L(i, Y) = kY(i - \rho)^{-\eta}\,. \tag{20}$$

This demand function involves an income elasticity of demand for money of unity, which is the usual assumption. This makes perfect sense for it implies that if incomes were to double the public would need twice as much M_1 to execute transactions. Note that if ρ were 0, η would be the interest elasticity of the demand for money.

[6]The symbol L is often used to denote the demand for money balances because money is said to be the most liquid of assets. An asset is liquid if it can be easily and quickly converted into another type of asset of the owners choosing. A house, for example, is an illiquid asset because considerable time and effort is required to sell it.

Numerical Example

Suppose that the demand function for M_1 is of the form

$$M_1 = pL(i, Y) = \frac{0.07pY}{(i - 2\%)} = \frac{0.07Y^n}{(i - 2\%)}, \quad i > 2\%, \tag{21}$$

where $Y^n = pY$ is the nominal level of GDP, $k = 0.07$, $\rho = 2\%$ and $\eta = 1$. Then we have $\partial M_1/\partial Y = 0.07p/(i - 2\%) > 0$ and $\partial M_1/\partial i = -0.07pY/(i - 2\%)^2 < 0$.

If the Fed wishes to peg the interest rate i at a particular level, demand for money equation (21), or more generally equation (20), will reveal the quantity of money, M_1, that the Fed will have to establish through open market operations, given ρ and Y.

Alternatively, if the Fed wishes to establish the money supply at a particular level, the resulting rate of interest can be derived from the inverse demand function for money obtained by solving (21) for i as a function of income and M_1. For example, demand for money equation (21) yields

$$i = L^{-1}(p, Y, M_1) = 0.07\frac{pY}{M_1} + 2\%. \tag{22}$$

Here $L^{-1}(p, Y)$ denotes the inverse of the demand for money function with respect to i. This says that demand for money function, equation (21), implies that the interest rate will be linearly related to the nominal level of GDP, given the quantity of money in circulation. Also, if the money supply grows at the same rate as nominal income, so that Y^n/M_1 is constant, then the rate of interest will be stable.

Further insight into the way income and interest rates interact in determining the demand for money is provided by Figure 10.6, which resembles the demand curve for any commodity, say wheat. The graph was generated with the data derived with equation (21) and reported on Table 10.9. It shows how the quantity of money demanded depends on the rate of interest, for two alternative levels of income. The graph plots the demand for money for GDP of $Y^n = 90$ and $Y^n = 360$. As the table helps to make clear, at lower levels of output (e.g., $Y^n = 90$) the demand for money balances would be lower at any given interest rate than it is with $Y^n = 360$; but the demand would be much stronger if the level of income were at a higher level, say $Y^n = 500$. Thus the demand curve for money as a function of the rate of interest shifts when income changes.

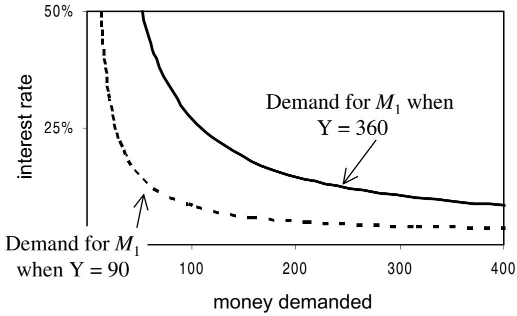


Fig. 10.6. Interest rates, income, and the demand for money
The lower the rate of interest, the less the cost of holding money and the greater the quantity of money the public wants to hold, given income. When income increases, the demand for money curve shifts to the right because more money is demanded at each rate of interest.

Table 10.9. The demand for money.

i	$M_1(i, 90)$	$M_1(i, 360)$
50%	13.1	52.5
40%	16.6	66.3
30%	22.5	90.0
20%	35.0	140.0
10%	78.8	315.0
5%	210.0	840.0
3%	630.0	2520.0

Note that while the Fed may be able to achieve a targeted rate of interest or a desired level of the money supply, it cannot determine both independently, given GDP; i.e., it must settle for a point on the demand for money function. For example, if the Fed wants to achieve an interest rate of 10%, equation (21) dictates that it will have to have a money supply of $M_1 = 79$ if $Y^n = 90$, $M_1 = 158$ if $Y^n = 180$, or $M_1 = 315$ if $Y^n = 360$. The rate of interest generated in achieving a money supply target of say 70 depends on the level of income, as can be seen from equation (21).

10.4.1 *Quantity theory of money*

We may manipulate demand for money equation (20) to obtain with (19) a relationship known as the quantity theory of money:

$$M_1 v(i) = pY\,, \tag{23}$$

where $v(i) = (i - \rho)^\eta/k$. Many economists have observed that at least over certain time periods velocity has been quite stable and appears to be insensitive to interest rates. And as will be explained later in this chapter, it can be argued that in the long run $dv/di = 0$, which yields the standard form of the quantity equation:

$$M_1 v = pY \tag{24}$$

The symbol v is called the income velocity of money in that it is a measure of how fast the money supply M_1 turns over in supporting the transactions involved in generating nominal income pY. Sometimes a broader definition of money, such as M_2, is used in applying the quantity theory.

Quantity theorists regard this equation as of fundamental importance in understanding what the monetary authorities are about. Under the assumption that velocity v tends to be stable, they point out that if the money supply M_1 grows at the same long-run rate as Y, the price level will also be stable. Inflation is generated when central banks, instead of following this simple rule, permit excessive growth of the money supply.

Data on the velocity of money for the United States, both $v_1 = pY/M_1$ and $v_2 = pY/M_2$, are plotted on Figure 10.7. The evidence suggests that neither velocity concept has been particularly stable, and v_1 was particularly erratic near the end of the 20th century.[7] Partly for this reason, the Fed currently focuses on controlling interest rates rather than the rate of growth of the quantity of money in circulation.

Some say that the rate of change in the money supply is a useful predictor of economic recessions. While it is clear from Figure 10.8 that the rate of growth of the money supply usually drops off in advance of recessions (as identified on our graphs by the NBER stripes), not all downturns in the rate of growth of the money supply have been followed by recessions. In the mid 1960s, the mid 1980s and the mid 1990s the money supply falsely signaled recession in that the money supply growth rate went through a substantial period of decline without a subsequent recession. Nobel Laureate Milton Friedman, the inspirational leader of the monetarist school of economists throughout the second half of the 20th century, stressed that

[7] Velocity is affected by a variety of institutional changes. For example, now that banks can sweep demand deposits into money market deposit accounts in order to reduce required reserves, the reported magnitude of M_1 is reduced, which means that v_1 increases. Because money market deposit accounts are included in M_2, this broader definition of money and v_2 are not affected by the sweeping of demand deposits.

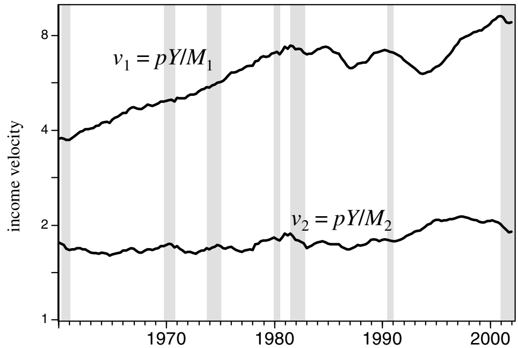


Fig. 10.7. Income velocity of money
The velocity of M_1 is the ratio $v_1 = pY/M_1$. Velocity measures how rapidly M_1 changes hands in executing transactions relating to the production and distribution of the GDP. Similarly, $v_2 = pY/M_2$ is the velocity of M_2. If nominal GDP (pY) were to increase at exactly the same rate as the money supply, velocity would remain constant.

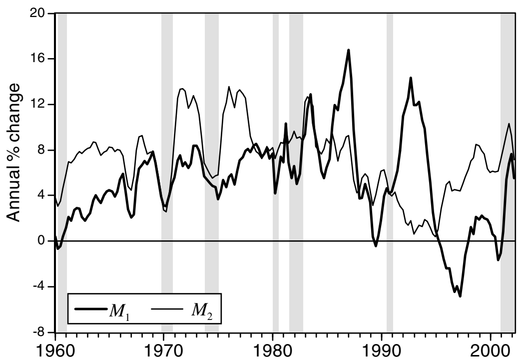


Fig. 10.8. Annual rate of change in M_1 and M_2
Economists of the monetarists school focus on the annual rate of change in M_1 and M_2 in evaluating whether the Fed has succeeded in stabilizing the money supply. As can be seen from the graph, the growth rates of both M_1 and M_2 have been far from stable.

the economy responds to changes in the money supply with a long and variable lag.

10.4.2 *IS-LM curve interaction (Model C)*

An alternative plot of the demand for money function is provided by Figure 10.9, which graphs equation (22) for given M_1, with Y on the abscissa and i on the ordinate. This graph was named the "LM Curve"

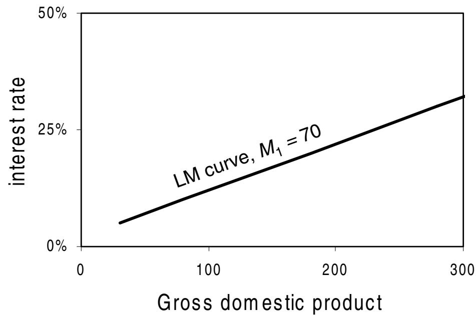


Fig. 10.9. The LM curve ($M_1 = 70$) — equilibrium in the money market
The LM curve shows the combinations of income and the interest rate at which the public will be willing to hold the quantity of money in circulation, given p. Above the LM curve the demand for money is less than the supply. Below the LM curve the demand for money exceeds the supply.

Table 10.10. The LM curve ($M_1 = 70$).

i	v	Y
50%	6.9	480
40%	5.4	380
30%	4.0	280
20%	2.6	180
10%	1.1	80
5%	0.4	30

by its inventor, Sir J. R. Hicks. Only at points on the LM curve can the money market be said to be in equilibrium in the sense that the demand for money is equal to supply, or as Hicks wrote, the demand for "*L*iquidity" is equal to the supply of *M*oney (that is why Sir Hicks called it the "*LM* curve"). At any point to the right or below the LM curve, the demand for money is greater than the supply. At any point above or to the left of the LM curve the quantity of money that the public wishes to hold is less than what the central bank has chosen to create. Note that an increase in the supply of money will shift the LM curve to the right because, at any given level of the interest rate, the public will be willing to hold the increased money supply only at a higher level of income.

Combining the IS and LM equations will yield a new model that is very helpful in understanding how monetary policy influences the economy. As can be seen from the flowchart on Figure 10.10, the IS-LM couples our earlier analysis of the IS curve with the LM curve. The rate of interest

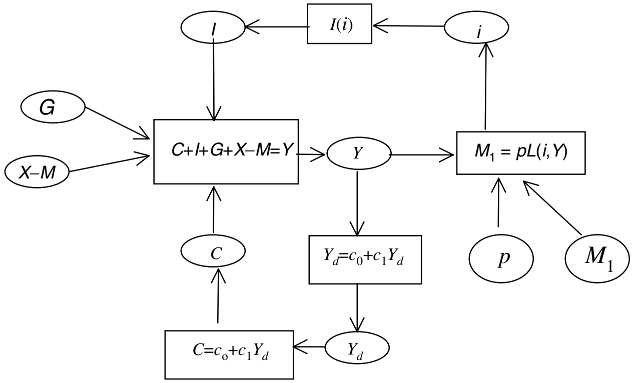


Fig. 10.10. The IS-LM flowchart (Model C)
The IS-LM flowchart adds the demand for money equation and two new exogenous variables: the nominal money supply and the price level. It creates a second monetary feedback loop capturing the effect of changing income on the demand for money and hence interest rates and investment spending.

is now an endogenous variable. A new exogenous variable, the nominal money supply, is controlled by the Fed. While initially we will regard the price level as fixed, $p = 100$, we will return to price movements once we have developed the basic IS-LM apparatus.

Plotting both the IS curve from Figure 10.3 and LM curve from Figure 10.9 on Figure 10.11 provides a deterministic system yielding simultaneously both the equilibrium rate of interest and the level of GDP. Only where the two curves intersect do we have equilibrium in both the goods market (because we are on the IS curve) and in the money market (because we are on the LM curve). At the intersection point we have the interest rate at the level at which the public will be willing to hold the quantity of money in circulation and, simultaneously, the output of the economy is equal to the sum of C, $I(i)$, G and $X - M$. Once we have found the equilibrium level of GDP and the interest rate it is easy to find the resulting levels of consumption, disposable income and investment spending by going back to equations (1), (2), and (7) that underlie the IS curve.

Fiscal-Monetary policy interaction

The IS-LM apparatus provides an explanation of how fiscal and monetary policies interact in the determination of economic developments. Figure 10.12 shows how an increase in government spending can stimulate the

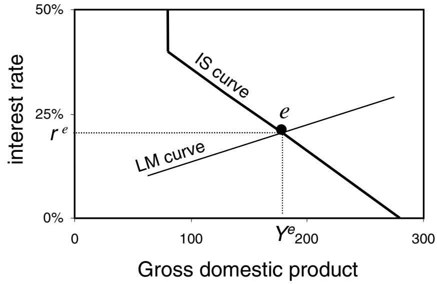


Fig. 10.11. IS-LM Interaction
At every point on the IS curve output equals demand — the goods market is said to be in equilibrium.
At every point on the LM curve the demand for money is equal to the supply — the money market is in equilibrium.
Only at equilibrium point e are both the goods and the money markets in equilibrium. This point simultaneously determines both GDP and the interest rate.

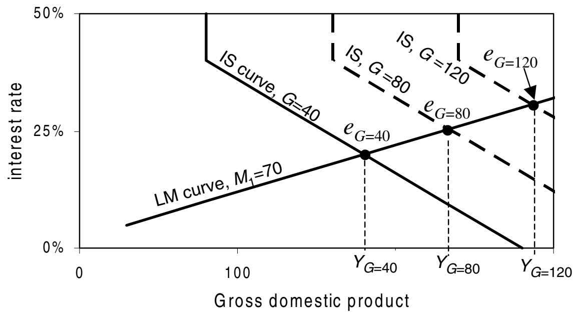


Fig. 10.12. Fiscal policy
An increase in government spending from $G = 40$ to $G^* = 80$ shifts the IS curve to the right by the multiplier times the increase in government spending. But the actual expansion in output will be less than what the government multiplier suggests because with expanding output the demand for money increases, forcing up the interest rate as we slide up the LM curve. The rising interest rate leads to a cutback in private investment spending that partially offsets the increase in government spending. Because of the crowding out of private investment, the level of output increases by less than the simple multiplier suggests.

economy, given the money supply. Note that while the IS curve has shifted to the right by the multiplier times the increase in government spending, the upward thrust to the economy is held in partial check by a rise in interest rates, which discourages private investment spending as determined by equation (7); i.e., the increase in government spending has crowded out some private investment. How much crowding out occurs depends on the slope of the LM curve and the sensitivity of private investment to interest rate movements. The expansionary effect of increased government spending will be particularly weak if private investment is very sensitive to the rate of interest (flatter IS curve) or the demand for money is quite insensitive to changes in the rate of interest (steeper LM curve).

As an alternative to a hike in G, a tax cut may be used to stimulate a depressed economy. A tax cut will have effects similar to a hike in G in that it also causes the IS curve to shift to the right, although the shift will be smaller than that sparked by a government spending hike of the same dollar magnitude because the tax multiplier is weaker, as we saw in Chapter 9.3.2.

Figure 10.13 reports what happens if the Fed attempts to stimulate the economy by increasing the nominal money supply through an open market operation, assuming that fiscal policy does not change. The LM curve shifts downwards to the right when the Fed increases the money supply. This is because the public is willing to hold more money at any given level of income only at a lower rate of interest. The economy adjusts to a new equilibrium at intersection point e_c on the old IS curve, where interest rates are lower and income is higher than before. The fall in interest rates has stimulated private investment spending, contributing to an increase in output.

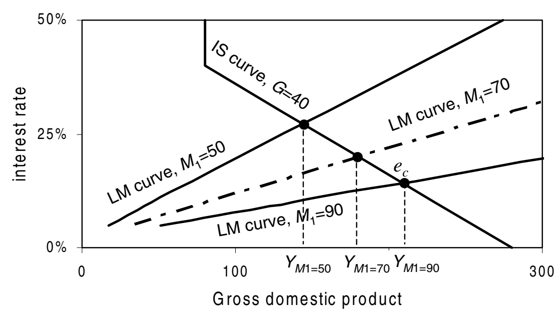


Fig. 10.13. Monetary policy
An increase in the money supply from $M_1 = 70$ to $M_1 = 90$ would push the LM curve downwards because, at any given Y, the public would be willing to held the larger quantity of money only if the rate of interest falls. The lower rate of interest encourages investment spending, leading to a multiplicative expansion in GDP to equilibrium point e .

A decrease in the money supply from $M_1 = 70$ to $M_1 = 50$ would shift the LM curve upwards to the left, which would lead to a higher interest rate. As a result, private investment would drop off and the pace of economic activity would slow.

The right policy mix

Should monetary or fiscal policy be used to stimulate a depressed economy? Part of the answer depends on the relative potency of the alternative measures. An increase in the money supply is more likely to have a strong expansionary effect on output if the demand for money is insensitive to the rate of interest, other things being equal. This is because a smaller $\partial M/\partial i$ means that the interest rate must fall by more in order to induce the public to hold the increased money balances engineered by the Fed. Further, the expansion in output induced by the fall in interest rates will also be stronger if investment spending is quite sensitive to the rate of interest, so that the IS curve is rather flat.

It turns out that in precisely these circumstances in which monetary policy is likely to provide a particularly potent stimulus, fiscal policy is likely to be weak. For example, if investment is particularly sensitive to the rate of interest, the ΔG induced expansion in Y and resulting rise in i is likely to crowd out considerable investment, which will partially offset the multiplier effect of ΔG. Conversely, fiscal policy will be more potent and monetary policy weaker if interest rate changes have little effect on investment.

Because expansionary fiscal policy (a rightward shift of the IS curve) tends to generate higher interest rates, it crowds out private investment spending, and the reduction in investment means that in the years to come worker productivity will be less than it otherwise would have been. This is so because reduced investment implies that in the future employees will be working with less machinery and equipment. In contrast, expansionary monetary policy (a rightward shift of the LM curve) works by lowering the rate of interest, which stimulates private investment spending on new factories and equipment that may serve in the longer run to increase worker productive and enlarge the nation's productive capacity. Thus the issue of monetary versus fiscal policy revolves at least in part on one's views about the appropriate size of the government sector and the desirability of achieving adequate investment for a more productive future.

Monetary and fiscal policy makers may be inclined to move in opposite directions in time of war. When military necessity forces a major expansion in government spending, the Fed might quite appropriately pursue a tight monetary policy in order to hold in check the tendency for the economy to over expand. But monetary and fiscal policy may move in opposite directions for other reasons. Traditionally the Fed has worried more about the dangers of inflation than the threat of unemployment. Since the Fed is designed to be at least partially independent of both the legislative and executive branches of government, it may at times push in a more conservative direction than the makers of fiscal policy, who may be worried about winning the next election.

Figure 10.14 shows what happens if the Fed, ever worried about inflation, responds to expansionary fiscal policy (tax cut or hike in government spending) by cutting the money supply: The IS curve had shifted to the right because fiscal policy is expansionary. The LM curve shifts to the left because of the reduction in the money supply. In terms of output the two policies have more or less offset each other, but that does not mean that the effect is neutral. There has been a shift in the output mix of the economy toward more G but, thanks to the hike in interest rates, reduced investment spending. This means that long run growth prospects are being held in check in favor of higher government spending.

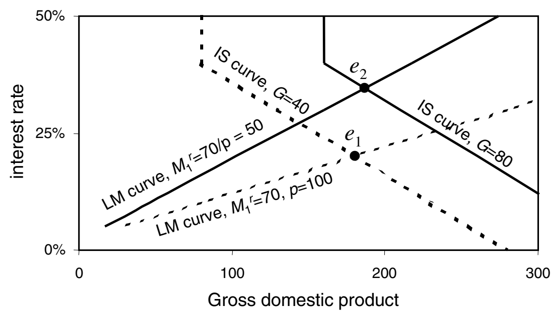


Fig. 10.14. Fiscal expansion plus monetary contraction
The initial equilibrium at point e_1 is disturbed by a massive increase in government spending from $G = 40$ to $G = 80$, which pushes the IS curve to the right. But if the Fed simultaneously cuts back on the nominal money supply, pushing the LM curve upwards to the left, the expansionary pressure will be limited because rising interest rates will cutback private investment spending, and the economy ends up at point e_2. Clearly, how fiscal policy affects the economy depends on the response of the central bank.

10.4.3 *International complications*

For the sake of analytical simplicity, it is helpful at times to pretend that we live in a closed economy that does not engage in international trade. But in fact economies are open, although some are obviously much more dependent on international trade than others. Central bankers are inclined to worry not only about how their actions affect inflation but also how they affect the foreign exchange rate. More than this, the international arena provides a second route by which changes in interest rates affect the pace of domestic activity.

Interest rates, the exchange rate and the pace of economic activity

International capital responds to differences among countries in the level of interest rates. And this provides a second channel by which the Fed affects the pace of economic activity. When the Fed raises interest rates it encourages foreign investors to shift their funds into the United States in order to earn a higher return. And the increased demand for the dollar will push up its value in the foreign exchange market, as can be seen on Figure 3.13. Now the rise in the value of the dollar will tend to encourage imports. Why? Because United States importers will find that their dollars go further in overseas markets. Simultaneously, United States exporters will find that they have more difficulty selling American goods abroad because they are now more expensive for foreigners to buy. This means that higher interest rates tend to reduce the foreign trade balance, $X - M$; i.e., $\partial(X - M)/\partial e < 0$, where e is the exchange rate. And because the effective demand for the nation's output is more sensitive to changes in interest rates, the IS curve is flatter.

How much the link between interest rates, the exchange rate and the balance of trade affects GDP depends on the magnitude of $\partial Y/\partial(X - M)$, $\partial(X - M)/\partial e$ and $\partial e/\partial i$. As was shown by equation (24) of Chapter 9.3.2, the foreign trade multiplier, $\partial Y/\partial(X - M)$ is of the same order of magnitude as the government spending and investment multipliers.[8] Both $\partial(X - M)/\partial e$ and $\partial e/\partial i$ are likely to be exceptionally strong for a small country, particularly if it is heavily dependent on international trade. This is so because only a small proportion of the available mobile international capital needs to be attracted into a small country by rising interest rates in order to have a substantial impact on its exchange rate. And only a small percentage change in the balance of trade will have a big impact on a country that has a high ratio of exports and imports to GDP.

Figure 10.15 adds open-economy complications to the IS-LM graph. The IS curve is drawn flatter than in the earlier graphs to indicate the greater sensitivity of output to changes in interest rates, thanks to the balance of trade effect. The new curve on the graph, labeled $de/dt = 0$, shows those

[8]See exercise 9.1 to learn how the multiplier is complicated when M is a function of Y.

combinations of the interest rate and output at which the current exchange rate, e^* will be stable. It is drawn under the assumption that the central banks are not entering into the foreign exchange market. This line has a positive slope because $\partial e/\partial i > 0$ but $\partial e/\partial Y < 0$. As indicated on the graph, above this line $de/dt > 0$ because the high interest rate, relative to the level of Y, is attracting so much in the way of foreign funds as to push the exchange rate above its current level. Below the line we have $de/dt < 0$ because Y is large relative to i, which places downward pressure on the foreign exchange rate. Since equilibrium point e is on the $de/dt = 0$ line, there is no movement in the foreign exchange rate.

While the exchange rate is stable, output is below Y^c. If the central bank expands the money supply in order to stimulate economic activity, a side effect of the fall in interest rates as the economy slides down the IS curve will be downward pressure on the foreign exchange rate. If instead, the government increases its spending, the IS curve will shift to the right and the equilibrium point will slide up the LM curve. But while output expands there will again be downward pressure on the foreign exchange rate. It may be possible to escape this dilemma by placing a heavier foot on the fiscal policy accelerator while simultaneously putting the brakes on

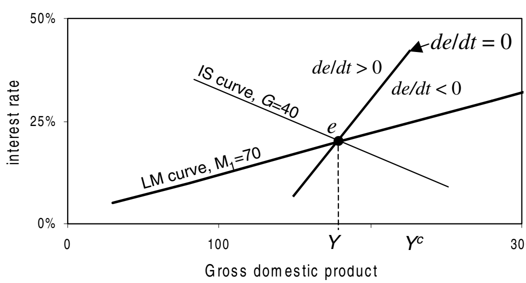


Fig. 10.15. Open-economy IS-LM analysis
At any point on the $de/dt = 0$ line the exchange rate is stable. To the left of the line the exchange rate strengthens ($de/dt > 0$); below or to the right of the line the exchange rate deteriorates ($de/dt < 0$).

At equilibrium point e output is substantially below capacity output Y . But because the equilibrium point is on the $de/dt = 0$ line, there is no tendency for the foreign exchange rate to change. If expansionary fiscal policy shifts the IS curve to the right, the equilibrium point will slide up to the right along the LM curve leading to greater employment. But this will place downward pressure on the foreign exchange rate; i.e., $de/dt < 0$. Expansionary monetary policy would also stimulate the economy by shifting the LM curve outward, but the fall in interest rates as the equilibrium point slides down the IS curve will cause the exchange rate to deteriorate; once again, $de/dt < 0$.

the money supply. The leftward shift in the LM curve coupled with a strong rightward shift of IS may allow equilibrium point e to slide up the $de/dt = 0$ line as the economy expands toward full employment. Alternatively, the central bank may purchase dollars in the foreign exchange market, but at the expense of running down its holdings of foreign currencies, gold and other international reserves.

10.4.4 *IS-LM analysis of demand pull inflation*

Although the price level is an exogenous variable in the basic IS-LM model, that framework can be extended to provide insight into the nature of the inflationary process, provided we take care in analyzing the effects of changes in the price level on the real money supply. Now the Fed may be able to determine the *nominal* quantity of money in circulation, M_1. But $L(Y, r)$ is the demand for *real* money balances — i.e., M_1/p. Looking back at equation (22), we can verify that the position of the LM curve depends on *real* money balances. Now suppose that given M_1, as determined by the Fed, the economy is initially operating close to its full potential, as indicated by point e_1 on Figure 10.16. Here Y^c denotes potential or capacity out-

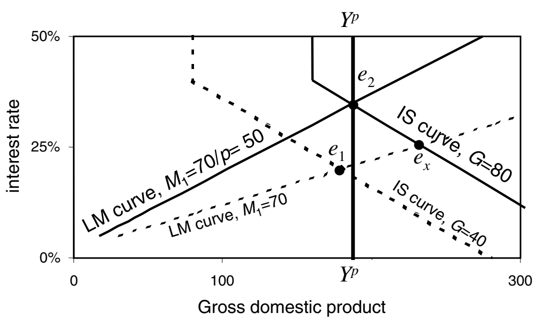


Fig. 10.16. Self-limiting inflation
When increasing government spending shifts the IS curve so far to the right as to push the economy beyond the capacity ceiling indicated by the vertical Y line, equilibrium at point e is impossible, and the inflationary gap makes inflation inevitable.

Provided the Fed keeps the nominal money supply M_1 fixed, the real money supply, $M_1 = M_1/p$, will contract when p goes up. The decline in the real value of the money supply will shift the LM curve upwards and to the left. The inflationary process will continue until the LM curve has shifted far enough to the left to eliminate the inflationary pressure. For this example, we require $M_1/p = 70/p = M_1 = 50$. Therefore, the price level must rise to $p^* = 70/50 = 140$ before the required reduction in the real value of the money supply is achieved and the inflationary gap eliminated at e_2.

put, perhaps as estimated with Okun's Law. And it may be assumed that initially the price level is at $p = 100$. If government spending increases, pushing the IS curve to the right, an "inflationary gap" may be said to be created because real output cannot expand to the point required to achieve equilibrium at point e_x. As a result of the shortage of goods, prices are bid up. The rise in prices reduces the real money supply to M_1/p_2, which shifts the LM curve to the left. Thus the rise in prices reduces the gap between capacity and the demand for the nation's output, which means that the inflationary gap is somewhat smaller than before. If the initial rise in prices reduces but does not suffice to eliminate the inflationary pressure, prices will continue to rise, but perhaps at a slower rate. This inflationary process will continue until the economy is restored to equilibrium with stable prices at point e_2.

The self-correcting process by which inflation itself tends to eliminate the inflationary gap requires that the monetary authorities keep the nominal money supply stable. If instead of keeping the money supply under control the central bank "validates the inflation" by letting the nominal money supply expand in step with rising prices, the real money supply will not decline and the LM curve will remain at LM_1 instead of shifting to the left. When the Fed fails to exercise restraint, there will be no tendency for the inflation to correct itself and prices will rise continuously in an inflationary spiral. That is why it is often said that sustained inflation is not possible unless the central bank allows the nominal money supply to get out of control.

10.5 Aggregate demand — Aggregate supply

In micro economics the interaction of demand and supply determine price. The IS-LM apparatus at best explains only the demand side of inflation, a type of inflation that develops when excessive demand pushes too hard against the nation's capacity to produce. The *demand-pull* theory of inflation, as it is sometimes called, cannot explain why inflation at times takes place when the economy is operating at less than full employment, as happened in so many countries during the 1970s and early 1980s. The aggregate demand and supply curve apparatus will provide an enriched understanding of the inflationary process by incorporating cost considerations directly into the analysis. The aggregate demand curve that we will derive will have a negative slope, just like the demand curves of Chapter 3. But its derivation will be completely different because this aggregate demand curve attempts to summarize how the demand for the nation's output depends upon prices, given government spending and the nominal quantity of money in circulation. The aggregate supply curve will have a positive slope, just like the supply curves of Chapter 3, but it will be relating the total supply, the nation's output, to the overall price level. Like Chapter 3, the intersections of the two curves will determine equilibrium price p and level of real output Q.

10.5.1 *Derivation of the aggregate demand function (Model D)*

The aggregate demand function that we will plot on Figure 10.17 specifies the demand for the nation's output (GDP) as a function of the price level, given government spending and the nominal money supply of $M_1 = \$90$. For any price level p we can find the point on the aggregate demand function by first computing the real money supply, M_1/p. This determines the location of the LM curve on the top panel of Figure 10.17. Next, the intersection of this LM curve with the IS curve determines output, thus providing one point on the aggregate demand curve on the bottom panel. Repeating this process for different values of p serves to map out the aggregate demand curve.

The first step in deriving the equation for aggregate demand is to substitute inverse demand for money function (22) and investment equation (7) into (4), obtaining

$$Y = c_0 + c_1 d_o + c_1 d_1 Y + k_o - k_1 \left(\frac{0.07Y}{M_1/p} + 0.02 \right) + G + X - M \,. \quad (25)$$

In this equation Y is real GDP, p is the price level, and M_1 is the nominal money supply. Algebraic manipulation eventually yields the reduced form equation for aggregate demand:

$$Y(p, M_1, G + X_M) = \frac{c_0 + c_1 d_o + k_o - 0.02k_1 + G + X - M}{1 - c_1 d_1 + 0.07k_1 p/M_1} \,. \quad (26)$$

The aggregate demand curve is the plot of $Y(p, M_1, G+X-M)$ on p, given M_1 and $G + X - M$. As with the demand curves of Chapter 3, p is plotted on the ordinate and output on the abscissa. The aggregate demand curve must have a negative slope, conforming to the Law of Demand of Chapter 3, because $\partial Y/\partial p < 0$; i.e., there is an inverse relationship between price level p and aggregate demand. Also, $\partial Y/\partial G > 0$ and $\partial Y/\partial M > 0$.

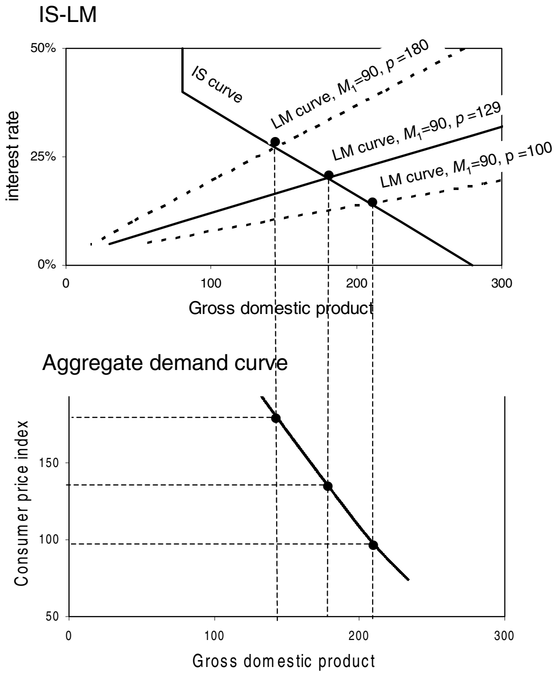


Fig. 10.17. Derivation of the aggregate demand curve
Given that the central bank keeps the nominal money supply fixed at $M_1 = 90$, the real money supply and hence the position of the LM curve depends upon the price level, as shown on the top panel.

For each price level and corresponding LM curve we can map out one point on the aggregate demand curve on the bottom panel. Note that when $p = 129$, the real money supply is 70 and when $p = 180$ the real money supply is 50.

10.5.2 *Aggregate demand and supply interaction (Model E)*

An aggregate supply curve is required in order to determine simultaneously both output and the price level. We will use the symbol Q to represent the aggregate output that will be produced by an economy of profit maximizing firms operating in a competitive environment. Our aggregate supply function, based on the assumption of profit maximization, is of the form $Q(w^m/p, K)$, where w^m is the nominal wage, $w^m/p = w^r$ is the real wage rate, and K is the stock of productive capital. We must also assume that an increase in the real wage rate will lead employers to cut back on hours of work; i.e., $\partial Q/\partial w^r < 0$. This follows from the requirement that the real wage rate equals the marginal product of labor, which it will be recalled from equation 11 of Chapter 7.3.2 is necessary for profit maximization under competition. We must also distinguish between short and long run adjustments in the money wage rate.

Short-run wage rigidity

In the short run it may be reasonable to regard the money wage w^m as fixed. After all, some workers have their money wages set by union contract. Also, employers are reluctant to cut money wages for fear that it would undermine worker loyalty, undermine morale, and reduce productivity, as emphasized by the theory of efficiency wages presented in Chapter 7.3.5. Further, the minimum wage may place a floor under the money wage of workers at the bottom of the scale. Assuming that money wages are indeed fixed in the short run, a fall in prices will increase the real wage, which will lead employers to cut employment and output. That is one explanation of why the short-run aggregate supply curve should be drawn with a moderately positive slope, as on Figure 10.18. The slope if the aggregate supply curve is thought to be steeper farther to the right, where the economy is closer to capacity, because rising output in this region may lead to shortages in some sectors of the economy, which puts additional upward pressure on prices.

Now consider the effect of changes in fiscal policy. As can be seen from the graph, a reduction in government spending or some other force leading to a leftward shift of the aggregate demand curve causes the equilibrium point to slide to the left along the aggregate supply curve. This will generate reduced output, idle productive capacity and unemployment. In spite of

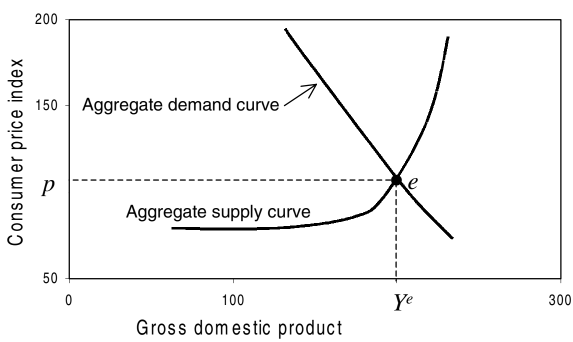


Fig. 10.18. Aggregate demand and supply interaction (short run)
The intersection of the aggregate demand and supply curves simultaneously determines both output and the price level.

the rising unemployment, rigid money wages mean that there will at best be only a slight reduction in prices when the economy slides along the aggregate supply curve. Alternatively, an increase in government spending will lead to a rightward shift of the aggregate demand curve and, as a result, an expansion of output, provided the economy has ample slack allowing for additional production. However the changes in output will be less than predicted by the simple multiplier mechanism of Chapter 9 for two reasons:

- As with the IS-LM model, the rise in output sparked by an expansion in government spending will contribute to an increase in the demand for money for transaction purposes. If the Fed keeps the nominal money supply stable, the tendency for the demand for money to rise when the economy expands must be held in check by a rise in interest rates; and the rise in interest rates will discourage private investment spending.
- Further, the aggregate demand and supply model recognizes that this crowding-out effect of a hike in government spending will be reinforced, given the nominal money supply. The expansion in output will involve rising prices, which will cause a reduction in the *real* money supply. Because, the decline in the real money supply shifts the LM curve to the left; interest rates will rise to choke off private investment spending and reduce output via the multiplier.

For these reasons, as well as the qualifications of Chapter 9.5, the multiplier mechanism overstates the short-run expansionary effects of an increase in government spending and tax cuts.

Cost-push inflation and the Fed's policy dilemma

Disruptions on the supply side may cause the aggregate supply curve to shift upwards, which leads to cost-push inflation. Such a shift will also come about if labor unions succeed in winning major wage concessions from their employers. The shift may be the result of abrupt changes in the costs of raw materials or major imports. The most notorious example of such cost-push inflation is provided by the success enjoyed by the OPEC (Organization of Petroleum Exporting Countries) cartel in pushing up the world price of petroleum in a series of price hikes beginning in 1973. The result was worldwide inflation that was of a magnitude not experienced since World War II. *Cost-push inflation* comes about when an exogenous price shock, such an OPEC engineered increase in the price of oil, pushes up the aggregate supply curve. An upward shift in the aggregate supply curve can generate both inflation and unemployment simultaneously, as is suggested by the movement from e_1 to e_2 on Figure 10.19. The term *stagflation* was coined to describe the simultaneous occurrence of inflation and unemployment in the 1970s and 1980s.

To the consumer all inflations may appear to be cost-push in origin. When the prices of groceries go up, the customer blames the grocer, the grocer replies that there was no choice but to raise prices because their supplier had pushed up prices, and the supplier will blame rising costs, including wage costs. The process by which rising prices permeate through the economy always appears to be cost-push, and this is so whether the inflation was generated by an exogenous price or wage shock or by too much demand pressing on productive capacity. Demand-pull inflation is sometimes described in the business press as "too much money chasing too few goods," but it might be better to say "too much purchasing power pushing against capacity," given the quite technical meaning that economists attach to the word "money."

Cost-push inflation confronts the central banker with most difficult choices:

- The central bank can fight the inflation by cutting back on the money supply, which will push the aggregate demand curve to the left. This will reduce the inflation at the cost of greater unemployment. Central bankers who pursue such a strategy run the risk of being accused of worsening the damage wrought on the economy by OPEC price hikes; they will be accused of adding higher unemployment on top of the disastrous inflation.

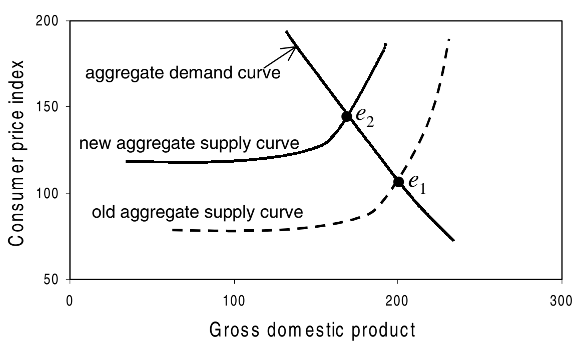


Fig. 10.19. Cost-push inflation — Stagflation
When a sharp jump in the price of oil or other imports or a substantial round of wage increases pushes up the aggregate supply curve, the consumer price index rises simultaneously with both a fall of output and an increase in unemployment.

- The central bank might instead focus primarily on the unemployment problem by expanding the money supply through open market operations in order to push the aggregate demand curve to the right. With this policy the central bank will without a doubt be blamed for contributing to inflation.
- The central bank may pursue a middle course, allowing only a moderate expansion of the money supply. With this strategy they will be blamed for both inflation and unemployment.

No matter what the Fed does when the economy is confronted with cost-push inflation, it is all too likely to be accused of making things even worse!

10.5.3 *Long run aggregate supply and demand interaction (Model F)*

It is only in the short run that money wages can be presumed to be rigid. When unemployment remains high for a protracted period, laid off workers may eventually become reconciled to accepting a lower wage. And when employment contracts expire, employers are likely to succeed in pushing through wage concessions if there is an army of unemployed competing for jobs. When money wages are flexible, real wages will adjust so as to equate the supply of workers with demand in long-run equilibrium. This is the process that is fundamental to the derivation of the economy's long-run aggregate supply function. And the properties of this long-run equilibrium, other than the price level, are determined in accordance with the micro-economic theory developed in the first several chapters of this text:

1. The quantity of labor supplied by each utility maximizing worker is determined by the influence of the real wage on the work-leisure choice, as was explained in Chapter 4.4.4.
2. Given the technology, employers operating in competitive markets will hire workers to the point where the real wage is equal to the marginal revenue product, just as was explained in Chapter 7.3.1.
3. In the long run the real wage adjusts so as to equate the demand for labor with the supply. In the absence of government intervention (for example, with the minimum wage), we will have the supply of labor matching the demand in the labor market, which gives us the level of employment.

4. Substituting the equilibrium employment level into the aggregate production function yields output, given the state of the technology and the existing stock of productive capital.

Now the level of output determined in the final step is long-run aggregate supply. Note that only the "real" side of the economy is utilized in the derivation of the long-run aggregate supply function. And because the supply and demand for labor both depend on the *real* wage, aggregate supply is independent of the absolute level of prices. For example, if equilibrium is obtained with a CPI of 150 and wages at $10 per hour, then the same quantity of labor would be supplied by workers and demanded by employers if prices and wages both doubled, the CPI climbing to 300 and money wages rising to $20 per hour. Because it is the real wage rather than the level of prices that determines the supply of labor, the long-run aggregate supply curve is drawn as a vertical straight line on Figure 10.20.

Note: The long run is not the longest run, in that the effect of changes in worker productivity arising from innovation and investment are neglected. Chapter 12 will consider a growth model taking these complications into account.

If in the long run output is determined by the vertical long-run aggregate supply curve, what is the role of monetary and fiscal policy? The answer to this basic question is very different from that for the short run. In the long run, output will return to the full employment level after a hike in government spending, but that is not the whole story. The rightward shift of the aggregate demand curve pushes the economy against capacity,

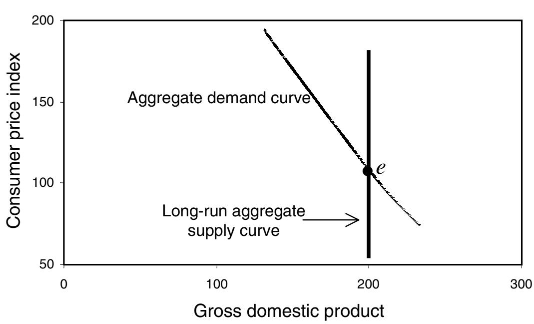


Fig. 10.20. Aggregate supply in the long run
In the long-run the aggregate supply curve is vertical and the output is determined by the level of employment at which the real wage equals the marginal productivity of labor.

generating inflationary pressure. Given the *nominal* money supply M_1, as determined by the central bank, rising prices serve to reduce the *real* money supply. As a consequence, interest rates are bid up because the cost of holding money must rise to the point where the demand for money balances is cut back to the available supply. And the rising interest rate in turn discourages investment in houses, new factories and equipment, as predicted by the IS curve. Thus, in the long run expansionary fiscal policy crowds out private investment spending, assuming the nominal money supply is held constant. If the central bank validates the inflationary pressure by expanding the money supply, there will be successive rounds of rising prices and an upward inflationary spiral.

10.5.4 *The classical dichotomy and the neutrality of money*

While in the short-run an increase in the nominal money supply may stimulate output, in the long run the quantity of money only determines the price level. The increase in the nominal money supply does push the aggregate demand curve upward to the right, as indicated on Figure 10.21. But the long-run equilibrium point slides up the vertical aggregate supply curve to point e. Note from equation (19) that the public's demand for real money balances, $M_1^r \equiv M_1/p$, is determined by real output Y and the rate of interest i:

$$M_1^r \equiv \frac{M_1}{p} = L(i, Y^r)\,. \tag{27}$$

While the central bank may determine M_1, it does not determine the *real* money supply $M_1^r = M_1/p$ because in the long run p adjusts to equate the supply of real money balances with the demand $L(i, Y)$ for real money balances. If M_1 doubles, equilibrium is restored when the price level doubles as well, just as is predicted by quantity theory, equation (24).

If the government has the central bank help finance substantial budget deficits year after year, the rise in prices generated by the expanding money supply means that more and more money will have to be created in order to finance a given level of real government spending. If the deficit is not brought under control, runaway inflation is inevitable.

The fact that the nominal money supply determines p while output is determined in the long run by all the factors underlying aggregate supply is often referred to as the *classical dichotomy*. It is a *dichotomy* because the task of explaining what determines the condition of the economy is separated into these two parts. And the proposition earns this

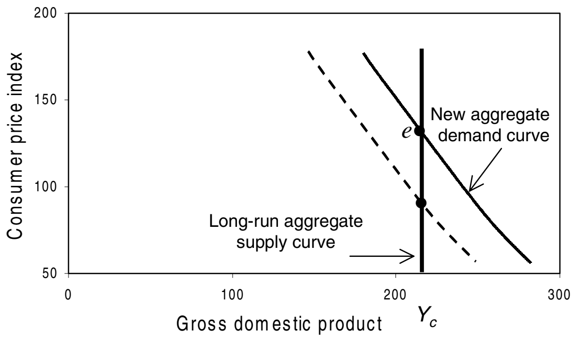


Fig. 10.21. Inflation generated by the expanding money supply.
When an expanding money supply pushes the aggregate demand curve upward, inflation results, but output is unaffected in the long run if the long-run aggregate supply curve is completely inelastic! Prices rise in the same proportion as the increase in the money supply, as predicted by the quantity theory (equation 27).

If expansionary fiscal policy pushes the aggregate demand curve upward, inflation results but there is no long run expansion in output. The multiplier is zero!

name *classical* because the proposition was understood and anticipated by classical economists more than two centuries ago, long before the aggregate demand and supply curve apparatus was developed. Indeed, philosopher-economist David Hume [1711–1776] explained:[9]

> Money is ... only the instrument which men have agreed upon to facilitate the exchange of one commodity for another. It is not one of the wheels of trade: It is the oil which renders the motion of the wheels more smooth and easy. If we consider any one kingdom by itself, it is evident, that the greater or less plenty of money is of no consequence; since the prices of commodities are always proportioned to the plenty of money ...

And Hume was absolutely right in saying that the price level will rise in proportion to the increase in the money supply. That is precisely the amount of inflation that is required to return the real value of the money supply to its former level. Changes in the nominal money supply lead to proportional changes in the price level so as to leave the real value of the money supply equal to the quantity that the public wants to hold at the long run equilibrium level of output.

[9]Robert Lucas quoted Hume at length in his 1996 Nobel Laureate address.

The concept of the classical dichotomy is sometimes characterized by the proposition that "money is a veil," meaning that the quantity of money in circulation does not affect the underlying conditions of the economy, such as real income, unemployment, etc., at least in the long run.

The short run

In the short run, changes in the nominal quantity of money in circulation can and do affect economic conditions. Hume explained:

> Alterations in the quantity of money, either on one side or the other, are not immediately attended with proportional alterations in the price of commodities. There is always an interval before matters be adjusted to their new situation; and this interval is as pernicious to industry, when gold and silver are diminishing, as it is advantageous when these metals are increasing. The workman has not the same employment from the manufacturer and merchant; though he pays the same price for every thing in the market. The farmer cannot dispose of his corn and cattle; though he must pay the same rent to his landlord. The poverty, and beggary, and sloth, which must ensue, are easily foreseen.

Or as Stanford Professor John Gurley has remarked: "Money is a veil, but when the veil flutters, real output sputters."

10.6 Monetarists versus Keynesians

Major controversy among macro economists has centered for more than half a century on the issue of price and wage flexibility. In his *General Theory of Employment, Interest and Money*, published during the Great Depression of the 1930s, economist John Maynard Keynes invoked the assumption of downward rigid money wages to explain the prevalence of massive unemployment. Keynes focused on the short-run analysis, captured on Figure 10.18. He dismissed the analysis of flexible wages, summarized on Figure 10.20, asserting that "in the long run we are all dead."

More than this, Keynes argued that in the absence of wage inflexibility and other rigidities, the economy might be caught in a downward deflationary spiral, citing Irving Fisher's "Debt Deflation" explanation of the great depression.[10] Falling wages would lead the public to anticipate still lower prices. As was explained in Chapter 8.5.2, $\dot{p} < 0$ will cause the real rate of

[10] Irving Fisher, "The Debt-Deflation Theory of Great Depressions," *Econometrica*, 1933.

interest to rise above the nominal interest rate. Or to put it another way, $\dot{p} < 0$ will lead investors to postpone investment purchases in anticipation of still lower prices, which means a greater demand deficiency and further price declines. The falling prices would put tremendous stress on debtor farmers and other borrowers who had not anticipated the collapse of prices when they incurred their mortgage obligations. While it is true that a lower price level means that the real value of the money supply, $M_1^r = M_1/p$, has increased, which would tend to push up investment and the equilibrium level of economic activity once prices stabilize, during the transition to the new equilibrium investment will contract. At best the process of adjustment will be painful. At worse, the process of adjustment may be unstable, degenerating into a downward deflationary spiral rather than adjusting to equilibrium.

The counter attack was led in the decades following World War II by Nobel Laureate Milton Friedmanof the University of Chicago. Friedman argued that changes in the money supply impact the economy with a long and variable lag. As a result, there is always the danger that attempts to stabilize the economy, however well intentioned, are likely to turn out to have been timed perversely — attempts at stimulating an economy suffering from unemployment may all too frequently have their punch only after the economy has already turned the corner. Given the limited state of current knowledge, Friedman argued, it is best to pursue a neutral monetary policy: the central bank should have the nominal stock of money grow at a constant rate in parallel with the long run growth rate of the economy. We should rely on the strong natural self-recuperating powers of the economy to restore long-run equilibrium rather than attempt to manipulate the economy with activist fiscal or monetary policy.

The battle lines between the Keynesians and the monetarists were clearly drawn in the decades immediately following World War II. Much of the debate hinged on the critical question of how rapidly the economy adjusts to long-run equilibrium. The next chapter will report on major contributions that were made by macro economists in analyzing these issues into the 21st century.

Summary

1. A concise summary of this chapter, starting with the multiplier as Model A, was presented on Table 10.1.

2. The IS curve (Model B) captures the effect of changing interest rates on GDP. When the central bank pushes interest rates down, investment spending is encouraged, which leads to an expansion in output equal to the change in investment times the multiplier. Expansionary fiscal policy, such as an increase in government spending or a tax cut, shifts the IS curve to the right. At every point on the IS curve the goods market is said to be in equilibrium in the sense that the sum of investment spending, determined by the rate of interest, plus consumption, government spending and the foreign grade balance all add up to the nation's output.
3. A nation's central bank, such as the Federal Reserve System of the United States (the Fed), controls the nation's money supply and influences interest rates. By the money supply we mean the stock of money in circulation. One measure of the nominal money supply, M_1, includes currency in circulation and demand deposits (funds placed by the public in checking accounts). M_2 is a broader concept of the money supply, including saving deposits and certain other assets in addition to M_1.
4. Banks in the United States are required to keep reserves equal to 10% of their demand deposits. Their holdings of currency plus their deposits at the Fed count in meeting their required reserves. Because banks do not earn interest on their reserves, they have an incentive to put any excess reserves to work by making loans to the public or buying government securities.
5. A bank with excess reserves may expand its lending activities to its regular customers. Or it may buy government securities, such as T-bills or government bonds. Or it may lend its excess reserves in the federal funds market to other banks suffering a reserve deficiency. The interest rate charged for such loans, determined by supply and demand, is called the federal funds rate.
6. The Fed's open market operations, the buying and selling of government securities, is the primary instrument for effecting monetary policy. When the Fed purchases government securities on the open market it pays for them by adding to the deposits at the Fed of commercial banks, which count as reserves. This increase in reserves of the banking system puts downward pressure on the federal funds rate and encourages lending activity. Contrariwise, when the Fed sells government securities, the reserves of the banking system are reduced, the federal funds rate moves upwards because banks have to pay more to borrow on the

federal funds market, and the commercial banks are led to curtail their lending activity.

7. Interest rates fluctuate so as to equate the demand for money with the supply. The LM curve specifies this inverse relationship between the interest rate and GDP at which the public is willing to hold the given money supply — at every point on the LM curve the money market is said to be in equilibrium in the sense that the demand for money equals the supply. If the Fed wishes to peg interest rates at a particular level it must shift the LM curve by adjusting the money supply so that it yields the targeted interest rate at the prevailing level of GDP. The Fed cannot simultaneously determine both the quantity of money in circulation and the price level.
8. The IS-LM diagram (Model C) combines these two functions in order to simultaneously determine the interest rate and GDP, given the real money supply. The point where the IS and LM curves intersect determines the level of GDP and the interest rate at which the goods market and the money market will both be in equilibrium. The diagram reveals that the economy may be stimulated by either monetary or fiscal policy.
   - An increase in the money supply will push the LM curve to the right, leading to a fall in interest rates and an expansion of GDP as the economy moves towards its new equilibrium by sliding down the IS curve.
   - Alternatively, an increase in G, by pushing the IS curve to the right, leads to an expansion of GDP and a rise in interest rates. Interest rates rise as the economy slides up the LM curve in order to offset the tendency for the higher pace of economic activity to increase the demand for money — the rise in interest rates holds the demand for money equal to the fixed supply.
9. The IS-LM model provides limited insight into the inflationary process. If the intersection of the IS-LM curve is beyond the capacity of the economy to produce, there will be an inflationary gap. The excessive purchasing power results in demand-pull inflation. If the Fed keeps the nominal money supply stable, the rising prices will reduce the real value of the money supply, pushing the LM curve to the left, thus limiting the inflationary pressure and slowing the economy. Thus demand-pull inflation will be eliminated if the quantity of money is kept under control.

10. The aggregate demand curve (Model D) is derived from the IS-LM apparatus, given the level of government spending and the nominal money supply. Since government spending is constant, the IS curve is stable. Given the nominal money supply the Fed has chosen to create, an increase in the price level will reduce the real money supply, pushing the LM curve to the left, and reducing GDP. This inverse relationship between the price level and the level of output, given M and G, is the aggregate demand curve.
11. The positively sloped short-run aggregate supply curve is derived from the production function under the assumption of profit maximization, and given the money wage rate. The higher the price level the lower the real wage, given the fixed money wage. And the lower the real wage the more workers the employers will want to hire according to marginal productivity theory. Thus we have a positively sloped aggregate supply curve, given the money wage rate.
12. Model E combines the aggregate demand and supply curves. An increase in government spending or the money supply will push the aggregate demand curve to the right. Output will expand more and prices rise less if the aggregate supply curve is relatively flat. When the economy approaches closer to capacity at higher levels of output, the aggregate supply curve becomes steeper and the primary effect of the expansion will be on prices rather than output.
13. An increase in the money wage rate (or other cost shock, such as a rise in the price of petroleum imports) will push the aggregate supply curve upward, which will stimulate the pace of economic activity but generate demand-pull inflation.
14. In the long run the aggregate supply curve is vertical. Output is at the full-employment level determined by the underlying real side of the economy. Money is a veil, determining only the price level. This division is called the "classical dichotomy."
15. What is the most appropriate mix of fiscal and monetary policy? One answer to this contentious question is provided by the observation that tight fiscal policy balanced with expansionary monetary policy will stimulate private investment spending, increasing worker productivity, and providing for a more prosperous future. Many economists — monetarists — have argued over the years that rather than attempting to manage the economy, the central bank should follow a neutral policy of constant money supply growth. These policy issues will be addressed again in the next chapter.

Key Concepts

Exercises

1. Consider the following equations for the Never-Never Land economy (all data in billions of \$):

$$C = 25 + 0.8Y_d \tag{1}$$

$$Y_d = 5 + 0.75Y \tag{2}$$

$$I = 1000 - 4000i \tag{3}$$

$$Y = C + I + G\,. \tag{4}$$

   a. Derive the reduced form equation for Y as a function of exogenous variable G and i.
   b. Suppose that the rate of interest is 10% and government spending is 400. Determine the level of investment, disposable income, consumption and GDP.

Table 10.11. The Never-Never Land IS curve.

	a	b	c	d	e
Interest rate (i)	5.00%	10.00%	______	20.00%	10.00%
Investment (I)	______	600	400	200	600
Disposable income (Y)	2,309	______	1,559	1,184	2,684
Consumption (C)	1,873	______	1,273	973	2,173
Government (G)	400	400	______	400	800
GDP ($Y = C + I + G$)	3,073	______	2,073	1,573	3,573

c. Fill in the blanks on Table 10.11.

d. Plot the IS curve on a neat graph. Show how the IS curve would shift if government expenditure increased to 800.

2. Suppose the demand for money equation is $M_d = Y/2.5 + 30/i$.

   a. Fill in the blanks on Table 10.12.

   b. Using the data on the table, plot the LM curve for $M_1 = 1,228$ on the graph you constructed for question 1.

   c. Now estimate, graphically if you like, the equilibrium rate of interest and the levels of consumption, and disposable income.

Table 10.12. The LM curve for Never-Never Land.

	a	b	c	d	e
GDP	1,570	2,320	2,570	2,695	______
Interest rate (i)	5.00%	10.00%	15.00%	______	25.00%
Money demanded (M_1)	______	______	1,228	1,228	2,448

3. Suppose that the Fed purchases four million dollars of government securities on the open market.

   a. Assume that commercial banks sold the government securities that are involved in the open market operation. The commercial banks are paid for the bonds by an increase in their deposits at the Fed. Show the direct effects of this open market transaction on the balance sheets of the Federal Reserve System and of the commercial banks.

      Hints: Write out only the *changes* in the balance sheet (use the same general format as on Table 10.6). Note that if and only if your *changes* sum to the same figure on both sides, will the balance sheet still balance after the changes, as required. How would this transaction affect the actual and required reserves of the commercial banks?

b. How would the change in the reserves posture of the commercial banks affect their willingness to make loans to the public, assuming that banks always strive to remain loaned up.
c. Estimate the magnitude of the change in the money supply resulting from the Fed's open market operation? Explain.
d. Briefly explain how you would expect the change in the money supply to affect employment, output and prices.

4.* In developing the IS curve in Chapter 10.2.2 it was assumed that the foreign trade balance, $X - M$, was exogenous but investment was sensitive to the rate of interest. In fact, imports may depend on the level of income, as was argued in exercise 1 of Chapter 9. But for that exercise investment was treated as exogenous.

Change that model in exercise 1 of Chapter 9 by making investment depend upon the rate of interest. Now derive the IS curve for this new model. How does the dependence of imports on the pace of economic activity affect the IS curve? Does this change strengthen the impact of interest rate changes on economic activity? Does it change the impact of changes in the level of government spending? Explain.

11

Dynamics, Expectations and Inflation

11.1 Introduction

This chapter looks at the process by which markets adjust to equilibrium. Both micro and macro policy issues are addressed, but the common thread uniting the diverse applications of this chapter is our concern with the dynamic processes of adjustment. We start by looking at a simple model purporting to explain how an agricultural market may have an inherent tendency to generate price fluctuations as it adjusts toward equilibrium. Then we shall look at the activity of speculators, finding that contrary to conventional wisdom they may help stabilize markets. We will also look at the consequences of government attempts to stabilize commodity markets. Later we will turn from micro to macro economic issues. We shall take a fresh look at the question of whether there exists an inflation unemployment tradeoff — is it possible to achieve low unemployment by putting up with a bit more inflation? Finally, we shall address in the light of a variety of dynamic complications the task of prescribing appropriate monetary policy.

11.2 Market dynamics

11.2.1 *The cobweb model*

Some critics of the market mechanism object that the price system is inherently unstable. For example, defenders of fixed foreign exchange rates have sometimes argued that in the absence of government intervention, currency markets will be subject to considerable fluctuation, as was displayed on Figure 1.6. As another example, agricultural interests, abroad no less than in the United States, frequently argue that government intervention — price supports or subsidies — is required in order to protect not only farmers but also their customers from the inherent instability of agricultural markets. In part, volatile prices may be the natural consequence of the vicissitudes of nature. When droughts, floods, disease or locusts cause crop failures, prices jump. But will markets fluctuate in the absence of such external shocks? In an article published in the 1930s on "The Cobweb Theorem," Mordecai Ezikiel presented a simple model suggesting that agricultural markets have an inherent tendency to fluctuate even in the absence of nature's disturbances.[1]

The key feature of the cobweb model is the assumption that production of agricultural commodities takes time. Farmers must decide how much to plant in the spring, long before they know the price p_t that will be offered by the market at harvest time. Let $\widehat{p_t}$ denote the anticipated price in year t and suppose that the market demand and supply curves are linear:

$$Q_t = d_0 - d_1 p_t \,, \tag{1}$$

$$S_t = k_0 + k_1 \widehat{p_t} \,. \tag{2}$$

[1]Mordecai Ezikiel, "The Cobweb Theorem," *Quarterly Journal of Economics*, 1938.

Here Q_t is market demand and S_t is supply, both measured in millions of bushels.

Everything depends on how the expected future price $\widehat{p_t}$ is determined. Ezikiel assumed that $\widehat{p_t} = p_{t-1}$. That is to say, farmers expect the price to remain the same as last year. Then substituting into (2) we have for supply:

$$S_t = k_0 + k_1 p_{t-1}\,. \tag{3}$$

What happens is illustrated by the numerical example recorded on Table 11.1, which uses demand equation $Q_t = 240 - 80p_t$ and supply equation $S_t = -10 + 70p_{t-1}$. Since the price in 1900 had been only 50¢, farmers plant only enough seed in the spring of 1901 to produce $S_1 = -10 + 70 \times 0.5 = 25$ million bushels. But since the inverse demand curve is $p_t = 3 - 0.0125Q_t$, when 25 million bushels are brought to market in the fall, consumer demand pushes the price up to \$2.69, as recorded on the 1901 row of the table and indicated by point a on Figure 11.1. Farmers, encouraged by the high price, plant enough seed the following spring to grow 178.1 million bushels, which is point b on the graph. The price plunged to 77¢ at point c on the graph when this huge crop was brought to market. The low price led farmers to cutback production to 44.1 million bushels the following year. Figure 11.2 reveals that the saw-tooth cycles of erratically fluctuating prices and quantities continue indefinitely, but the amplitude of the movements become smaller and smaller as the system converges toward the equilibrium where the demand and supply curves intersect. In the long run this stable system converges to the equilibrium price of \$1.67 and output of 106.7 predicted by the static supply and demand model.

Table 11.1. The cobweb model.

t	p_{-1}	S	p	D	$R = pq$
1900			0.50		
1901	0.50	25.0	2.69	25.0	67
1902	2.69	178.1	0.77	178.1	138
1903	0.77	44.1	2.45	44.1	108
1904	2.45	161.4	0.98	161.4	159
1905	0.98	58.8	2.27	58.8	133
1906	2.27	148.6	1.14	148.6	170
1907	1.14	70.0	2.12	70.0	149
1908	2.12	138.7	1.27	138.7	176
1909	1.27	78.6	2.02	78.6	159
1910	2.02	131.2	1.36	131.2	178

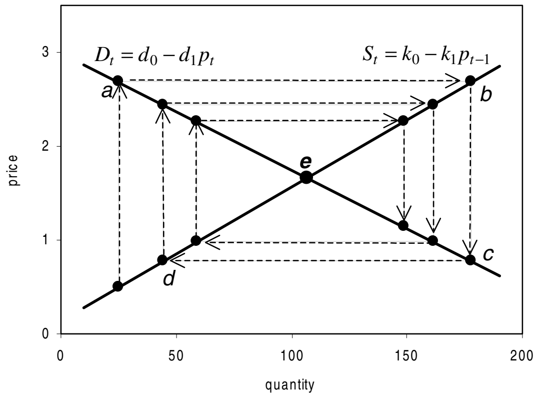


Fig. 11.1. The cobweb
The cobweb model assumes that farmers base their spring planting decision on the assumption that this year's price will be the same as last year's. A low price last fall leads farmers to plant a very small crop, which at harvest time leads to a high price at point a. Anticipating a high price, farmers plant a huge crop expecting to end up at point b. But the large crop pulls the price down, causing farmers to wind up with a low price at point c. The cycle persists, but converges to equilibrium point e.

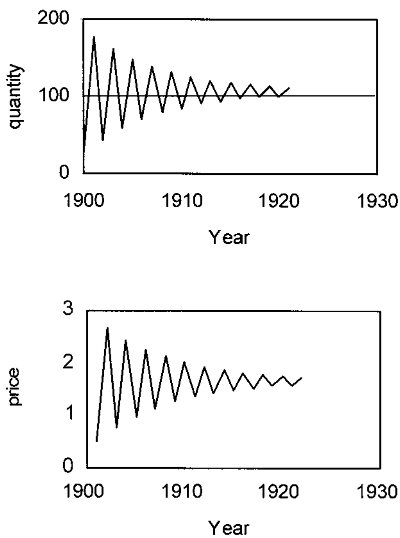


Fig. 11.2. Price and output fluctuations
The cobweb model generates saw tooth fluctuations in quantity and price.

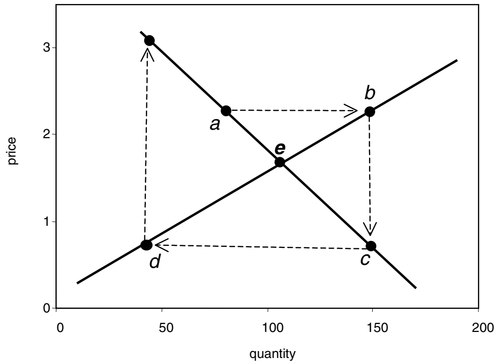


Fig. 11.3. Explosive cobweb
Unlike the cobweb on Figure 11.1, this one spirals outwards in an explosive cycle instead of converging to equilibrium! Why is this cobweb unstable? Is it because the demand curve is steeper than the supply curve?

The first point to note about this simplified dynamic process is that the equilibrium price p^e and output Q^e are precisely the values predicted by the static demand and supply apparatus, as can be verified by substituting p^e for p_t and $\widehat{p_t}$ in equations (1) and (2). Second, while the cobweb was stable for the particular numerical parameter values chosen for illustrative purposes, the model is capable of generating explosive oscillations, as shown by Figure 11.3. Instead of converging to equilibrium, the oscillations in prices and output become larger and larger.[2] We will find that the system will converge to equilibrium if and only if $k_1 < d_1$; i.e., the supply curve must be steeper than the demand curve.

To verify the condition for stability, note that each fall when the harvest S_t is brought to market, demand must adjust so that the market clears. Therefore, we must have $S_t = Q_t = d_0 - d_1 p_t$. Hence the price in year t is given by the inverse demand function

$$p_t = \frac{d_0}{d_1} - \frac{S_t}{d_1}\,. \tag{4}$$

Substituting (3) into (4) yields an equation explaining the evolution of prices through time:

[2]The cobweb model is an example of a first-order linear difference equation. Appendix 11.1 explains how to solve models of this class.

$$p_t = \frac{d_0}{d_1} - \frac{k_0}{d_1} - \left(\frac{k_1}{d_1}\right) p_{t-1} \,. \tag{5}$$

To solve this first order linear difference equation, note that the equilibrium price, p^e, must satisfy this equation; i.e.,

$$p^e = \frac{d_0}{d_1} - \frac{k_0}{d_1} - \left(\frac{k_1}{d_1}\right) p^e \,. \tag{6}$$

Subtracting (6) from (5) yields the homogeneous first order difference equation

$$p_t - p^e = -\left(\frac{k_1}{d_1}\right)(p_{t-1} - p^e) \,. \tag{7}$$

For $t = 1901$ we have $p_{1901} - p^e = -(k_1/d_1)(p_{1900} - p^e)$. Multiplying by $-k_1/d_1$ yields $p_{1902} - p^e = (-k_1/d_1)^2(p_{1900} - p^e)$. Indeed, since equation (7) holds for all t, we have by repeated substitution

$$p_t - p^e = \left(\frac{-k_1}{d_1}\right)^{(t-\tau)} (p_\tau - p^e) \,, \tag{8}$$

where τ denotes the initial year, which for our example is 1900. Prices will converge to equilibrium price p^e for any initial condition p_τ only if $(-k_1/d_1)^t$ approaches 0 as $t \to \infty$, which means that $(-k_1/d_1)$ must be less than unity in absolute value. Since k_1 and d_1 are both positive, equation (8) establishes that the cobweb model is stable if and only if $k_1 < d_1$. What is the economic implication of this stability condition? Since k_1 and d_1 are the reciprocals of the slopes of the supply and demand curves, we can say that the market is stable only if the supply curve is steeper than the demand curve. That is to say, supply must be less responsive than demand to changes in price.

Critique

It is essential to note that a fundamental limitation of this model is its reliance on the grossly unrealistic assumption that farmers never learn from repeated experience how prices oscillate. It seems obvious to expect that some farmers would, sooner or later, come to recognize from historical experience that if prices were low last period they should produce more in the following year because prices are likely to be especially high. Conversely, if prices were high last year they should produce only a small crop because prices are likely to be low in the following year.

While the cobweb model is of interest because it provides a simple introduction to dynamic models and shows that agricultural markets could conceivably be prone to cyclical movements even if harvests were not subject to the vagaries of the weather, it is grossly unrealistic because it assumes that farmers do not learn from experience in forming their expectations about future price movements. Later we shall consider a more interesting theory of expectations. But first we must consider how speculators affect the stability of markets.

11.2.2 *The corn story: Government stabilization versus speculation*

Speculation is the purchase of assets for later resale in the hope of profiting from an anticipated price change. Speculators buy assets not for use but with the anticipation of selling them in the future at a higher price. The asset may be real estate (e.g., Florida land or a San Francisco apartment house) or commodities (e.g., wheat, heating oil, pork bellies, gold) or financial (e.g., corporate stock or foreign currency).

Professional speculators are no more popular than landlords. Critics object that speculators are parasites who waste their energies in unproductive activity — they don't produce anything useful! Worse, speculators stand accused of rocking the economic boat by destabilizing markets. When the Mexican peso and the Indonesian rupiah collapsed in value in the foreign exchange markets in the late 1990s, many government officials and some commentators in the financial press were quick to blame speculators for the problem. When the price of copper, silver, government bonds, and the stock market move erratically it is easy to blame the speculators.

Contrary to the conventional wisdom, many economists feel that speculators get a bum rap. Successful speculators may actually serve a useful function by stabilizing markets! Markets inhabited by speculators may fluctuate a lot, but this does not mean that speculators cause the fluctuations. Quite the contrary, the argument goes, markets would be subject to even more serious fluctuations if it were not for the efforts of successful speculators.

A simple numerical example will suffice to illustrate the basic principles involved.

Case 1: Free markets

Suppose that the demand for corn in Never-Never Land is given by the equation

$$q = 200 - 50p\,, \tag{9}$$

where q is millions of bushels and p is price per bushel. Suppose, only for simplicity, that there is a drought every other year — even numbered years are good crop years but every odd year is bad. Farmers plant the same amount each spring (no cobweb here), but total farm output is only 40 million bushels in a bad crop year versus 140 million in a good year. Farmers complain that whenever they get a good crop the bottom drops out of the market and prices fall to the point where farm revenue is not that much higher than in bad crop years. The outcome is recorded as Case 1 at the top of Table 11.2.

Table 11.2. Hypothetical corn market.

	Good Year	Bad Year	Average
Annual Output (millions of bushels)	140	40	90
Case 1: Free Market			
Price	\$1.20	\$3.20	\$2.20
Farm Revenue (\$millions)	168	128	148
Consumption (mill of bushels)	140	40	90
Case 2: Government Price Supports			
Price	\$2.20	\$2.20	\$2.20
Farm Revenue (\$millions)	308	88	198
Consumption (mill of bushels)	90	90	90
Government Purchase (mill of bushels)	50	−50	0
Cost to Government (\$millions)	110	−110	0
Case 2b: Government Price Supports at "parity"			
Price	\$2.30	\$2.30	\$2.30
Farm Revenue (\$millions)	322	92	207
Consumption (mill of bushels)	85	85	85
Government Purchase (mill of bushels)	55	−45	5
Cost to Government (\$millions)	126.5	−103.5	11.5
Case 3: Speculation			
Price	\$1.70	\$2.70	\$2.20
Farm Revenue (\$millions)	238	108	173
Consumption (mill of bushels)	115	65	90
Speculator Purchases (mill of bushels)	25	−25	0
Return to Speculator (\$millions)	−\$42.50	\$67.50	\$12.50

Case 2: Government price supports

Suppose that the legislature, concerned about the plight of the farmers, instructs the Department of Agriculture to stabilize prices. The government will offer to buy corn from anyone who is willing to sell it at $2.20 per bushel.[3] And the government also offers to sell corn at $2.20 a bushel to all comers. In years of bounty, the government's offer to purchase corn at $2.20 will pull the price up to this level. As indicated on the table, the government will have to buy 50 million bushels of corn in order to support this target price. The corn will be added to the government's stockpile where it will be available in years of drought. When the drought does come the small crop will threaten to drive up the price of corn. As soon as the price starts to rise above $2.20, the public will start buying from the government at the offer price of $2.20 a bushel. Thus the government's offer to buy or sell corn at $2.20 effectively stabilizes the price.

Who will gain and who will lose from this price-support policy? As can be seen from Case 2, the government's stabilizing program has not stabilized farm revenue, which in percentage terms fluctuates more violently than before! Averaging good years with the bad, however, farmers are ahead for farm revenue is up substantially, averaging $198 million with price supports rather than the $148 million average that farmers received in the absence of government intervention.

Case 2b: Fair price for farmers

Farm blocs often have substantial political clout. Suppose that the farm lobbyists convince the Congress to support a "fair price" of $2.30 a bushel. This is Case 2b on the table. It is clear from the table that average farm revenue is up as a result of this more favorable legislation. Equally obvious, a higher support price means that consumers are hurt. And because the price is above the free market average, the government is adding more to its stockpile in years of surplus than it is selling in drought years. Over the years the government's hoard of corn grows and grows, leading to mounting storage costs and growing concern about the unbalanced farm policy.

[3]Farm price supports in the United States were discussed in Chapter 3.6.3.

Case 3: Speculation

For Case 3 there is no government intervention, but there is a smart speculator. Speculators make their profits by buying when prices are low in order to sell when prices are high. In this example, our speculator buys 25 million bushels in good crop years, which pulls the price up towards its long-run equilibrium value. But when the corn is unloaded in high price years the speculator's selling pushes the high price down toward equilibrium. Thus our brilliant speculator not only makes a profit. Our successful speculator substantially reduces the range of price movements — she helps stabilize the market!

Note that our speculator would make zero profits under two conditions. Profits would be zero if she did not enter the market at all, purchasing zero in both good years and bad. They would also be zero if our speculator eliminated all price oscillations by purchasing 50 million bushels in good years and selling it off in bad years, for then the movements in price required for profits would be eliminated.

Unfortunately for our speculator, there is nothing to keep others from also speculating in this market. And the competition among speculators will lead to a larger and larger carryover from good years to bad, which reduces further the price movements. The end result of free entry will be to drive economic profit to zero! The price spread will be virtually eliminated, except for the margin that may be required to cover the cost of storage, the internal cost of funds invested in the grain inventory, and other expenses incurred by the speculators.

Stabilizing speculation

One moral of this numerical example is that successful private speculators help stabilize markets, contrary to the conventional wisdom. In our numerical example, prices still fluctuate and there is a sizable profit to be reaped by anyone with the courage to speculate in corn. If more speculators enter the market, and if they indeed succeed in buying when prices are low and selling when they are high, they will not only enjoy a profit. Their purchases will pull low prices up and their sales will push high prices down. The prices fluctuate as much as they do in Case 3 because there were not enough speculators. If the cost of storing corn from one year to the next is negligible, speculators can make profits as long as prices fluctuate. Even when uncertainty and storage costs are taken into account, speculators may substantially reduce the range of price fluctuations. Thus the availability of people of money with an inclination to speculate in the market will tend to stabilize prices close to the same level that was generated by government price supports, Case 2b. A primary effect of government intervention in the market to stabilize prices may be to crowd private speculators out of the market by making speculation less rewarding.

Critique

The simplified cobweb example was presented for the unrealistic case in which the market swings are perfectly predictable, every other year being a bad crop year. In practice, of course, the vicissitudes of the weather and other shocks prevent precise predictions of crop yield, although the probability of a crop failure may be calculated from the historical record with some precision. The presence of uncertainty is not enough to undermine the validity of the argument. Speculators only make money when they succeed, more often than not, in buying low and selling high; and buying low tends to raise the price in depressed markets and selling high tends to depress high prices. Thus a successful speculator, even in the case of uncertainty, tends to stabilize the market. Unsuccessful speculators, the argument goes, will have to eat their losses and move on to some other line of work. Thus the survival of the fittest insures that the remaining speculators will be helping to stabilize the market.

Of course, when unskilled individuals try trading they are likely to make too many mistakes. The argument that speculation tends to stabilize markets rests on the assumption that unsuccessful speculators will abandon the activity once they realize they are losing their shirts. Even so, speculation may still destabilize if there is a perpetual supply of relative uninformed or inexperienced speculators who retire on losing money only to be replaced by a new generation of inferior speculators (suckers).

11.2.3 *Options, forward markets and hedging*

How can one cope with unpredictable price movements brought on by crop failures, political surprises or other unexpected shocks? Fortunately, a variety of different types of contracts help market participants cope with uncertainty:

Options

An option is the right to execute a specific transaction at a future date.

- A manufacturer considering a real estate parcel for a new warehouse does not want to purchase the property pending the resolution of various zoning issues and the determination of the expense of hazardous waste removal; but our manufacturer does not want to undergo the expense of evaluating these potential problems only to find that the seller has raised the price or sold the property to someone else. The seller is reluctant to take the property off the market while the manufacturer decides whether to go ahead with the transaction. The impasse may be resolved if the manufacturer, for a fee, is able to obtain from the seller an *option* to buy the property at a particular price. If the prospective buyer decides, after all, not to buy the property, the fee is forfeited as compensation to the seller for holding the property off the market.
- A *stock option* is the right to purchase a share in the issuing company's stock at a specified exercise price on or before a specified date. Some speculate by purchasing options in the marketplace. Many companies provide key employees with stock options as part of their deferred compensation package. Assuming that the company prospers and the market value of its stock rises in value above the exercise price, the employee will eventually cash in the shares and pocket the excess of the market value over the exercise price, less income taxes. The employees will have to pay personal income tax but not Social Security or Medicare taxes on their stock option gains. The company may mention the options in a footnote in its annual report but is not legally required to count it as an expense in reporting profits in the year the option is issued, although some do so voluntarily. Rewarding employees with stock options instead of salary allows both the employer and the employee to postpone tax payments — it's an interest free loan from the Internal Revenue Service.

Forward contract

Closely related to an option, the forward contract is an agreement to deliver, or to accept delivery of, a specified quantity of an item at a specified price on a particular future date.

- For example, farmers must worry not only about how the weather will affect their harvests. They also worry about what price their crop will bring when they take it to market in the fall. Consider a wheat farmer who is planting crops in early May. No one knows for sure what price the wheat will be selling for at harvest time. It may be possible for the wheat farmer to "lock up" the price of wheat in the spring by signing a contract to deliver a specific quantity of wheat in September at a specified price.

Who will be willing to take the other side of this transaction, committing themselves in advance to buy the wheat at the specified price? That is to say, who would promise in the spring to buy the grain at a particular price rather than waiting to see what the price turns out to be at harvest time?

- Perhaps a bakery firm, having made a commitment to deliver bread to a college dining hall at a particular price, is concerned about what will happen to the price of flour. If our baker buys wheat for future delivery in October it will have hedged its commitment. If the price of wheat climbs as a result of a crop failure, the profit made from the increase in the price of the wheat purchased for delivery in October will offset the increase in the price of flour that the baker will have to buy to make the bread he has committed to deliver in the fall. Thus the baker who purchases such a contract will be protected against the uncertainty of future fluctuations in the price of wheat.

Both sides to the transaction will have reduced their exposure to risk. They are said to have hedged their positions; that is to say, *hedging* is the act of undertaking transactions to offset risk.

Futures

While forward contracts date from the very beginning of commerce, future markets are characterized by standardized contracts that are bought and sold on organized markets. The seller of a futures contract has the right and obligation to deliver the specified item at the date in the future that is specified in the contract. The buyer has the obligation to purchase the commodity on the specified date. The price of the contract equates demand with supply, as determined in the market place — e.g., at the Chicago Board of Trade. Futures Trading got its start with the founding of the Chicago Board of Trade in 1848. On the financial pages of today's *Wall Street Journal* are posted the prices of futures in a host of commodities, including heating oil No 2, gold, copper, soybean oil, frozen orange juice and flax seed. One can also buy or sell for future delivery Japanese Yen, Eurodollars, treasury bills and other financial instruments. But the transactions are in sizable lots; for example, to trade in heating oil you must buy or sell in 42,000 gallon lots, which obviously is beyond the reach of an individual home owner.

Hedgers, speculators and other traders may choose "settlement by offset" rather than taking actual delivery of the commodity they have purchased in the futures market. That is to say, if the spot price turns out to be below the price specified in the future contract, the buyer pays the holder of the contract the difference. If, however, the spot price is above the future price, it is the seller who must pay the difference.[4]

Application

Middletown's Didato Heating Oil Company allows its customers to prepay for oil. Customers can protect themselves against price increases during the coming heating season by paying for a fixed quantity of heating oil in advance of the heating season at an agreed price per gallon. This is a convenience for the customers, who know precisely how much they should budget per gallon of anticipated oil consumption. But how can Didato Oil protect itself from an increase in the price of oil? If it has the storage capacity, it could purchase the oil in advance to cover its prepayment commitment. Alternatively, it could hedge its commitment by purchasing heating oil for delivery on the futures market. The other side of the contract, the sale of futures, might be undertaken by an oil refiner interesting in locking up the price of oil that it expects to be delivering in the months ahead. Or it might be picked up by a speculator who is willing to gamble that the price of heating oil will not increase above the quoted price. The market price of the oil contract will be that price which equates demand for futures with the supply.

Efficient markets

If you think the market price for the futures contract is too low, if you think you are smart enough to know something that is not recognized in the market place, then you can put your money on the line by buying futures. If you think the price quoted in the futures market is too high, then sell futures. Of course, many others may also put their money where their mouth is. The market equilibrium price, equating demand and supply, is sort of a consensus forecast of what the future will bring. Economists

[4]Options are similar but not identical to futures contracts. A call option gives the right to buy at the specified price at the specified date, but if the spot price turns out to be below the option price the holder of the option does not have to exercise it. Similarly a put option gives the right to sell at a specified price on a certain date. In 1972 the Chicago Board Options Exchange standardized the terms of the contracts and introduced pit trading in options. Two types of options are distinguished: The "American option" gives the right to buy at the strike price anytime before the expiration date while a "European option" can be exercised only on the expiration date.

who study futures markets usually find that they tend to be efficient or nearly so. An efficient market is a market in which the price established by the market incorporates all the information that is currently available.[5] The stock market provides another example of an efficient market.

11.2.4 *The stock market*

In 1792 twenty-four prominent New York business leaders, meeting under a buttonwood tree on what is now 68 Wall Street, signed an agreement to trade securities in common. This historic pact marks the establishment of the New York Stock Exchange. Nowadays, stocks of more than 2,800 corporations are listed on the New York Stock Exchange. On a typical business day, a billion or more corporate shares worth more than $40 billion may trade hands on the NYSE. Many other corporations are listed on the American Stock Exchange. The NASDAQ (National Association of Securities Dealers) lists the shares of more than 5,000 corporations. Regional markets in Philadelphia, San Francisco and elsewhere trade stocks that are of more interest to regional than national investors.

The Dow Jones Industrial Average summarizes what has been happening to stock values. Its fluctuations are reported minute by minute on the web, are listed in the financial press, and are often cited on the evening television news. When Charles H. Dow introduced his stock market average on May 26, 1896, he used a simple pencil and paper procedure for constructing the index: he added up the prices of 11 major stocks and divided by 11. Today, the Dow Jones average is based on 30 stocks representing about a fifth of the $8 trillion-plus market value of all U.S. stocks.[6] A competitor to the Dow, the S&P 500 stock market index, provides a much more comprehensive measure of stock market performance because it is based on the performance of 500 stocks. But when people talk about stock market gyrations they are usually referring to the Dow — it is the traditional measure of stock market performance.

The plot of the Dow on Figure 11.4 provides a graphic summary of stock-market performance. While stocks have, on average, had a strong

[5] The term "efficient markets" has a meaning in the field of finance that is quite different from the use of the word "efficiency" when talking about market efficiency, as in Chapter 2.5.2.

[6] Calculating the average is a bit more complicated than simply dividing the sum of their prices by the number of stocks, in part because of the need to keep the index from being affected by stock splits.

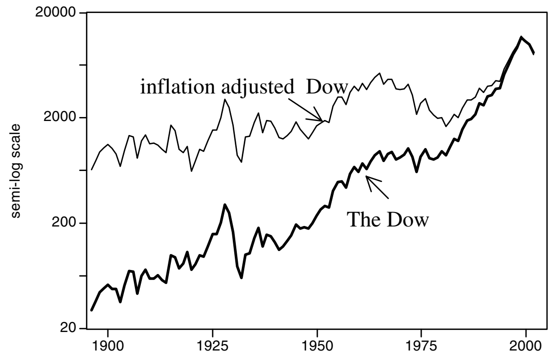


Fig. 11.4. The Dow, nominal and real
The Dow, nominal, and after deflation by the consumer price index (2000 = 100, annual observations).
Source: Dow Jones Indexes, http://www.djindexes.com/jsp/index.jsp

upward trend, even when corrected for inflation by dividing by the CPI, there obviously have been both good years and bad. Purchasers of a stock must worry that they are exposed to the risk of a capital loss — when the time comes to sell the stock its price may be substantially below what was paid for it. This risk is more dramatically revealed by Figure 11.5, which reports the annual percentage change in the Dow.

More detail is provided by Table 11.3, which compares the total return offered on the U.S. stock markets with that which could have been obtained by investing in government securities. The total return includes the capital gain plus the dividends paid to the investor. Consider someone who had the perspicacity to invest \$1,000 in the 30 stocks of the Dow in 1972 and then reinvested all the dividends back in the market. As is clear from Figure 11.4, that investment would have looked a bit shaky during the next few years. But the total return figure for the Dow of 14.1% per annum from 1972 to 1999 reported on Table 11.3 implies that the \$1,000 would have grown in value to \$35,214 by 1999. Taking inflation into account, that brilliant \$1,000 investment would have grown to almost \$8,000 in dollars of constant purchasing power! In contrast, a cautious investor who purchased \$1,000 in U.S. Treasury bonds in 1972 would have accumulated only \$9,510, or \$2,402 net of inflation by the turn of the century.

Some gain and some lose in the stock market, but it is clear from Table 11.3 that on average those who chose to invest in the stock market rather than holding bonds or other debt instruments have been rewarded over

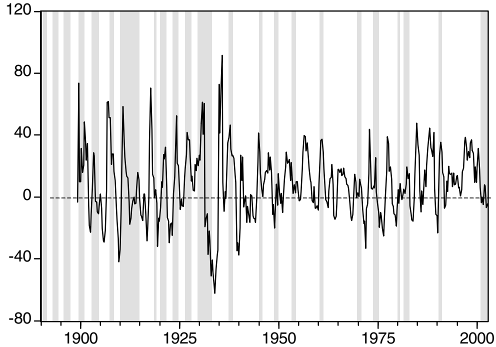


Fig. 11.5. The Dow, annual percentage rate of change
Source: See Fig. 11.4.

Table 11.3. Historical returns on U.S. financial assets (% per annum).

Period	Nominal returns: Stocks	Nominal returns: U.S. Treasury Securities 20-year bonds	Nominal returns: U.S. Treasury Securities 1 month T-bills	Real returns (net of inflation): Stocks	Real returns (net of inflation): U.S. Treasury Securities 20-year bonds	Real returns (net of inflation): U.S. Treasury Securities 1 month T-bills
1802–1999	8.4	4.8	4.3	7.0	3.5	2.9
1926–1945	6.5	4.7	1.1	6.4	4.6	0.9
1926–1999	11.3	5.1	3.8	7.5	1.9	0.7
1945–1999	13.3	5.4	4.7	8.4	1.1	0.5
1972–1999	14.1	8.7	6.8	8.0	3.3	1.5

Source: Ravi Jagannathan, Ellen R. McGrattan and Anna Scherbina, "The Declining U.S. Equity Premium," Federal Reserve Bank of Minneapolis *Quarterly Review*, Fall 2000, p. 17.

the long run with a much higher rate of return than could be earned by investing in fixed income securities, such as government bonds. This excess return is known as the *equity premium.* But in the short run anything can happen. While the Dow started off the new century right by reaching an all time high of 11,722.98 on January 14th, 2000, it dropped to 7,286.27 in October of 2002, a fall in value of 38%. Only time will tell whether investors should count on the equity premium continuing in the years ahead.

Diversify

The graphs and tables on market performance report only the averages. Individual stocks are subject to much greater fluctuation. A *diversified* *portfolio* composed of many stocks will offer protection from idiosyncratic movements in the price of individual stocks. Let us consider a grossly simplified example explaining the advantages of portfolio diversification:

Example:

Suppose you have \$90 to invest. You could purchase shares in Stock A, which costs \$45 per share. You judge that there is a 50% chance that a share of this stock will be worth \$75 at the end of the year, but there is also a 50% chance that it will be worth only \$25. Stock B offers the same prospects as Stock A and also costs \$45 per share. But the outcomes are totally independent of each other; in particular, the probability that a share of Stock B will be worth \$75 is completely unrelated to what happens to Stock A.

With your \$90 you could buy two shares of Stock A or two of Stock B; but you could also buy one share of each stock. Which portfolio should you buy?

Table 11.4. Diversification example.

Possible Portfolios	Possible Outcomes	Probability
2 shares of A	\$50	1/2
	\$150	1/2
1 share A & 1 of B	A & B	
	$25 + 25 = \$50$	1/4
	$25 + 75 = \$100$	1/4
	$75 + 25 = \$100$	1/4
	$75 + 75 = \$150$	1/4
2 shares of B	\$50	1/2
	\$150	1/2

Analysis:

Table 11.4 shows possible outcomes for each of the three alternative portfolios. As can be seen from the table, regardless of which of these portfolios you buy, on average you will end up with \$100 — each of these portfolios has an expected value of \$100. But the likelihood of extreme outcomes is reduced if you buy the mixed portfolio. If you were to buy two shares of stock A, there would be a 50% probability that you will end up with only \$50. The same loss is equally likely if you buy 2 shares of Stock B. If you buy the diversified portfolio —

> one share of each stock in this example — you run only a 25% probability of ending up with only $50; on the other hand, you also have only a 25% probably of ending up with $150. The advantage of the diversified portfolio is that it lowers the likelihood of extreme outcomes. If you want to limit your risk, you will buy the diversified portfolio. But a risk lover might prefer to purchase two shares of Stock A (or two of Stock B) in order to have a higher probability of ending up with $150.

The moral of this simplified example is clear: risk adverse investors should not put all their eggs in one basket.[7] Holding a diversified portfolio of assets reduces the risk of extreme outcomes.

Many investors are "stock pickers," trying to discern which corporation's securities are likely to increase the most in value. This effort can be time consuming and hard work. As an alternative to trying to pick your own winning stocks, one may purchase shares in a mutual fund. The mutual fund will pool the funds you have given it to invest with the funds provided by others and invest them in the stock market. Most funds are actively managed, trading one stock for another in a continuing effort to obtain a higher return. The hope is that the professional investment managers at your mutual fund will succeed in picking stocks that will yield high rates of return.

Index funds are an alternative type of mutual fund. Instead of being actively managed by stock pickers, indexed funds buy and hold a representative portfolio of stocks. This strategy was pioneered by the Vanguard Group, an investment management company that began selling their S&P 500 index fund in 1972. This fund has essentially the same composition as the S&P 500 stock market index itself, which means that it fluctuates with the market. Purchasers of this fund are locked in to earning the average rate of return obtained by investors in S&P 500 stocks, less modest management expenses. In any given year, some actively managed mutual funds do better than Vanguard's indexed fund, but the majority usually do worse, in part because they have the added expense of hiring managers to pick the stocks.

---

[7] This is precisely what James Tobin, who pioneered research in this area, told reporters he had proved when he was asked to explain the contribution that won him the Nobel Prize.

Random walk hypothesis

The dominant model of stock market behavior is Brownian motion, otherwise known as the random walk hypothesis.[8] This is the proposition that changes in stock market prices are purely random and unpredictable. Why should stock prices appear to move in a random fashion? The answer is that there are many sophisticated investors studying the market and attempting to profit from price changes. Their trading activity takes advantage of all available information. And their decisions to buy or sell stocks in the aggregate affect stock prices. Their buy and sell decisions generate an efficient market. If a stock is under priced, speculators buy it, and the buying activity will push up the price. Because today's price effectively sums up all the available information, subsequent changes in the stock market must be unpredictable; hence future price movements must be purely random. Because of the costs of executing transactions, brokerage fees and so forth, there may be minor departures from random behavior, but they are so small as to not provide an opportunity for profit once trading costs are taken into account.

The hypothesis that stock market prices move randomly or nearly so, has been subject to considerable debate. Princeton Professor Burton Malkiel presents the argument in favor of the hypothesis in *A Random Walk Down Wall-Street*, a national best seller first published in 1973. A number of empirical studies questioning the validity of the hypothesis are presented in *A Non-Random Walk Down Wall Street* by Andrew W. Lo and A. Craig MacKinlay, 1999. The reader who wishes to pursue this topic further would do well to consult these two volumes. It is now time for us to move on to a consideration of dynamic issues in macroeconomics.

11.3 Inflation, unemployment and the Phillips curve

It is not true that economists do not know how to cure unemployment, and it is not true that we do not know how to control inflation. The problem is that the remedies for unemployment have as a side effect an unfortunate tendency to generate inflation. And the remedies for inflation have as an

[8] Brownian motion is named after biologist Robert Brown, who observed with his microscope in 1827 that suspended particles of pollen appeared to move in a totally unsystematic or random manner. Such movements are referred to as a random walk, like that of a drunken sailor who can still stagger but whose direction of movement is unpredictable.

unfortunate side effect a tendency to generate unemployment. The problem is not so much that there are no cures as that the side effects of the remedies are sometimes worse than the diseases they are designed to cure.

11.3.1 *The promised tradeoff*

The Phillips curve purports to clarify the nature of the relationship between unemployment, plotted on the abscissa, and inflation, plotted on the ordinate. The relationship is named after Professor A. W. Phillips [1914–1975] of the London School of Economics, who published in 1958 a pioneering study based on British data that revealed an inverse relationship between the unemployment rate and the rate of increase in the money wage rate. The Phillips curve is not to be confused with the aggregate supply curve, explained in Chapter 10. The aggregate supply curve relates the long run equilibrium *level* of the price index to the level of output. The Phillips curve is a statement about how rapidly wages and/or prices *change* as a function of the unemployment rate.[9]

Figure 11.6 reproduces an estimate of the Phillips curve for the United States that MIT Professors Paul A. Samuelson and Robert M. Solow presented at the 1959 meetings of the American Economic Association.[10] Samuelson and Solow suggested that their empirical Phillips curve, "as roughly estimated from the last twenty-five years of American data," showed the menu of choice between different degrees of unemployment and price stability. Point A on the graph, corresponding to price stability, is seen to involve about 5 1/2% unemployment, whereas point B, corresponding to 3% unemployment, involves a price rise of about 4 1/2% per annum ..." Both Samuelson and Solow subsequently won the Nobel Prize in economics, but not for their Americanized Phillips curve.

[9]Phillips' pioneering study, "The Relation Between Unemployment and the Rate of Change of the Money Wage Rate," *Economica*, 1958, was based on data for the United Kingdom covering the years 1861 through 1957. Of New Zealand birth, Phillips shipped out on as a deckhand on a merchant ship when he couldn't find other work in the depths of the Great Depression. He studied to be a radio operator on his first trip to sea and on a subsequent trip, while serving as radio officer; he studied electrical engineering. In subsequent economic research he drew on his electrical engineering background by applying control theory to clarify the task of stabilizing dynamic models of the economy.

[10]"Analytical Aspects of Anti-Inflation Policy," *American Economic Review*, May, 1960.

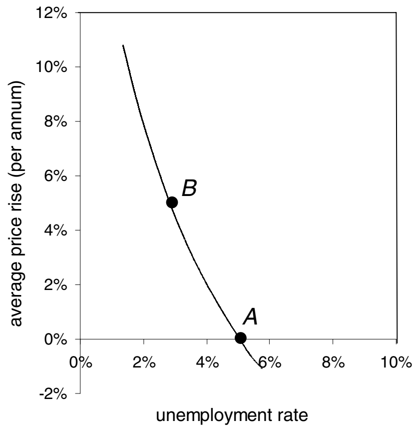


Fig. 11.6. The Samuelson-Solow Phillips curve
The Samuelson-Solow 1959 estimate of an American Phillips curve provided a choice between price stability and 5% unemployment at Point A or 3% unemployment with 4 1/2% inflation at Point B.

Is it really necessary to make such a painful choice between high employment versus low inflation? The basic argument underlying the original Phillips curve is that while unemployment is painful, it does serve to hold worker wage demands in check. When threatened by unemployment, workers refrain from asking for too much at the bargaining table — they worry that if they ask for too big a wage increase their boss may replace them with unemployed workers or, alternatively, be driven out of business by high labor costs. Thus unemployment is said to discipline workers into accepting smaller annual wage increases. At the opposite extreme, if the policy makers push the unemployment rate too low, employers will bid against each other for scarce workers, and the resulting rise in wage costs will ultimately be passed along in higher prices to consumers. More than this, when the unemployment rate is pushed too low, factories are likely to be operating closer to capacity, and producers, finding it difficult to meet orders for their products, will be tempted to raise prices. How rapidly prices will rise depends in part upon how close the economy is to capacity and whether bottlenecks develop in key industries.

When they reported their empirical results in 1959, Samuelson and Solow cautioned that their Americanized Phillips curve did not precisely describe historical experience. They emphasized that in many years the mix of inflation and unemployment rates had been far from their estimated curve.[11] They reported that they were presenting only a "best guess" rather than a precise description of historical experience. They said it was conceivable that the economy "might build up within itself over the years larger and larger amounts of structural unemployment" that would result in "an upward shift of our menu of choice with more and more unemployment being needed just to keep prices stable ..."

11.3.2 *The Phillips curve in practice*

The Samuelson-Solow paper had a key impact on economic policy at the time. The Samuelson-Solow argument provided intellectual support for President Kennedy's program of cutting taxes to get the economy rolling again. Point B on the Samuelson-Solow graph, tolerating moderate inflation in return for a substantial reduction in unemployment, was the policy choice. The expansion of the 1960s might have been even more successful, liberal economists felt at the time, if only Congress had been more receptive to expansionary fiscal policy. Kennedy could not get Congress to pass the massive cut in the personal income tax that constituted a key component of his program for economic expansion. Not until after Kennedy's assassination did the reluctant Congress finally pass the tax cut medicine.

Verification of the inflation-unemployment tradeoff predicted by the 1959 Samuelson-Solow Phillips curve appeared to be provided by the experience of the subsequent decade. Macro economic policy was judged a success in the early 1960s, for the American economy marched up the Phillips curve, tolerating minor increases in the rate of inflation in exchange for a gradual fall in the unemployment rate. The dots that generated the Phillips curve scatter plot on Figure 11.7, one for each year of the 1960s, trace out a negatively sloped Phillips curve that is remarkably close to the predicted relationship. Samuelson and Solow may have been slightly optimistic, for their Phillips curve underestimated the extent of inflation generated at each unemployment rate; but given the carefully qualified nature of their

[11] Two anomalous years, 1955 and 1956, were excused on the grounds that shifts in the composition of output at the end of the Korean War imposed inflationary strains on the economy. They suggested that the substantial increase in money wages in the face of unprecedented unemployment in the mid-1930s may have resulted from the spread of labor unions and the government effort to support rising prices through the National Recover Act.

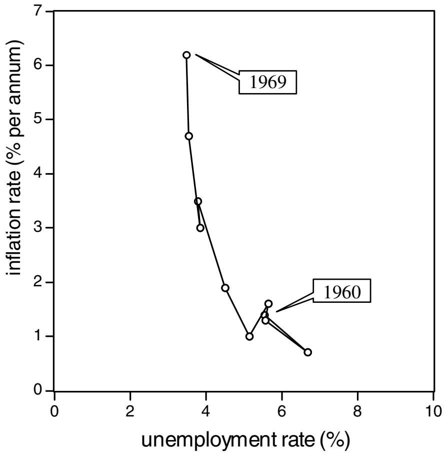


Fig. 11.7. The Phillips curve in the 1960s

presentation, events conformed more closely to their empirical estimate than they had any right to expect.[12]

In subsequent decades the simple Phillips curve turned out to have promised too favorable a tradeoff. Indeed, the 1970s and 80s were characterized by high unemployment coupled with excessive inflation! The promised Phillips curve tradeoff is not to be found in the forty year's of data plotted on Figure 11.8. A review of how inflation and unemployment unfolded during the last four decades of the 20th century will reveal how theory had to be repeatedly modified in response to the stimulus provided by unexpected economic developments that could not be explained, even after the fact, with established theory.

[12]The Phillips curve tradeoff was helped by government policy, so to speak. In the first part of the 1960s, President Kennedy's Council of Economic Advisers attempted to mobilize public opinion in the battle against inflation by promulgating "Wage-Price Guideposts" prescribing a "non-inflationary code of wage and price behavior." The President pushed big steel to role back a substantial price increase that violated the guideposts.

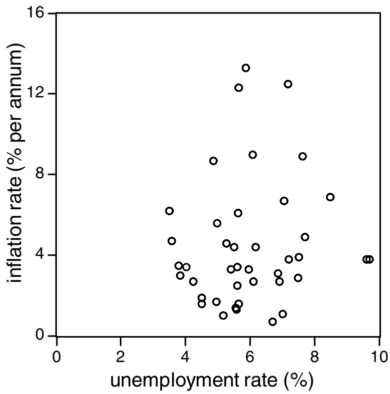


Fig. 11.8. Where is the Phillips curve tradeoff (1960–2000)?

11.3.3 *The Phillips curve boomerang*

Near the end of the decade of the 1960s, it was generally felt that inflation had become a problem of major proportions. The Vietnam War had over-stimulated the economy, pushing unemployment too low and carrying the economy too far up the Phillips curve. It was widely felt by economists at the time that it was necessary to put on the brakes, to slow down the economy, and to deliberately generate unemployment in order to limit the inflation by sliding back down the Phillips curve. On assuming office in 1969, President Richard Nixon, encouraged by his economic advisers, tried to stem the inflation at the cost of a moderate rise in unemployment.

The austere fiscal program did slow the economy. The unemployment rate did increase in accordance with Nixon's "game plan," as he called it, but the painful unemployment medicine did not work. The Nixon administration was plagued with a perverse mix of accelerating inflation coupled with unacceptable high unemployment. As can be seen from Figure 11.9, throughout the 1970s the easy tradeoff promised by the Phillips curve did not hold up. The perverse combination of high unemployment, $U = 7.1\%$, coupled with high inflation, $\dot{p} = 13.5\%$, helped defeat Jimmy Carter's bid for reelection in 1980. In 1982, during President Reagan's first term as president, the unemployment rate hit 9.7%, the highest rate in the United States since the Great Depression of the 1930s. High unemployment rates did eventually contribute to a reduction in inflation, but the massive doses of the unemployment medicine were much more painful than the Samuelson-Solow Phillips curve predicted.

In contrast to the painful 1970s and 1980s, in the last decade of the 20th century inflation averaged less than 4% per annum. The decade started with recession, which was a major factor in preventing George H. W. Bush from winning reelection in 1992. The Clinton years were blessed with a substantial decline in unemployment without serious inflation as the economy enjoyed the longest boom in U.S. history!

What went wrong with the Philips curve? This was a major question for debate among macro economists at the time. There was no shortage of explanations:

Fig. 11.9. The Phillips curve betrayal — the 1970s

Fig. 11.10. Inflation unwinds in the 1980s

Fig. 11.11. Full employment without inflation — the 1990s

1. *Cost-push inflation?* Much of the inflation was said to be "cost-push" in origin. It was generated by the success of the Organization of Petroleum Exporting Countries (OPEC) in pushing through major increases in the price of petroleum during the 1970s. Because oil is a major import, the unprecedented increases in the price of oil generated substantial inflation by pushing up the aggregate supply curve, as was explained by Figure 10.18.
2. *Central bankers run amuck?* An alternative argument blamed the central bankers for allowing excessive increases in the money supply. In terms of Figure 10.20, it was a rightward shift in the aggregate demand curve that generated the inflation. According to this view, if only the Fed and other central banks had kept the brakes on the rate of growth of the money supply, the inflation would have been short lived. If central bankers in Great Britain, Japan, and the United States had kept a tighter reign on the rate of growth of the money supply, the resulting world wide weakness in aggregate demand would have undermined the efforts of the OPEC Cartel to push up the price of oil. If the public had been convinced that the central bankers would hold down the rate of growth of the money supply no matter what the unemployment rate, the public would have recognized that the inflation would not continue, further inflation would not have been anticipated, and the inflation would never have dominated the story of the 1970s and early 1980s.

3. *Shifting demographics?* It was thought that the increase in the female labor-force participation rate (recall Figure 8.3) contributed to the rightward shift of the Phillips curve because women are more likely than men to suffer from unemployment. Women were said to have a higher unemployment rate because (more so in the 1970s and 1980s than today) women drop out of the labor force to have children. Later, when the children are older and the mother reentered the workforce, she was likely to encounter a spell of unemployment before finding appropriate employment. Since women had greater job turnover, they naturally tended to have a higher unemployment rate. This argument makes less sense nowadays than it did then because the female unemployment rate is no longer higher than the male rate.

Two other explanations of the shifty Phillips curve will require more detailed consideration.

4. *Productivity slowdown?* The slowdown in productivity growth pushed the Phillips curve to the right, adding to the inflationary pressure.
5. *Misspecification?* The simple Phillips curve relationship is fatally flawed because it fails to take account of the fact that when the public comes to anticipate inflation they will demand a larger increase in their money wage at any level of unemployment — anticipated inflation shifts the simple Phillips curve.

These last two explanations will receive close attention.

11.3.4 *Productivity, labors share, and the Phillips curve*

The link between the unemployment rate and inflation can be decomposed into a two-step process:

#1 The first step relates the rate of increase in the money wage rate, $\dot{w} = (dw/dt)/w$ to the unemployment rate. When the unemployment rate is low, employers are more likely to acquiesce to the demands of their workers for wage increases — otherwise their workers may move on to other jobs. When jobs are scarce, workers will not be able to bargain as effectively with their employers. That is why A. W. Phillips in his original study of British data postulated that the rate of increase in the money wage rate is inversely related to the unemployment rate. We will let $\dot{w}(U)$ denote the Phillips wage-change unemployment relationship, with $d\dot{w}(U)/dU < 0$. $\dot{w}(U)$ is plotted on Figure 11.12.

#2 For the second step, the extent to which more rapidly rising money wages are in turn translated into faster increases in the general price level depends critically on the rate of change in worker productivity.

One explanation for the shifty link between the unemployment rate, real GDP and the rate of price inflation focuses on this second step, the link between money wage increases and the rate of inflation. A critical concept in this regard is the rate of growth of *worker productivity*. When macro economists refer to productivity they usually mean output per hour of work (i.e., the average not the marginal product of labor), or $y = Y/h$, where Y is real GDP and h is total hours worked. The Samuelson-Solow Phillips curve estimate was based on the explicit assumption that the rate of productivity growth would remain at 2.5% per annum, the rate that had prevailed since World War II. But starting around 1973 the United States, and indeed most industrialized nations of the world, experienced a marked slowdown in the rate of productivity growth. The productivity slowdown itself will be examined in more detail in Chapter 12.4. The immediate task is to show how the slowdown in the rate of productivity growth contributed to the perverse upward shift of the Phillips curve relationship that was portrayed on Figure 11.9.

Recall from Chapter 7.3.3 the discussion of labor's share, which was the fundamental concept stressed by Cobb-Douglas in their pioneering empirical investigation of the aggregate production function. Labor's share is the ratio of labor income to total income,

$$\lambda = \frac{wh}{pY}\,. \tag{10}$$

In this equation the numerator wh, the product of money wage rate w times the number of hours h worked during the year, is labor income. The denominator pY is the value of nominal GDP, the product of price level p times Y (real GDP). Substituting worker productivity $y = Y/h$ into (10) yields $\lambda = w/py$ or

$$p = \frac{w}{\lambda y}\,. \tag{11}$$

To get to an explanation of $\dot{p}$, the rate of inflation, we first take logs to the base e of both sides,

$$\ln p = \ln w - \ln y - \ln \lambda\,. \tag{12}$$

Let $\rho = d \ln y/dt = (dy/dt)/y$, the rate of productivity growth.[13] Then differentiating (12) with respect to t yields

$$\dot{p} = \dot{w}(U) - \rho - \dot{\lambda}\,, \tag{13}$$

where the dots indicate annual rates of changes in the variables and the function $\dot{w}(U)$ is the Phillips wage-change curve relationship showing how the rate of change in the money wage rate is affected by the unemployment rate.

Now labor's share, λ, has been fairly stable over time. Indeed, Cobb-Douglas argued that labor's share was stable because it was technologically determined by their aggregate production function.[14] This suggests that it is not unreasonable to presume that the perennial struggle between workers and management for a larger share of the economic pie may result in rounds of money wage increases being more or less offset by rising prices so as to leave a stalemate in terms of the shares of total output received by labor and capital. With constancy of labor's share ($\dot{\lambda} = 0$), equation (13) implies that the rate of change in prices depends upon both the unemployment rate as embodied in the wage-change function and the rate of productivity growth:

$$\dot{p}(U, \rho) = \dot{w}(U) - \rho\,. \tag{14}$$

The linkage between the rate of growth in worker productivity and inflation is illustrated on Figure 11.12. How far the price change curve, $\dot{p}(U, \rho)$, showing that link between the unemployment rate and inflation, lies below the wage change curve, $\dot{w}(U)$, depends on ρ, the rate of labor productivity growth. Also displayed on Figure 11.12 is the NIRU, or Non-Inflationary Rate of Unemployment. This is the level of unemployment that is required to keep prices stable, given the rate of productivity growth. As is clear from the graph, when the unemployment rate is at NIRU, $\dot{w}(U^{\text{NIRU}}) = \rho$ and $\dot{p} = 0$; i.e., prices are stable and wages grow at the same rate as worker productivity.

[13] To verify the second equality, recall from your study of calculus that $d \ln y/dy = 1/y$; then application of the chain rule yields $d \ln y(t)/dt = (dy/dt)/y$.

[14] As explained in Chapter 7.3.3, Cobb and Douglas argued that their production function together with the marginal productivity theory of wages explained the constancy of labor's share.

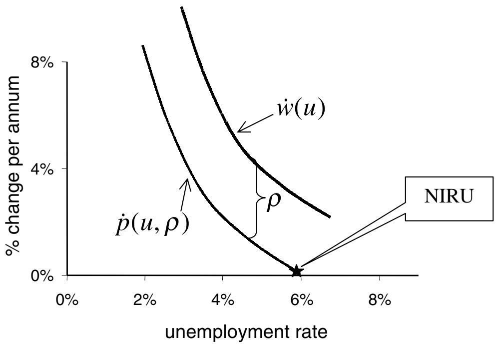


Fig. 11.12. The wage change curve, productivity changes, and inflation
Subtracting the rate of productivity growth ρ from the wage change curve $\dot{w}(u)$ yields the price change curve $\dot{p}(u,\rho)$. The Non-Inflationary Rate of Unemployment (NIRU) is that unemployment rate at which the rate of price change is zero.

President Kennedy

The labor-share arithmetic embodied in equation (13) was invoked by President Kennedy's Council of Economic Advisers in the *1962 Economic Report of the President* (p. 186):

> If hourly labor costs increase at a slower rate than productivity, the share of non-labor incomes will grow or prices will fall, or both. Conversely, if hourly labor costs increase more rapidly than productivity, the share of labor incomes in the total product will increase or prices will rise, or both. It is this relationship among long-run economy wide productivity, wages and prices which makes the rate of productivity change an important benchmark for non-inflationary wage and price behavior.

The president's advisers were arguing that if money wages grew no more rapidly than output per hour of labor, and if labor's share in total output remained fixed, then the price level would remain stable, which is the prediction of equation (14).

President Nixon

In the *1972 Economic Report of the President* (p. 356), President Nixon's Council of Economic Advisors invoked precisely the same labor-share arithmetic that Kennedy's Council had employed ten years earlier:

> If compensation per hour of work rises by 5 1/2% per annum, and if output per hour of work rises by 3% per annum . . . then prices will rise by 2 1/2% per annum.

Alan Greenspan

In the late 1990s Alan Greenspan, the Chairman of the Federal Reserve Board, is reported to have reasoned that increases in the rate of productivity growth, generated in part by the contributions of the Internet and the information technology revolution, meant that the economy could enjoy an exceptionally low level of unemployment without inflation. Greenspan, who was highly regarded for his ability to interpret economic indicators, correctly perceived the increase in the rate of growth of worker productivity in the late 1990s, long before it was being generally recognized. His insight gave him the fortitude to resist during the 1990s those who were arguing that it was necessary to slow down the economy more aggressively because unemployment was so low that high inflation was sure to follow. He correctly called the turn, only moderate increases in the Federal funds rate were implemented, the unemployment rate was allowed to fall to 4.0% in 2000, and Chairman Greenspan received much of the credit for what turned out the be the longest period of economic boom in the history of the United States.

Productivity slowdown

Beginning in the early 1970s and persisting into the 1990s, the United States, and indeed practically every industrialized country in the world, experienced a substantial slowdown in the rate of productivity growth. In the 1950s and 1960s worker productivity in the United States had grown at about 2.6% per annum, but starting around 1974 the rate of productivity growth decelerated to about 1.4% a year in the United States. Finally, around 1995 the growth rate accelerated to about 3.0% per year.[15] The size of the slowdown, a difference between 1.4% after 1974 compared with 2.6% before, may appear to be small, but it contributed to the perverse upward shift of the Phillips curve.

[15]Estimates from the *Economic Report of the President*, U.S. Government Printing Office, 2002, p. 10. The sources of productivity growth will be discussed in more detail in Chapter 12.4.

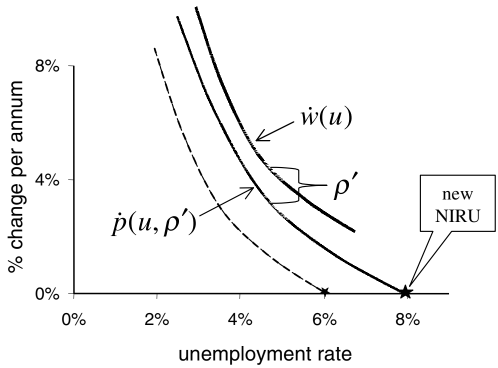


Fig. 11.13. A productivity slowdown shifts the Philips curve.
A slowdown in the rate of worker productivity growth to ρ', because it narrows the gap between the price change curve and the wage-change curve, increases the Non-Inflationary-Rate of Unemployment (NIRU).

How the slowdown in productivity growth affects the non-inflationary rate of unemployment (NIRU) is made clear on Figure 11.13. The reduction in ρ, by narrowing the gap between the rate at which wages increase and the rate of inflation, pushes the $\dot{p}(U,\rho)$ curve up toward the $\dot{w}(U)$ wage-change curve. As can be seen from the graph, this leads to an increase in the *non-inflationary rate of unemployment (NIRU)*. Thus the productivity slowdown at least partially explains the worsening tradeoff between unemployment and inflation in the late 1970s and 1980s. The accelerated growth of productivity in the second half of the 1990s reversed the process and improved the tradeoff between unemployment and inflation as the economy moved into the 21st century. The favorable unemployment-inflation tradeoff was reinforced by the strength of the dollar in the foreign exchange market, which made imports cheaper, plus the continued low world price of petroleum imports.

11.3.5 *Anticipated inflation shifts the Phillips curve*

While changes in worker productivity help us to understand the course of inflation in recent decades, they are not enough to explain more than partially the painful path of the inflation that plagued the United States in the 1970s and early 1980s. The next step will be to elaborate upon the wage-change curve, $\dot{w}(U)$.

A fatal weakness of the simple Phillips curve is that it assumes that workers do not bargain with employers on the basis of anticipated inflation. It assumes that the increase in the money wage is unaffected by the rate of inflation. This might happen in the very short run if workers fail to note what is happening to prices. Once inflation is generally recognized and anticipated to prevail into the future, employers can plan on paying higher wages because they can expect to sell the resulting output for more. More than this, it seems reasonable to suppose that when inflation has not been fully anticipated in past wage negotiations, workers will make a major effort to negotiate catch-up wage increases in order to restore their real wage.

Short-run Phillips curves

One way to analyze such complications is to introduce the anticipated rate of inflation, $\hat{\dot{p}}$, directly into the money-wage-change-curve. Suppose, using a functional form introduced in early empirical work by George Perry[16], that

$$\dot{w}(U, \hat{\dot{p}}) = \omega_0 + \omega_1 \frac{1}{U} + \omega_2 \hat{\dot{p}}\,. \tag{15}$$

If workers do not succeed in catching up with inflation, we will have $0 < \omega_2 < 1$. Assuming in accordance with equation (14) that prices rise by less than money wages because of productivity growth, we have

$$\dot{p}(U, \hat{\dot{p}}, \rho) = \dot{w}(U, \hat{\dot{p}}) - \rho = \omega_0 + \omega_1 \frac{1}{U} + \omega_2 \hat{\dot{p}} - \rho\,. \tag{16}$$

This equation explains the shifty short-run Phillips curve displayed on Figure 11.14. There are four short-run Phillips curves on the graph, each drawn for a different anticipated rate of inflation. These are short-run Phillips curves because each is drawn under the assumption that the anticipated rate of inflation, $\hat{\dot{p}}$, is constant.

The long-run Phillips curve

The more steeply sloping curve on Figure 11.14 is the long-run Phillips curve derived by assuming that if the inflation rate stabilizes in the long run at some level $\dot{p}_L$, the public must eventually learn what that rate of inflation is; i.e., inflation will be fully anticipated in long-run equilibrium,

[16] "Wages and the Guideposts," *American Economic Review*, September, 1967, pp. 897–904.

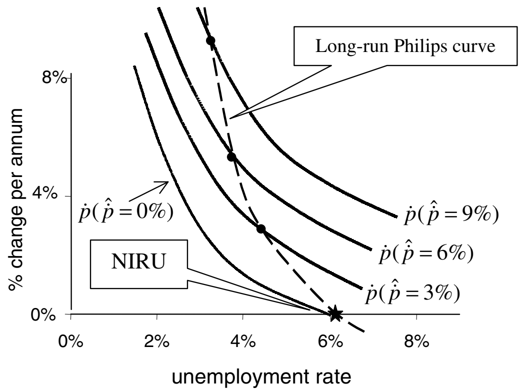


Fig. 11.14. The expectations augmented Phillips curve
When inflation has persisted, the public will anticipate that it will prevail in the years ahead and push for bigger wage increases. The family of short run $\dot p$ curves on the graph indicate that the higher the anticipated rate of inflation the greater the upward shift of the price change curve. On each of the short-run Phillips curves there is one "truth point" where the anticipated inflation rate turns out to be accurate, For example, the graph suggests that with 3% anticipated inflation, the steady state unemployment rate would have to be about 4 1/2 percent. The long run Phillips curve connects all these truth points.

which means that $\hat{\dot p} = \dot p = \dot p_L$. Therefore, to find the nature of the long-run relationship between unemployment, productivity changes, and inflation, we substitute $\dot p_L$ for $\dot p$ and $\hat{\dot p}$ into equation (16), which yields

$$\dot p_L = \omega_0 + \omega_1 \frac{1}{U} + \omega_2 \dot p_L - \rho \,. \tag{17}$$

Therefore, with $0 < \omega_2 < 1$,

$$\dot p_L = \frac{\omega_0}{1-\omega_2} + \frac{\omega_1}{(1-\omega_2)U} - \frac{\rho}{1-\omega_2} \tag{18}$$

Because this long-run relationship makes the rate of inflation more sensitive to the unemployment rate than the short-run Phillips curve, the long-run Phillips curve has a steeper slope on Figure 11.14.

Special interest obviously attaches to the rate of unemployment that is compatible with long-run price stability, called the Non-Inflation Rate of Unemployment, or NIRU. With the functional form of inflation equation (18), we find on setting $\dot p_L = 0$ that NIRU is

$$U^{\text{NIRU}} = \frac{\omega_1}{\rho - \omega_0} \,. \tag{19}$$

Clearly, the larger ρ (the rate of productivity growth), the lower the unemployment rate can be pushed without generating inflation. This is how the rise in productivity growth in the 1990s helped make it possible for the American economy to enjoy remarkably low unemployment without substantial inflation.

The Phillips curve loop

When the reactions of policy makers are taken into account, the expectations augmented Phillips curve mechanism can generate a dynamic inflation-unemployment cycle illustrated by the circular arrows on Figure 11.15. Our economy's initial equilibrium at Point A (with $U = 6\%$) is disturbed by the misguided application of fiscal stimulus. Let us analyze the various phases of the resulting adjustment process:

Phase 1: The unemployment rate is pushed down below U^{NIRU}, and initially only a moderate inflationary penalty is imposed, as at Point B.

Phase 2: Once the public learns from experience to anticipate the inflation, wage bargaining become more intense and the rate of inflation rises more and more rapidly, given the unemployment rate, as at Point C.

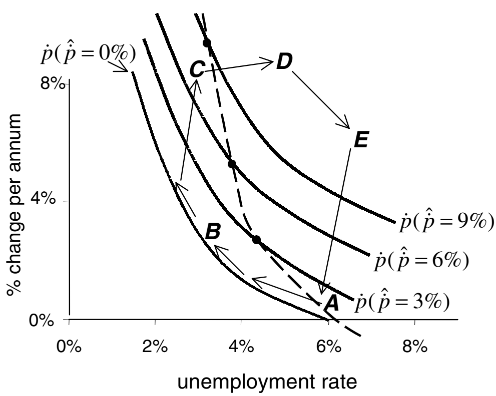


Fig. 11.15. The Philips curve loop
Point B may appear more favorable than point A, but it is not sustainable. Once the public begins to anticipate inflation, the short-run Phillips curve shifts upward and the pace of inflation quickens, as in the movement from B to C. Massive unemployment, as at points D and E, is likely to be required in order to wipe out inflationary expectations. The cycle is all too likely to repeat.

Phase 3: Only after the unemployment increases to the point where the economy is well to the right of the long run Phillips curve, as at points D and E, will inflation begin to slow and eventually be brought under control.

Political business cycles

The Phillips curve loop suggests the possibility of a political business cycle. Anyone planning to become President would be well advised to pick an easy act to follow by assuming office after an administration that has suffered through a period of high unemployment. Presidents John Kennedy and William Clinton were both blessed in this way, for high unemployment in the preceding administrations had reduced inflationary expectations toward zero, as at Point A. Indeed, it is said that the stress of high unemployment helped get them elected. A president so fortunate as to inherit a low inflationary expectations economy can push the unemployment rate down with only moderate inflation, moving to point B. It can be left until after the next election for the next administration, taking office at point C of the cycle, to confront the problem of the inflation generated by the low unemployment policy. Unless rescued by a spurt in the rate of productivity growth, the next President may have to face up to the task of imposing a massive dose of unpopular unemployment at point E in order to move the economy around full circle toward point A, just in time for the next election. Thus the timing of the business cycle may be influenced by the efforts of incumbent politicians to push the unemployment rate down in advance of elections with the realization that the painful inflationary consequences will not become manifest until after the voters have gone to the polls.

11.3.6 *The vertical long run Phillips curve*

The negative slope of the long run Phillips curve generated by equation (18) and plotted on Figure 11.14 implied that there is a tradeoff between inflation and unemployment, even in the long run when everyone has had a chance to adapt to the upward spiral of prices. The source of the tradeoff becomes apparent when we subtract $\dot{p}$ from both sides of equation (15) in order to obtain the rate of change in the real wage. We have, since $\hat{\dot{p}} = \dot{p}$ in the long run,

$$\dot{w}^r(U, \dot{p}) = \dot{w}(U, \dot{p}) - \dot{p} = \omega_0 + \omega_1 \frac{1}{U} + \omega_2 \dot{p} - \dot{p}\,. \tag{20}$$

This equation states that with $\omega_2 < 1$, which we assumed in deriving equation (18) and the long-run Phillips curve plotted on Figure 11.14, workers will suffer a reduction in the rate of growth of their real wage when there is more inflation, given the unemployment rate. That certainly may happen in the short run when inflation is not fully anticipated, but it does not make much sense for the long run when workers and management have had plenty of time to adapt to the inflationary process.

To avoid the implication that workers are misled by the inflationary process into accepting a reduction in the rate of growth of their real wages, given the level of unemployment, we must set the parameter $\omega_2 = 1$ in equation (15). This change does not have a major effect on the short-run Phillips curves plotted on Figure 11.15. But equation (18) explaining the long-run relationship is obviously no longer valid. Moving back to equation (17), which was based on the assumption that in the long run $\hat{\dot{p}} = \dot{p} = \dot{p}_L$, we note that with $\omega_2 = 1$ we can eliminate $\dot{p}_L$ by subtracting it from both sides of the equation. We are left with $\omega_0 + \omega_1/U - \rho = 0$, which yields as the unique level of unemployment that can prevail in the long run:

$$U_n = \frac{\omega_1}{\rho - \omega_0} \tag{21}$$

This level of unemployment is known as the "*natural rate of unemployment.*" Sometimes it is called the "non-accelerating inflation rate of unemployment" or NAIRU and sometimes it is called the "benchmark unemployment rate." The natural rate is said to be the level of measured unemployment that would be generated by the market mechanism in long-run equilibrium, but taking into account market imperfections and making proper allowances for frictional unemployment that naturally arises as a result of normal job turnover and the entry of new job hunters into the labor force.

The nature of the inflation mechanism if $\omega_2 = 1$ is plotted on Figure 11.16. Although this graph appears similar to Figure 11.15, there is a fundamental difference. The pace of inflation will accelerate whenever the level of unemployment is pushed below the natural rate (NAIRU). As long as U is held below the natural unemployment rate, the pace of inflation becomes faster and faster, first $\hat{\dot{p}} = 3\%$, then 6%, then 9%, and so on in an ever accelerating upward spiral. The inflation rate will continue to accelerate in this way as long as the unemployment rate is kept below the natural rate. Pushing the unemployment rate back up to the natural rate achieves $d\dot{p}/dt = 0$. That is why the long run Phillips curve is a vertical line,

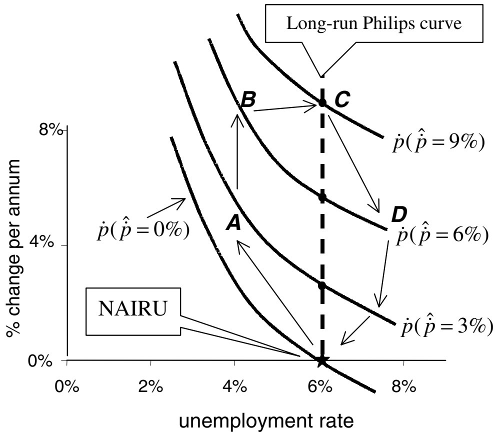


Fig. 11.16. The vertical long-run Phillips curve
Suppose the public is not misled by inflation. For example, if workers and management would settle on a zero wage increase when there is 6% unemployment and zero anticipated inflation, then they they will settle, if the unemployment rate remains at 6%, on a 3% wage increase if 3% inflation is anticipated or a 6% money wage increase if 6% inflation is anticipated.

The Non-Accelerating Inflation Rate of Unemployment (NAIRU), 6% in this example, is that rate of unemployment at which the current rate of inflation will continue, what ever it may be, neither accelerating or decelerating.

as on Figure 11.16.[17] This long-run Phillips curve shows that at measured unemployment rate U_n or NAIRU the current rate of inflation, whatever it is, will persist. It will prevent further acceleration in the rate of inflation, but it won't suffice to reduce $\hat{\dot{p}}$ or $\dot{p}$. To slow the inflation it is necessary to push U temporarily above the natural rate.

If one believes that the distorting effects of inflation on the real wage cannot persist in the long run (i.e., $\omega_2 = 1$), then one must conclude that the tradeoff between inflation and unemployment promised by Samuelson and Solow is an illusion. If the unemployment rate is pushed below this natural rate, which is said to have happened in the 1960s, the resulting inflation will become faster and faster. In order to undo the inflation generated by excessive expansion, it is necessary for the central bank to further slow the

[17] This argument was developed independently in two separate papers: Milton Friedman, "The role of Monetary Policy," *American Economic Review*, March, 1968, and Edmund S. Phelps, "Introduction" in E. S. Phelps, ed., *Micro Economic Foundations of Employment and Inflation Theory*, Norton, 1970.

economy and push the unemployment rate temporarily above its natural rate, as in the late 1970s and 1980s. The period of high unemployment is required in order to convince the public that the monetary authorities will indeed take whatever contractionary steps are necessary in order to bring the inflation under control; only then will $\hat{\dot{p}} = \dot{p} = 0$, as required for the restoration of the natural rate of unemployment without inflation.

How long it will take to beat the inflation out of the system depends upon whether the public believes that the central bank will take a resolute stand against inflation. If the public doubts that the Fed has the intestinal fortitude to persevere with high unemployment, then $\hat{\dot{p}}$ will remain positive and inflation will not be brought under control. If the central bank establishes credibility, if it can convince the public that it will pursue a high unemployment anti-inflationary policy for however long may be required, then the public will be convinced that the inflation will be stopped and we will have $\hat{\dot{p}} = 0$. Once the Fed has beaten inflationary expectations back to zero it can allow the unemployment rate to fall to the natural rate of unemployment without inflation.

11.4 Rational expectations

The derivation of the expectations augmented Phillips curve and the resulting Phillips curve loop of Figure 11.15 relied on the assumption that the public makes systematic errors in forecasting the rate of inflation. Fundamentally different results are generated if it assumed that the public is not so dumb. That is the basic assumption of the theory of rational expectations pioneered by John F. Muth.[18] Rational expectation theorists assume that the public does not make systematic errors in forecasting future economic developments, even in the short run. This is not to say that forecasts are perfect. In an uncertain world, rational forecasters will make mistakes, but they will not consistently over or consistently underestimate what is going to happen. If, for example, the public found from experience that on average they had underestimated inflation by 5%, they would improve their future predictions by henceforth adding 5% onto their forecasts in order to correct the systematic error. More precisely, if experience revealed that with existing forecasting procedures on average $\dot{p} = \hat{\dot{p}} + 5\%$, then the public would eliminate the systematic underestimation error by

[18] John F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July, 1961.

hence forth applying a correction to their forecasts: $\hat{\dot{p}}^R = \hat{\dot{p}} + 5\%$. Or, to take a slightly more complicated case, if on average $\dot{p} = \alpha_0 + (1 + \alpha_1)\hat{\dot{p}}$, then henceforth the public would offset the systematic error by using the corrected forecast $\hat{\dot{p}}^R = \alpha_0 + (1 + \alpha_1)\hat{\dot{p}}$.

According to the theory of rational expectations, the forecasts used by the public have been corrected in this way for systematic errors. This does not mean that people forecast perfectly — they do make errors, but they are not systematic. According to the assumption of rational expectations,

$$\dot{p} = \hat{\dot{p}}^R - \varepsilon_t \,, \tag{22}$$

where ε_t, the error of forecast, is positive when the rate of inflation is underestimated and negative when inflation falls short of the anticipated rate. If the forecasts are indeed rational, the forecast errors ε_t must average out to zero and be unpredictable rather than systematic; i.e., they are purely random. But more than this is required for expectations to be truly rational. Expectations are truly rational only if forecast errors are not systematically related to *any* variables that are observable at the time the forecasts are made. Why? Because if there were a systematic relationship, a rational forecaster would be exploiting that information in order to derive an improved prediction, contradicting the assumption that the forecast with the systematic error is rational.

11.4.1 *Rational expectations and the Phillips curve*

The assumption that expectations are rational implies that two fundamental adjustments must be made in the Phillips curve relationship. First, if it is reasonable to assume that workers do not make systematic errors in anticipating inflation, then they will not be misled by inflation. Thus it also seems reasonable to assume that they succeed in preventing their bosses from pushing their real wages down in times of inflation. This means that we have $\omega_2 = 1$ in equation (15) and (16). Second, we substitute the rational forecast $\hat{\dot{p}}^R = \dot{p} - \varepsilon_t$ for $\hat{\dot{p}}$ in equation (16), which yields with $\omega_2 = 1$:

$$\dot{p} = \omega_0 + \omega_1 \frac{1}{U} + \dot{p} - \rho - \varepsilon_t \,. \tag{23}$$

This equation has a startling implication: when we solve the equation for U we find that with rational expectations there is, even in the short run, a unique unemployment rate that is not affected by $\dot{p}$:

$$U_t = \frac{\omega_1}{\rho + \varepsilon_t - \omega_0}\,. \tag{24}$$

Now consider the gap between this equation for unemployment and the natural rate of equation (21):

$$U_t - U_n = \frac{\omega_1}{\rho + \varepsilon_t - \omega_0} - \frac{\omega_1}{\rho - \omega_0}\,. \tag{25}$$

If expectations are truly rational, the random forecast error ε_t in the denominator of the first fraction following the equals sign generates a random gap between the actual and the natural unemployment rate. When it happens that $\varepsilon_t = 0$, we will have $U_t = U_n$; otherwise, U departs randomly from U_n.

11.4.2 *The Lucas aggregate supply function*

Invoking the assumption of rational expectations, Nobel Laureate Robert Lucas advanced the famous proposition that even in the short run departures from the vertical long-run aggregate supply curve that was plotted on Figure 10.20 must be random! As a first step toward appreciating this surprising result, let us review the properties of the long-run aggregate supply curve. Then we will examine the reason for short-run departures from the long-run equilibrium.

To review, the short-run aggregate supply function linking the price level to aggregate output was derived in Chapter 10.5.3 under the assumption that the money wage rate is rigid, it being assumed that in the short run workers would succeed in preventing reductions in their money wage. In the long run, the supply of labor will be determined by the real wage, as was explained in Chapter 4.4.4. And given the technology, employers operating in competitive markets will hire workers until the real wage equals their marginal product, just as was explained in Chapter 7.3.1.[19] In the long run the real wage adjusts so as to equate the demand for labor with the supply. In the absence of government intervention, as with the minimum wage, we

[19] "Long run" aggregate supply and demand analysis neglects the effects of capital investment and technological advance on the nation's capacity. The next chapter presents a growth model showing the effects of capital investment, technological change, and population growth on a nation's output.

will have full employment. Given the level of employment and the stock of capital, the level of output is determined by the aggregate production function. Thus we have a vertical aggregate supply curve in the long run.

Lucas rejected the assumption that wage or price rigidities keep the economy from adjusting promptly to the full-employment equilibrium, even in the short run. Instead, he attributed short-run departures from the vertical long-run aggregate supply curve to misperceptions about price and wage developments. Consider an economy initially in long-run equilibrium at point e on Figure 11.17. Now suppose a representative business firm observes an upward shift in the demand for its product. The upward shift might result from an underlying change in the demand for the firm's product, such as would be generated by a shift in tastes. If so, the firm should increase production. But suppose instead that the increase in demand were the result of an inflationary upward shift in the aggregate demand function, which might be caused by the Fed's decision to allow excessive money supply growth. In this case, the firm should hold production stable because the increase in demand will be offset by rising labor and material costs. The problem for the firm is to decide how to respond to the increase in demand before it knows whether the supply function of labor and the costs of other inputs will shift upward or remain stable. In the absence of other information, firms may interpret the increase in demand as at least in part a favorable shift in the market for their product that will not be offset by rising costs; therefore, they expand production. Then the economy will slide up the short run aggregate supply curve to point e_s on Figure 11.17.

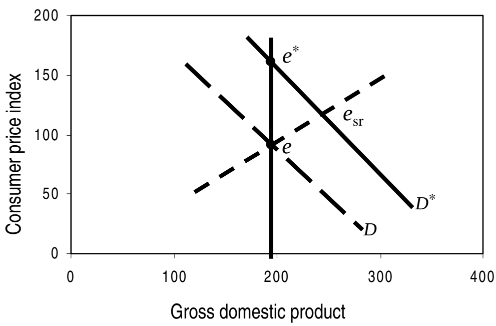


Fig. 11.17. Lucas aggregate supply
Monetary policy affects the economy only if it is not anticipated. Rationality implies that expectation errors are random. Therefore, policy measures can only cause random fluctuations in economic activity.

Things work out differently if the increase in the money supply is anticipated. Then firms will correctly recognize that the increased demand for their product is part of a general rise in costs and prices; therefore, they have no reason to increase either their workforce or their output. This means that if price and wage movements are correctly anticipated the economy will adjust immediately to point e^* without even a short run increase in supply. However, Lucas did not assume that firms anticipate develop ments with complete accuracy. Instead, he invoked the rational expectations argument that expectation errors are random, which means that the short run departures of output from long-run equilibrium will be random as well.

Policy ineffectiveness

A startling implication for economic policy follows from the assumptions that (1) expectations are rational and (2) workers are not led by anticipated inflation to accept a reduction in their real wage (i.e., $\omega_2 = 1$ in equation (15)). We shall find that these two assumptions fundamentally limit the scope for stabilization policy.

Suppose, for example, that the central bank increases the rate of growth of the money supply in an attempt to stimulate the economy. If this policy action is anticipated, there will be no effect on employment and output will be unaffected. Or suppose the policy action is a complete surprise, as represented by ε_t in equation (22). Then the effect of the policy surprise on prices will be unanticipated, and the real wages of workers will be reduced. Thus the ε_t shock leads to an expansion of employment and output. But if expectations are truly rational, the expectation error ε_t in equation (25) must be random and so must be the effects of monetary policy on unemployment, on the supply of labor, and on output. Thus economic policy has an effect on the economy only to the extent that it is unanticipated. But according to rational expectations theory, to be unanticipated the policy must be random, and so must be the resulting fluctuations in output. To summarize, the assumption that expectations are rational implies that only the random component of monetary policy will affect the economy, but that will generate random fluctuations in output instead of contributing to economic stability.[20] The argument does not rest on the

[20]Robert E. Lucas, Jr., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April, 1972, developed his argument within a consistent general equilibrium framework based on maximizing behavior. A digestible presentation of the policy ineffectiveness argument is developed by Thomas J. Sargent and Neil Wallace, "Rational Expectations and the Theory of Economic Policy," *Journal of Momentary Economics* 2 (1976) 169–183.

assumption that policy makers inevitably make mistakes, given the limited state of current knowledge. Rather, it relies on the assumption that the public's anticipations are indeed rational, which is an issue on which there is less then complete agreement.

Critique

The policy ineffectiveness proposition and the Lucas supply function on which it is based rely on two critical assumptions. Expectations must be rational and prices and wages must adjust promptly so as to equate demand with supply. It is this second assumption that has been subjected to the most serious criticism.

Many economists argue that wages are slow to adjust to shifting demand because of long term "overlapping" labor contracts. If the workers for American Airlines have just signed a three year contract, a decrease in demand, perhaps sparked by an unexpected decrease in the money supply, will be met with a contraction in output and employment rather than a fall in wages that would facilitate a reduction in fares; workers with enough seniority to keep their jobs may enjoy an increase in their real wage. And if rival United's contract comes up for negotiation a few months later, their workers are likely to be reluctant to accept much of a reduction in their money wages below what American employees are receiving. Thus it is a gradual process by which money wage adjustments take place, and as a result, part of the adjustment to shifts in aggregate demand will involve a temporary fall in employment and output.

There is another reason for questioning the policy ineffectiveness proposition. Federal Reserve Chairman Alan Greenspan has received considerable credit for engineering the longest economic expansion in U.S. history and for successfully combating the 2001 recession. Yet as we shall see, his policy was not unpredictable but confirmed fairly closely to a simple equation for setting monetary policy known as the "Taylor Rule."

11.5 What should central bankers do?

Central banks worry about preventing inflation, preserving the value of their country's currency in the foreign exchange market, contributing to long run economic growth, and maintaining an appropriate level of employment. Which of these objectives should have the highest priority? What strategy is most appropriate in attempting to achieve the bank's goals? Or, given the limited state of currently available knowledge about how the economy actually works and the notorious difficulties in forecasting future economic developments, should the Fed pursue a neutral "hands off" policy and rely on the self-recuperating powers of the economy to keep the system on an even keel? These questions have worried generations of economists, and each generation has had somewhat different answers to the recurring questions. More than this, economists of each generation have been divided on these issues.

11.5.1 *Rules versus discretion*

Should a country's central bank deliberately adjust its policy on the basis of its appraisal of current economic conditions or should it follow a set of rules prescribing how it should behave? And if we are to have a central bank governed by rules, what rule should they be?

Early in its history, the Fed's behavior was influenced by the "productive credit" argument. This rule, also known as the real bills doctrine, was designed to make sure that the Fed would not find itself lending funds that the commercial banks would use to make loans for speculative purposes, inflaming the financial markets. Therefore, funds loaned by the Fed to the commercial banks should only be used to finance productive uses of credit, such as to help finance the payment of wages and suppliers while goods were being manufactured or held by retailers as inventory before purchase by the final consumer. This meant that the Fed would be loaning more in boom times when there was more business borrowing. Since Fed lending increases the reserves of the commercial banks, this led to an expansion of the money supply during the boom. When business slowed in recessions, so would Fed lending to the commercial banks, and so the bank reserves, commercial bank lending and the money supply would all contract. Critics of the Fed argue that this policy was destabilizing because it led to a pro-cyclical movements in the money supply, expanding in boom times and contracting in recession.

In the 1950s and beyond Professor Milton Friedman of the University of Chicago and other monetarists argued that the Fed had historically done the wrong thing, allowing the money supply to contract when the economy went into the Great Depression, which had only made the situation worse. Friedman argued that given the limited state of current knowledge, the best thing for the Fed to do was to pursue a neutral policy of keeping the money supply growing at a constant rate per annum regardless of the state of the economy. Monetarists have long complained that the Fed has failed to stabilize the rate of growth of the money supply. Only for one brief period, when confronted with massive inflation, has the Fed focused on stabilizing the rate of money supply growth. In August of 1979 Paul Volcker, the newly appointed chair of the Federal Reserve Board, announced that the Fed would focus on achieving a target rate of growth of the monetary aggregates rather than targeting the federal funds rate. But in practice the rate of growth of M_1 was not stabilized, and the experiment was abandoned within just a few years.

11.5.2 *Leaning against the wind — Taylor rules*

Central bankers have long described what they do as "leaning against the wind," attempting to slow the economy when it is expanding too rapidly and stimulating it when times are slow. A simple policy equation capturing the spirit of leaning against the wind has been analyzed by Stanford Professor John B. Taylor, who has served as Treasury Undersecretary for International Affairs under President George W. Bush. Taylor's equation focuses upon the Federal funds rate, which is the interest rate banks charge each other when they loan funds they have on deposit at the Fed.

Taylor's equation says that the Fed should raise the real (inflation adjusted) federal funds rate, $i^{\mathrm{fed}} - \dot{p}$, by 1/2 of the excess of the annual rate of inflation over 2%. But the Fed should lower the real rate by 1/2 of the GDP gap (measured as a percent). More precisely,

$$i^{\mathrm{fed}} - \dot{p} = 2\% + 0.5(\dot{p} - 2\%) - 0.5Y^{\mathrm{gap}}\,. \tag{26}$$

Or to put it another way,

$$i^{\mathrm{fed}} = 2\% + \dot{p} + 0.5(\dot{p} - 2\%) - 0.5Y^{\mathrm{gap}}\,. \tag{27}$$

If inflation is $\dot{p} = 2\%$ and $Y^{\mathrm{gap}} = 0$, then the real federal funds rate, $i^{\mathrm{fed}} - \dot{p}$, should be 2% and the nominal federal funds rate 4%. But if the rate of inflation rises to $\dot{p} = 6\%$, then the Fed should put the brakes on the economy by raising the real federal funds rate to 4%, which would make the nominal rate $i^{\mathrm{fed}} = 10\%$. The Taylor rule prescribes that it is not enough for the Fed to raise the nominal interest rate when inflation threatens; it must raise the nominal rate by more than the rate of inflation so that the real cost of borrowing increases. The Taylor rule also prescribes that if the Y^{gap} has increased by 2% but the rate of inflation remains unchanged, then the nominal federal funds rate should be lowered by 1%.

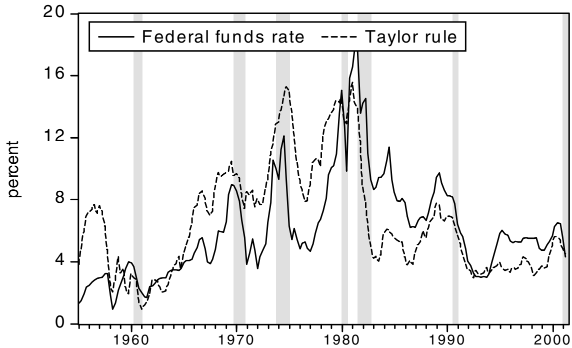


Fig. 11.18. The federal funds rate and the Taylor rule
The Taylor rule, based on the unemployment rate and inflation, tracks much of the movement in the federal funds rate.

Figure 11.18 contrasts the federal funds rate predicted by a simple Taylor rule with the actual federal funds rate.[21] The line showing the federal funds rate prescribed by this Taylor rule is calculated using the most recently revised data on inflation. In contrast, only preliminary estimates are available for the policy makers at the time they were making their decisions. Further, there is more than one procedure for estimating Y^{gap}. Nevertheless, the graph suggests that the Taylor equation does a pretty good job of describing what the Fed actually has been doing in recent years. In the more distant past, the fit is less precise because the Fed was not responding as vigorously to excessive inflation and unemployment as equation (26) requires.

11.5.3 *Simulating policy with econometric models*

Computer simulations provide an environment for experimenting with alternative economic policies. The strategy is to trace out the time paths for GDP, unemployment, inflation and other variables of interest that are generated by a carefully constructed computer model of the economy. This line of research was pioneered, before the advent of computers, by Jan Tinbergen in the 1930s and by Lawrence Klein, beginning in the 1940s. Both scholars earned the Nobel Prize. Over the decades improvements in computer technology coupled with the great increase in the availability

[21] The GDP Gap is calculated with an Okun coefficient of 2 and a benchmark unemployment rate of 6%.

of economic data and the development of improved estimation procedures facilitated the construction of larger and more sophisticated macro-econometric models.

The reliability of computer simulations hinges upon the validity of the underlying econometric model and the accuracy with which the parameters of the models have been estimated from historical data. An obvious test of a macro-econometric model is to ask how well it predicts future economic development. As mentioned in Chapter 8.6.3, the predictive ability of macro-econometric models has been a major disappointment. The models generate suggestive simulations, but the outcomes must be taken with a grain of salt. In particular, in judging simulation outcomes it is important to know what economic assumptions have been built into the behavioral equations.

The FRB/US model of the U.S. economy, developed at the Federal Reserve Board, is one of the more recent in a long line of sophisticated macro econometric models. It is based on more than 350 empirically estimated equations and identities. In contrast to many earlier models, it incorporates the assumption that business firms and consumers are maximizers with rational expectations. However, decision makers are assumed to face significant frictions slowing the speed with which they adjust prices and quantities, such as arise from the costs faced by firms in changing their work force. As a result prices and wages are sticky. They do not adjust quickly enough to insure full resource use at all times. This means that during economic downturns a high percentage of the workforce may be willing to work at current wages but unable to find a job.[22]

The first policy simulation reported on Table 11.5 asks what would happen if the federal government permanently increased its purchases of goods and services by 1% of GDP. The answer depends upon how the Fed responds. In Chapter 9.5.4 we reported multiplier simulations generated with this model under the assumption that the Fed keeps the real federal funds rate constant. But suppose instead that the Fed responds more aggressively, raising the federal funds rate by more than the rate of inflation in accordance with the Taylor rule? Comparison of simulation 1a (the simulation that was reported in Chapter 10), and simulation 1b confirms that what happens does indeed depend on the Fed's response.

[22]The basic features and some simulations generated by the model are reported by David Reifshneider, Robert Tetlow and John Williams in "Aggregate disturbances, monetary policy, and the macroeconomy: the FRB/US Perspective," *Federal Reserve Bulletin*, January, 1999.

Table 11.5. Policy simulations: FRB/US macro-econometric model simulations.

	Response (%) at end of 1st year	2nd year	3rd year	10th year
#1: Federal government spending on goods and services increased by 1% of GDP				
#1a: Real Federal Funds Rate Constant				
GDP (real)	1.4	1.4	1.1	1.1
Unemployment rate	−0.5	−0.7	−0.7	−0.6
Consumer price inflation	0.1	0.5	0.7	1.4
#1b: Taylor Rule				
GDP (real)	1.1	0.5	0	−0.6
Unemployment rate	−0.3	−0.3	−0.2	0
Consumer price inflation	0	0.1	0.1	0
#2: Permanent increase in personal income tax equal to 1% of GDP				
#2a: Real Federal Funds Rate Constant				
GDP (real)	−0.4	−0.8	−1	−1.5
Unemployment rate	0.1	0.3	0.4	0.7
Consumer price inflation	0	−0.1	−0.3	−1.2
#2b: Taylor Rule				
GDP (real)	−0.3	−0.5	−0.5	−0.1
Unemployment rate	0.1	0.2	0.2	0.1
Consumer price inflation	0	0	0	0
#3: Permanent $10 per barrel increase in the price of oil				
#3a: Real Federal Funds Rate Constant				
GDP (real)	−0.2	−0.4	−0.2	−0.3
Unemployment rate	0.1	0.2	0.1	−0.3
Consumer price inflation	0.5	0.3	0.2	0.4
#3b: Taylor Rule				
GDP (real)	−0.2	−0.4	−0.2	−1.1
Unemployment rate	0.1	0.2	0.1	0
Consumer price inflation	0.5	0.3	0.1	−0.1

Source: David Reifschneider *et al.*, "Aggregate Disturbances, Monetary Policy and the Macroeconomy: The FED/US Perspective," *Federal Reserve Bulletin*, January, 1999.

The more aggressive hike in the federal funds rate called for by the Taylor rule stabilizes unemployment and prevents inflation at the expense of a much smaller increase in GDP.

If the Fed reacts by increasing the Federal funds rate in proportion to the inflation, keeping the real federal funds rate constant, real GDP during the first year that the policy is in effect is 1.4% higher than it would have been without the fiscal stimulus, suggesting a government spending multiplier of $\Delta Y/\Delta G = 1.4$. But as can be seen from the table, the effect slackens off after a couple of years to a 1.1% gain in GDP, implying that the longer run multiplier is only 1.1. Observe that initially there is only a small inflationary penalty to be paid for the fall in unemployment, which is in accordance with the short-run Phillips curve. The inflation-unemployment tradeoff deteriorates with the passage of time, but even after ten years there is some tradeoff of reduced unemployment at the expense of inflation. The long-run Phillips curve tradeoff generated by this model is steep but not vertical.

Simulation 1b shows how the economy responds to the same government spending shock if the Fed follows the Taylor rule, raising the federal funds rate by more than the rate of inflation so as to discourage borrowing. As anticipated, with this more aggressive monetary policy the multiplier effect is dampened and the unemployment rate does not fall as much, but inflation is kept under control.

Simulation 2a considers a hike in the personal income tax equal to 1% of GDP. Since personal income tax revenue is about 9% of GDP, this treatment amounts to an increase in the personal income tax receipts of a bit more than 10%, or bad news for the tax payers. Because this is a tax hike, it pushes the economy in the opposite direction from the G increase of Simulation 1a. The downward force of higher taxes is stronger than the expansion generated by the increase in G, but the G effect may have been limited because the economy was pushing against the capacity ceiling.

Comparison of Simulation 2b with 2a reveals that when the Fed is pursuing a more vigorous Taylor Rule policy of lowering interest rates in response to the slowdown generated by the tax hike, the deflationary pressure is much less and the contractionary pressure is muted.

The third pair of policy simulations looks at the effects of a rise in the price of oil. Under either FED policy, there is a fair amount of inflationary pressure, a protracted decline in GDP and an initial increase in unemployment. For what it is worth, the 10 years of computer simulation output suggests that the Taylor rule is more successful in dampening inflation than the less restrictive monetary policy, but the success on the inflationary front is paid for with an increasingly severe decline in real GDP.

Summary

1. The discussion of the dynamic process of adjustment presented in this chapter began with an analysis of the stability properties of the cobweb model. Output and prices oscillate. Only for certain values of the model's parameters will output and prices converge to the levels predicted by static demand and supply analysis.

2. We investigated how governments can buy, store and sell commodities in order to stabilize price. Excessive stockpiles of "surplus" commodities are generated if the government, perhaps responding to political pressures, sets the support price too high.
3. Market participants may try to limit risk by hedging. Successful speculators try to buy low and sell high, and in doing so they may contribute to market stability.
4. An efficient market is one where the current price reflects all the information that is now available. Information on past sales and price will not help to predict future price movements because the implications of past events have already influenced market participants buy and sell decisions, thereby impacting the current price. Prices in efficient markets change randomly, following a random walk. The stock market, financial markets, and the foreign exchange markets are efficient, or nearly so. Most departures from randomness are so small as to preclude profitable speculation, once the costs of executing transactions are taken into account.
5. The Phillips Curve, estimated for the United States by Professors Samuelson and Solow in 1959 (Figure 11.6), promised a reduction in unemployment in return for moderate inflation. While the tradeoff appeared to be exploited effectively in the 1960s, the relationship broke down in the inflationary 1970s and beyond. The Phillips Curve relationship broke down because it neglected the effect of anticipated inflation on wage demands. The inflation expectations augmented Phillips Curve showed a much less favorable inflation unemployment tradeoff. Indeed, the long run Phillips Curve may be vertical, implying that pushing the level of unemployment too low will generate inflation that can be stopped only by letting the unemployment rate rise above its natural rate.
6. Expectations are said to be rational if they are based on all available information. Rational expectations are free of bias and systematic errors that could be corrected by adjusting the forecasts. That is why their errors are random.
7. The Lucas short-run aggregate supply function allows for departures from long run equilibrium output only when errors are made in forecasting the extent of price changes. Under the assumption of rational expectations, such forecast errors must be random. Therefore, the departures from long-run equilibrium output must be random as well. The policy ineffectiveness proposition invokes the Lucas aggregate

supply function. The Fed can affect output only if its actions are not foreseen, and its actions will be unforeseen only to the extent that they are random. But a random policy will have the perverse effect of generating random fluctuations in output rather than stabilizing the economy.

8. Sticky wages and prices rather than expectation errors provide a more widely accepted explanation of how the economy responds to changes in monetary policy. For example, overlapping wage contracts may keep money wages from falling promptly when aggregate demand slacks off, contributing to a decline in output and an increase in unemployment.
9. The Taylor rule, equation (26), is a simple prescription for monetary policy: the Fed should hike the *real* rate of interest when inflation threatens; it should lower it in response to rising unemployment. This rule does a reasonably good job of describing the policy pursued by Alan Greenspan while serving as Fed Chairman from 1987 into the 21st century.
10. The effects of economic policy may be simulated on a computer, but the predictions are no more accurate than the underlying economic model.

Appendix 11.1. Solving first order linear difference equations

An equation of the form

$$X_t = \beta_0 + \beta_1 X_{t-1} \tag{28}$$

is known as a first order linear difference equation with constant coefficients. Since this equation holds for all t, including $t = 0$, it is possible to calculate future values of X_t given the coefficients β_0 and β_1 and the initial condition, X_0. For example,

$$X_1 = \beta_0 + \beta_1 X_0\,, \tag{29}$$

and applying (28) again,

$$X_2 = \beta_0 + \beta_1 X_1\,. \tag{30}$$

Applying (28) again we can calculate X_3. Repeating this recursive procedure t times will yield X_t.

We shall seek a solution to the difference equation. By a solution we mean a function of the form $X_t = f(X_0, t)$, $t = 1, 2, \ldots$. The solution yields

X_t without resort to recursive calculations. Solving for the solution involves two easy steps:

Step 1: We seek an equilibrium solution X^e satisfying

$$X^e = \beta_0 + \beta_1 X^e\,, \tag{31}$$

If the system ever attained an equilibrium value X^e there will be no tendency to change. The solution for (31) is:

$$X^e = \beta_0/(1 - \beta_1)\,, \tag{32}$$

Thus there exists an equilibrium solution, and it is unique.

Step 2: Finding the general solution: Subtracting equation (31) from (28) yields

$$\begin{aligned} X_t - X^e &= (\beta_0 + \beta_1 X_{t-1}) - (\beta_0 + \beta_1 X^e) \\ &= \beta_1(X_{t-1} - X^e)\,. \end{aligned} \tag{33}$$

Since this equation holds for all t, including $t-1$, we also have $X_{t-1} - X^e = \beta_1(X_{t-2} - X^e)$, and substituting into (33) yields $X_t - X^e = \beta_1^2(X_{t-2} - X^e)$. Proceeding repeatedly in this way, we find for any t,

$$X_t - X^e = \beta_1^t(X_0 - X^e)\,. \tag{34}$$

Adding X^e to both sides of (33) yields the desired solution to our difference equation:

$$X_t = \beta_1^t(X_0 - X^e) + X^e\,. \tag{35}$$

Stability condition

The dynamic properties of this system depend critically on the magnitude of β_1. Observe that $Limit_{t\to\infty} = \beta_1^t = 0$ if and only if $-1 < \beta_1 < 1$.

The system is stable, X_t necessarily converging to its unique equilibrium value, only if this condition on β_1 is satisfied. If $\beta_1 = 1$ output will remain constant at whatever its initial value, X_0, happens to be. If $\beta_1 > 1$, output will diverge without bounds from its equilibrium. Also $X_t - X^e$ will oscillate in sign from one period to the next if and only if $\beta_1 < 0$ as was the case with the cobweb model.

References

Procedures for solving second order difference equations of the form $X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2}$ are explained in many texts, including Alpha C. Chiang, *Fundamental Methods of Mathematical Economics*, Chapter 17 and Michael Hoy *et al.*, *Mathematics for Economists*, Chapter 20.

Key Concepts

anticipated inflation, 530
business cycle
  political, 533
cobweb theorem, 498
cost-push inflation, 523
diversified portfolio, 514
efficient markets, 511
equity premium, 513
forward contract, 508
futures, 509
hedging, 509
index funds, 515
Lucas aggregate supply function, 539
NAIRU (Non-Accelerating Inflation Rate of Unemployment). *See* Natural rate of unemployment
natural rate of unemployment, 534
NIRU (Non-inflationary Unemployment Rate), 526
options, 507
Phillips curve, 517
  long-run, 530
  loop, 532
  short-run, 530
Phillips, A. W., 517
political business cycle, 533
policy simulations, 544
productivity slowdown, 525
random walk, 516
rational expectations, 536
speculation, 503
stock options, 508
Taylor rule, 543
total return, 512
worker productivity, 525

Exercises

1. Suppose that the demand curve for wheat is $q_t = 1000 - 20p_t$ and the supply curve is $s_t = 20p_{t-1}$. Describe the dynamic properties of this model. Is it stable?
2.* Consider the data on Table 11.2, Case 3. Determine the profit maximizing quantity for a speculator to purchase in good crop years to carry over to years of drought, assuming that there are no other speculators and that the government does not participate in the market.

3. The Phillips curve for Simpleland is

$$\dot{p} = \left(\frac{0.0012}{u - 0.03}\right) - 0.02\,, \tag{36}$$

where $\dot{p}$ is the rate of inflation and u is the unemployment rate.

a. If $u = 4\%$, how rapidly will prices rise.
b. What is NIRU (non-inflationary rate of unemployment)?

4. Econoland has the following expectations augmented Phillips Curve:

$$\dot{p} = \left(\frac{0.0012}{u - 0.03}\right) - 0.02 + \hat{\dot{p}}\,, \tag{37}$$

where $\hat{\dot{p}}$ is the expected rate of inflation.

a. What is Econoland's natural unemployment rate (also known as NAIRU, the non-accelerating inflation rate of unemployment)?
b. What is the significance of the difference between the two alternative formulations of the Phillips Curve?

12

Growth and Development

12.1 Introduction

Why some nations grow while others stagnate is a question that has captured the interests of generations of economists. And economists have not always been optimistic about the long run destiny of nations. The classical school of economists, led by Adam Smith and David Ricardo, worried that

Table 12.1. International comparisons of growth of GDP per capita.

GDP per capita (1990 $)						
	1700	1820	1870	1950	1973	1998
China	600	600	530	439	839	3,117
Germany	894	1,058	1,821	3,881	11,966	17,799
Italy	2,100	1,921	2,753	5,996	13,082	20,224
Japan	520	570	737	1,926	11,439	20,413
Mexico	568	759	674	2,365	4,845	6,655
United Kingdom	1,250	1,707	3,191	6,907	12,022	18,714
United States	527	1,257	2,445	9,561	16,689	27,331
Africa	400	418	444	852	1,365	1,368
World	615	667	867	2,114	4,104	5,709
GDP per capita as a percent of U.S. GDP in 1998						
	1700	1820	1870	1950	1973	1998
China	2%	2%	2%	2%	3%	11%
Germany	3%	4%	7%	14%	44%	65%
Italy	8%	7%	10%	22%	48%	74%
Japan	2%	2%	3%	7%	42%	75%
Mexico	2%	3%	2%	9%	18%	24%
United Kingdom	5%	6%	12%	25%	44%	68%
United States	2%	5%	9%	35%	61%	**100%**
Africa	1%	2%	2%	3%	5%	5%
World	2%	2%	3%	8%	15%	21%

Source: Angus Maddison, *The World Economy: A Millennial Perspective*, Table 5-21, p. 264.

a maturing economy will inevitably approach a "stationary state" characterized by a bare subsistence standard of living and zero economic growth. No wonder economics has been called "the dismal science."

The gloomy predictions of the classical economists were wrong! As is clear from the data presented in Chapter 1.5.1, growth has been the big economic story of the last two centuries. The data on Table 12.1 tell more of the growth story. The evidence, which is measured in international dollars of constant purchasing power so as to permit comparisons both among countries and over time, displays a mixed picture. Some nations have stagnated, but much of the world has enjoyed a dramatic increase in real income rather than the decline predicted by the classical school. In some countries, the average citizen's living standard has doubled and then doubled again in a single lifetime!

This chapter begins by presenting the "classical" argument concerning the inevitability of the stationary state. Then we shall construct a "neo-classical" growth model explaining how technological advance and capital accumulation can, under certain conditions, lead to a continuing improvement of living standards for a growing population in spite of the law of diminishing returns. Later in this chapter we will also ask whether we can count on the market mechanism to allocate petroleum and other exhaustible resources appropriately over time or whether we risk squandering our limited resources to the detriment of future generations. We will also consider a simple model of "over-fishing."

12.2 Malthusian population dynamics

In 1798 the Reverend Thomas R. Malthus [1766–1834] anonymously published *An Essay on the Theory of Population.* Malthus warned the readers of his best seller that while the population tended to grow like a geometric series $(2, 4, 8, 16, \ldots)$, the food supply only grows arithmetically $(1, 2, 3, 4, \ldots)$. As a result, Malthus argued, population growth has an inevitable tendency to outstrip the world's food supply. Therefore, a deteriorating standard of living and widespread hunger are inevitable, unless moral restraint holds reproduction in check. Throughout the 19th and well into the 20th century, many economists, and indeed the public generally, worried about the pessimistic Malthusian prediction.

The Law of Diminishing Returns (recall Chapter 5.4.2) provides a link between population growth and the food supply that supports the prediction of Malthus. Since the amount of land in the world is a fixed resource (Holland, thanks to its dikes, being a notable exception), the law of diminishing returns implies that, if the world's population keeps growing, the supply of food available per worker must eventually decline, as illustrated on Figure 12.1. Here is a "scientific" case for population control.

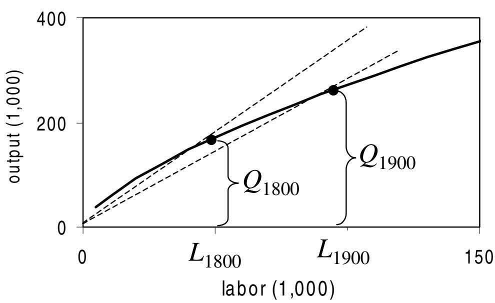


Fig. 12.1. Diminishing returns
The increase in the supply of labor leads to greater output in Econoland. But because of the law of diminishing returns, the increase in the labor supply, other things being equal, leads to a decline in the ratio Q/L, or output per capita.

While the prediction of Malthus appears to this day to be all too true in many sectors of the globe, the evidence on per capita output growth presented in Chapter 1.5.1 and on Table 12.1 makes it clear that the law is far from a universal truth. There are many hungry in the world, but on average the world's population is better fed than at anytime in history — contrary to Malthus, the food supply has grown more rapidly than the population.

12.3 A classical growth model, simplified

Let us begin by considering a model that captures certain essential features of the classical analysis of the growth process as expounded by Adam Smith, David Ricardo and their followers. This model predicts, as did the classical economists, that the law of diminishing returns means that there will be a gradual decline in living standards. The economy will inevitably mature into a *stationary state* characterized by a stable population, zero output growth and a subsistence standard of living. Obviously, this prediction went wrong, but much is to be learned from finding the source of the prediction error. The discussion will set the stage for the subsequent construction of a more optimistic model of the growth process.

The subsistence theory of wages assumes that the rate of growth of the labor force depends on how the average real wage rate w_g compares with the subsistence wage w^s. Here the subscript g indicates the generation; i.e., we use as the unit for recording time the number of years required for a generation to replicate itself — perhaps 25 years equals one generation. If the wage w_g that parents of generation g receive is above the subsistence wage w^s that is required for subsistence, they will have more children and, as a result, the next generation will be larger and so will the work force. If the wage rate is below the *subsistence wage*, the population and hence the workforce will shrink. The conjecture that it is the gap between the wage w_g that workers of each generation actually receive and the subsistence wage w^s that determines the rate of population growth is captured by the equation

$$\frac{L_{g+1} - L_g}{L_g} = \left(\frac{w_g}{w^s}\right)^{\alpha} - 1\,, \quad \alpha > 0\,, \tag{1}$$

or

$$L_{g+1} = \left(\frac{w_g}{w^s}\right)^{\alpha} L_g\,, \quad \alpha > 0\,. \tag{2}$$

Let us also suppose that the production process involves only two inputs, land which is in fixed supply, and labor. To be concrete, suppose that a function of Cobb-Douglas form determines output for generation g,

$$Q_g = \rho L_g^{\lambda} R^{1-\lambda}\,, \quad 0 < \lambda < 1\,, \tag{3}$$

where L_g is the labor force and R (for resources) measures the fixed supply of land. Assuming that the market for labor is competitive, workers will be hired up to the point where the marginal product of labor is equal to the real wage; as was explained in Chapter 7.3; i.e.,

$$w_g = \frac{\partial Q_g}{\partial L_g} = \lambda \rho L_g^{\lambda-1} R^{1-\lambda} = \frac{\lambda Q_g}{L_g}\,. \tag{4}$$

Suppose we are given the subsistence wage w^s, the values of the parameters of the system and the initial size of the labor force for a particular generation, say L_0. Then it is possible to determine the future time path of the labor force. First we calculate Q_0 using equation (3). Then (4) yields w_0 so that we can finally calculate L_1 from L_0 using equation (2). Once we have L_1 we can repeat the procedure to calculate Q_1 and w_1 and then L_2 and so on into the indefinite future. Figure 12.2 indicates what happens. Note that both population and output will continue to grow, but at

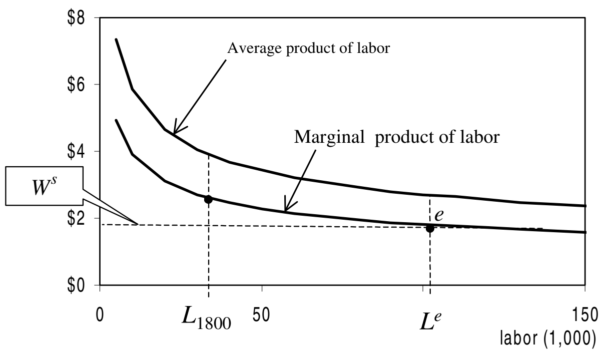


Fig. 12.2. Diminishing returns and the classical stationary state
Because of diminishing returns, the average and marginal products of labor are decreasing functions of the labor input. Since the supply of labor was initially 30,000 at L_{1800}, the wage will equal the marginal product of labor of approximately \$3.33. Because this wage is above subsistence wage w = \$2.00, the labor force expands, pushing down both the average and marginal product of labor. The labor force will continue to grow as long as $w > w$, which pushes us toward an unhappy equilibrium at point e where the wage is at the subsistence level of \$2 but output per worker is \$3.00.

slower and slower rates as the system gradually approaches the stationary equilibrium labor force L_e with wage w^s in the limit. Since output grows less rapidly than labor, thanks to the law of diminishing returns, the wage rate will inevitably be driven down to its subsistence level. The long-run equilibrium for this model is not a pretty sight.

Prediction fallacies

The pessimistic predictions of Malthus and the classical economists have been contradicted by history. Three factors account for the failure of this model to predict what happened.

1. The model fails to capture the upward shift in worker productivity brought about by the twin contributions of invention and capital accumulation. The steam engine, the internal combustion engine, electric power and now computers are but four of the inventions that have made decisive contributions to greater worker productivity. Malthus and the classical economists grossly underestimated the contributions of investment and technological progress.
2. Capital accumulation, made possible by thrift or abstinence from consumption, contributed to increased worker productivity.

   How technological progress can offset the effect of diminishing returns is illustrated by the total product curves plotted on Figure 12.3. The growth in population allows more and more workers to be employed with the passage of time. Output per capita would decline if the total project curve had remained unchanged at its 1800 level. But over time the development of better production techniques and the accumulation of physical capital shifted the total product curve upwards, thereby enabling a gradual increase in output per capita and rising living standards.
3. The assumption that the rate of population growth is governed by the gap between the wage received by workers and the subsistence wage, a critical assumption invoked to explain why the wage would be driven to the subsistence level, proved to be grossly inaccurate.

The law of diminishing returns may hold, given the technology and fixed supplies of natural resources and productive capital, but the historical record is clear: The unanticipated pace of technological advance coupled with substantial investment in productive capital has enabled output in the majority of countries to outstrip the growth of the workforce and has provided welcome increases in average living standards.

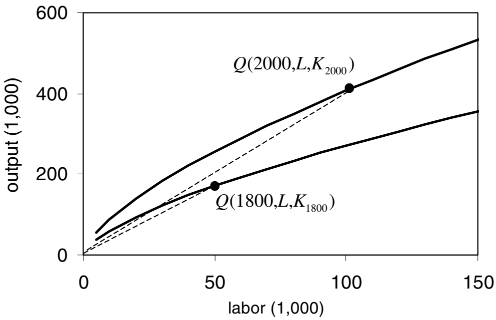


Fig. 12.3. Technological progress offsets the law of diminishing returns
Technological improvements and capital accumulation caused an upward shift in the total product curve from 1800 to 2000 that more than offset the tendency of the growing labor force to reduce the average and marginal products of labor. As a result, real wages increased rather than declining toward the subsistence level.

12.4 Growth accounting — The sources of economic growth

Output grows as a result of an increase in hours of work. It also increases if workers are able to work more efficiently because they are equipped with more capital equipment. And it will also grow because of the development of better production techniques. But how can we determine the relative importance of each of these factors in explaining economic growth? The question is of considerable policy interest. Definitely establishing that technological change plays the major role would imply that an increase in taxes to subsidize research and development might make a decisive contribution to greater growth. But if it turns out that investment is the decisive factor, then higher taxes, by discouraging thrift and investment, might slow the pace of economic development. If investment is the critical determinant, rapid growth might be encouraged by government subsidies and tax benefits promoting private investment spending.

Unfortunately, the contribution of technological improvements is hard to quantify. We have real GDP as a measure of total output and we can count the number of hours worked during the year, but how can we measure the contribution to worker productivity of the internal combustion engine, the assembly line, and the computer revolution? Almost a half century ago Nobel Laureate Robert Solow pioneered a procedure for estimating the contribution of technological progress that is still in use today.[1]

[1] Robert Solow, "Technical change and the aggregate production function," *Review of Economics and Statistics*, 1957.

Let us suppose that the aggregate output of the economy at time t is determined by the production function

$$Q(t) = f(t, L(t), K(t))\,, \tag{5}$$

where $L(t)$ is the number of labor hours and $K(t)$ is the stock of capital in year t. We include t as an argument in the production function to indicate that output would increase with the passage of time as a result of technological change, even if K and L were to remain constant.

Differentiating (5) with respect to time yields the total derivative:

$$\frac{dQ(t)}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial L}\frac{dL}{dt} + \frac{\partial f}{\partial K}\frac{dK}{dt}\,. \tag{6}$$

This total derivative says that the change in output is the sum of three components: first is the upward shift due to technological advance, second is the increase in output due to the growth of labor and third is the increase due to the availability of additional capital equipment. Note that the contribution of growing labor to increased output is the increase in labor dL/dt times the marginal productivity of labor.

We can manipulate (6) to obtain

$$\frac{dQ(t)/dt}{Q} = \frac{\partial f/\partial t}{Q} + \frac{\partial f}{\partial L}\frac{L}{Q}\frac{dL/dt}{L} + \frac{\partial f}{\partial K}\frac{K}{Q}\frac{dK/dt}{K}\,. \tag{7}$$

This simplifies to

$$q = \rho + \frac{\partial f}{\partial L}\frac{L}{Q}n + \frac{\partial f}{\partial K}\frac{K}{Q}k\,, \tag{8}$$

where $q = \frac{dQ(t)/dt}{Q}$ is the rate of growth of output, $n = \frac{dL/dt}{L}$ is the rate of growth of the labor-force, $k = \frac{dK/dt}{K}$ is the rate of growth of the capital stock, and $\rho = \frac{\partial f/\partial t}{Q}$ is the contribution of technological change. Thus we have decomposed the rate of growth of output into three components: the first is the contribution of technological change, the second is the contribution of labor-force growth, and the third is the contribution of the increased stock of capital.

In order to make the task of estimating the contribution of technological change manageable, Solow made two fundamental assumptions: He assumed that markets are competitive. He also assumed that the production function is homogeneous of degree 1 in capital and labor, which means that given the technology, doubling the quantities of both capital and labor will double output. Recall, once more that under competition the real wage equals the marginal product of labor; therefore, the coefficient of labor force growth in (8) is $wL/pQ = \lambda_L$, which is the concept of labor's share discussed in Chapter 7.3.3. By a parallel argument, the coefficient of the rate of growth in the capital stock k is equal to $rK/pQ = \lambda_K$, or capital's share, where r is the rental cost of capital. Now the assumption that the production function is homogeneous of degree 1 in capital and labor implies that $\lambda_L + \lambda_K = 1$.[2] If we follow Solow in invoking these two assumptions we have for the rate of output growth:

$$q = \rho + \lambda_L n + (1 - \lambda_L)k\,. \tag{9}$$

The rate of growth of per capita output, Q/L,[3] is

$$q - n = \rho + (1 - \lambda_L)(k - n)\,. \tag{10}$$

Solow recognized that the only unobservable in equation (9) is the rate of technological progress, ρ. So he calculated by subtraction what has ever since been known as the *Solow residual*:

$$\rho = q - \lambda_L n - (1 - \lambda_L)k\,. \tag{11}$$

Solow's residual estimate equals what is left over after the contributions of labor and capital growth are subtracted from the rate of growth of output.

Solow's Estimates

Applying residual equation (11) to annual data on q, n k and λ_L covering the period 1909 to 1949, Solow reported that ρ was about 1.2% per annum from 1909 to 1929 and about 1.9% per annum from 1929 to 1949. He concluded that about 7/8ths of the 80% increase in output per hour of work over the 40 year period was due to technological improvement and only 1/8th to an increase in the capital/labor ratio.

---

[2]Homogeneity of degree 1 in capital and labor means that, given the level of technological development at any point of time t, if we double labor and capital we will double output. More precisely, for any coefficient $\varphi, \varphi Q = f(t, \varphi L, \varphi K)$; differentiating both sides of the equality with respect to φ yields $Q = (df/dL)L + (df/dK)K$,which is an example of Euler's Theorem. Dividing by Q gives us $1 = \lambda \quad + \lambda \quad$, as required.

[3]To see why we just subtract n to get the rate of per capita output growth, note that $\ln Q/L = \ln Q - \ln L$. Hence, differentiating both sides with respect to t yields

$$\frac{d(Q/L)/dt}{Q/L} = \frac{dQ/dt}{Q} - \frac{dL/dt}{L} = q - n\,.$$

Productivity Slowdown/Productivity Spurt

Productivity growth is not a smooth and predictable process, as can be seen from the top row of Table 12.2, which summarizes evidence developed by Dale W. Jorgenson and Kevin J. Sitroh.[4] From the end of World War II until about 1973, productivity growth in the United States took place at a remarkable clip. But around 1973 the economy floundered in what is known as the "slowdown in the rate of productivity growth," or simply the "productivity slowdown." Because of this productivity slowdown, output per hour of work grew at a much slower rate for the next two decades than it had in the preceding quarter century. Starting around 1995, there was a substantial spurt in productivity growth.

The differences in terms of annual percentages are not large. But life would be so much brighter if there had not been a slowdown in the rate of productivity growth.

1. A simple exercise in counterfactual history shows that the slowdown in productivity growth resulted in a substantial loss of output. Suppose that from 1973 to 1998 the rate of growth in output per hour of work had remained at the earlier 2.948% per annum clip reported in the first column of the table. Then by the simple equation for compound interest, in 1998 output per hour of work would have been $(1.02948)^{98-73} = 2.07$ times what it was in 1973. Instead, because of the slowdown, output per hour of work grew to 1.46 times its level in 1973.[5] Or to put it another way, if there had been no slowdown in the rate of productivity growth, output in 1998 would have been $2.07/1.46 = 1.41$ times its actual level of 1998, given the number of hours worked. With the same work effort, 41% more would have been produced!
2. More rapid productivity growth would have substantially reduced inflationary pressure during the last quarter of the 20th century. The analysis of Chapter 11.3.4 suggests that inflation would have been less of a problem and the natural unemployment rate (or NAIRU) would have been lower if productivity had been growing more rapidly. It is fair to say that the productivity spurt in the late 1990s encouraged Fed Chairman Alan Greenspan to allow the unemployment rate to drop to 4% without imposing substantial monetary constraint.

[4]Dale W. Jorgenson and Kevin J. Stiroh, "Raising the speed limit: U.S. economic growth in the information age." *Brookings Papers on Economic Activity*, I:2000, p. 151.
[5]We have $1.02948^{(98-73)} = 2.067$ and $1.01437^{(90-73)} \times 1.01366^{(95-90)} \times 1.02271^{(98-95)} = 1.46$

Table 12.2. Sources of U.S. labor productivity growth.

	(percent per annum)			
	1959–1973	1973–1990	1990–1995	1995–1998
Growth in labor productivity (Q/L)	2.948	1.437	1.366	2.271
Components:				
Capital deepening (K/L)	1.492	0.908	0.637	1.131
Labor quality	0.447	0.200	0.370	0.253
Total factor productivity	1.009	0.330	0.358	0.987

Source: Jorgenson and Stiroh, p. 151 (see text)

The bottom section of Table 12.2 breaks the growth in labor productivity into three components. According to Jorgenson and Stiroh, capital deepening, the increase in capital per worker, accounted for at least half of the increase in output per worker in each of the time periods recorded on the table. Jorgenson and Stiroh show that improvement in labor quality has been a contributing if somewhat erratic factor in the growth process. Growth in total factor productivity constitutes the remaining source of productivity growth. The authors report that information technology — computer hardware, software and communications — made a significant contribution to the growth in productivity in the last decade of the 20th century.

12.5 A neo-classical model of the growth process

What determines in the long run whether an economy will grow or decay? What determines the rate of growth? And why do some countries remain dormant while others take off into self-sustained growth. We will consider a pioneering contribution toward the resolution of such questions that is provided by the neo-classical model of economic growth developed in the 1950s by Robert Solow.[6] In order to focus on the essential issues of the growth process, we shall assume that technological change takes place at a constant rate. We will also assume that the labor force will grow at a

[6] Robert M. Solow, "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, February, 1956. The model presented here differs from Solow's in several respects. In particular, Solow did not restrict the production function to be of Cobb-Douglas form, but he did require constant returns to scale in labor and capital. There are no fixed resources in the original Solow model. Also, the analysis here is further simplified by using discrete rather than continuous time.

constant rate forever more. To further simplify, the simple model presented here leaves out both international trade and the role of government.

Notation: Lower case letters will denote rates of growth. For example, $q_t = \ln Q_t - \ln Q_{t-1} \doteq (Q_t - Q_{t-1})/Q_t$, where Q_t is Net Domestic Product in period t.

12.5.1 *Assumptions*

Let us assume that output Q_t is a function of labor L_t, capital K_t, and land R, where the level of output at any point of time is determined by the following production function:

$$Q_t = \alpha(1+\rho)^t L_t^{\lambda} K_t^{\lambda'} R^{1-\lambda-\lambda'} . \tag{12}$$

This elaborates on the Cobb-Douglas production function (equation (6) of Chapter 5) in two fundamental respects: First, it includes R for resources in fixed supply, such as land, as an additional input. With $\lambda + \lambda' < 1$, the function is not homogeneous of degree one in capital and labor: we have diminishing returns to scale in the two variable inputs, implying that a doubling of labor and capital would not double output. Second, technological progress is captured by the term $(1+\rho)^t$.[7] With $\rho > 0$, this means that if L_t, K_t, and R were to remain unchanged output would still grow with the passage of time because of improving techniques of production.

It is also assumed, for simplicity, that the population grows at constant rate n:

$$N_t = N_0(1+n)^t . \tag{13}$$

Further, a constant portion γ of the population is employed. Presumably, the labor force participation rate is constant and the employed proportion of the labor force does not vary, either because of the economy's natural self-recuperating powers or because the central bankers succeed in keeping the economy moving along its full-employment growth path. Therefore, the labor supply grows at rate n:

$$L_t = \gamma N_t = \gamma N_0(1+n)^t = L_o(1+n)^t . \tag{14}$$

[7] Readers familiar with elementary differential equations may prefer to work in continuous rather than discrete time, substituting e for $(1+\rho)$ in equation (12) and e for $(1+n)$ in equations (13) and (14). See also footnote 8.

In addition, suppose that a constant fraction s of output is saved. Then consumption is $C_t = (1 - s)Q_t$. Since there is no government or foreign trade, $Q_t = C_t + I_t$ and we have net investment

$$I_t = sQ_t \,. \tag{15}$$

12.5.2 *Analysis*

As a first step toward determining the laws of motion for this dynamic model, we ask whether output can grow at some constant exponential rate, call it q^e. To find out, let us first take logs to the base e of (12), with the approximation $\ln(1 + \rho) = \rho$:[8]

$$\ln Q_t = \ln \alpha + (1 - \lambda - \lambda') \ln R + t\rho + \lambda \ln \gamma + \lambda \ln N_t + \lambda' \ln L_t \,. \tag{16}$$

This equation holds for all t, including $t - 1$:

$$\ln Q_{t-1} = \ln \alpha + (1 - \lambda - \lambda') \ln R + (t - 1)\rho + \lambda \ln \gamma + \lambda \ln N_{t-1} + \lambda' \ln L_{t-1} \,. \tag{17}$$

Subtracting equation (17) from (16) yields

$$\ln Q_t - \ln Q_{t-1} = \rho + \lambda(\ln L_t - \ln L_{t-1}) + \lambda'(\ln K_t - \ln K_{t-1}) \,. \tag{18}$$

Invoking the approximation that the difference in the logs of a variable is its rate of change [e.g., $\ln Q_t - \ln Q_{t-1} \doteq (Q_t - Q_{t-1})/Q_{t-1} = q$], we have:

$$q = \rho + \lambda n + \lambda' k \,, \tag{19}$$

where n is the constant rate of growth of the labor force and k is the rate of growth of the capital stock. This equation says that if output is to grow at a constant rate q^e then k, the rate of growth of the capital stock, must also be constant. More than this, from (15) we have

$$sQ_t/K_t = I_t/K_t = k \,. \tag{20}$$

This means that the capital stock can grow at a constant rate k only if the output capital ratio, Q_t/K_t is constant, but that requires that Q_t and K_t grow at the same rate; i.e. $k = q^e$ if output grows at a constant rate. To find q^e, substitute it for q and k in (19) to obtain:

---

[8] For example, if $\rho = 3\%$, $\ln(1 + \rho) \doteq 2.95588\%$ using the approximation discussed in Chapter 8.4.2. Alternatively, working in continuous time, as mentioned in footnote 7, we can differentiate (16) with respect to t to obtain equation (19) directly.

$$q^e = \rho + \lambda n + \lambda' q^e = \frac{\rho + \lambda n}{1 - \lambda'} . \tag{21}$$

The rate of growth of output per capita is $q^e - n$. Per capita output will increase along the equilibrium growth path, output growing faster than the population, if and only if

$$q^e - n = \frac{\rho + \lambda n}{1 - \lambda'} - n > 0 , \quad \text{or} \quad \rho > (1 - \lambda - \lambda')n . \tag{22}$$

The properties of this growth equilibrium are clarified with the aid of Figure 12.4, which plots the output/capital ratio on the abscissa and rates of growth on the ordinate. The ray emanating from the origin denotes the equation $k = sQ_t/K_t$, from (20). The line labeled q is obtained by substituting $k = sQ_t/K_t$ into (19) to obtain

$$q = \rho + \lambda n + \lambda' sQ_t/K_t . \tag{23}$$

The intercept of the q line is $\rho + \lambda n > 0$. The slope of the k line is s, which means that it is steeper than the q line, whose slope is only $\lambda' s$. Hence the two lines must intercept. At the point where the q and k lines cross, marked e on the graph, we obviously have $q = k$. With output and capital both growing at the same rate there is no tendency for the (Q_t/K_t) ratio to change. This equilibrium point is characterized by $q^e = k^e = I/K$.

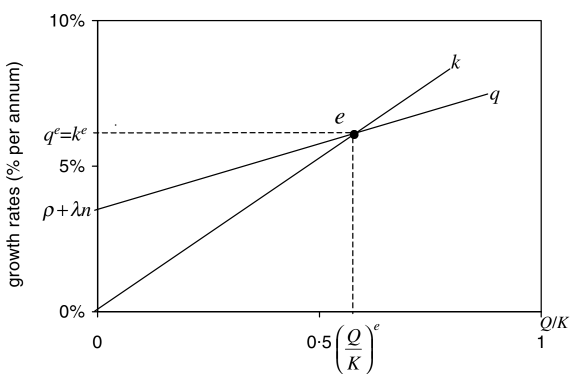


Fig. 12.4. Growth equilibrium
Both the growth rate of output (the q line) and the growth rate of the capital stock (the k line) depend on the Q/K ratio, plotted on the abscissa.

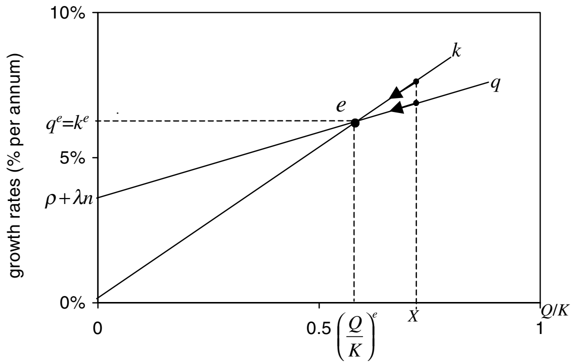


Fig. 12.5. Convergence to growth equilibrium
To see why the growth equilibrium at point e is stable, consider an emerging nation with a Q/K ratio above the equilibrium ratio, as at point X on the graph. Because the Q/K ratio is above the equilibrium value, K must be growing faster than Q (i.e., $k > q$), which means that the ratio Q/K must be falling toward the equilibrium value as indicated by the arrows on the graph.

Equations (20) and (21) imply that the corresponding equilibrium output/capital ratio is

$$\left(\frac{Q}{K}\right)^e = \frac{q^e}{s} = \frac{\rho + \lambda n}{s(1 - \lambda')} . \tag{24}$$

This growth equilibrium is stable. To see why, consider a country that has yet to realize its full development potential. Suppose initially the output/capital ratio is $(\frac{Q}{K}) > (\frac{Q}{K})^e$, as illustrated by point X on Figure 12.5. Since its output/capital ratio is high, $q > q^e$; our country will be growing above its equilibrium rate, as can be seen from equation (23). But $k > q$ implies that the Q/K ratio is falling with the passage of time. Thus Q/K will approach its equilibrium value as a limit, as indicated by the arrows on the graph.

12.5.3 *Growth or stagnation?*

The time path by which a developing nation may gradually move toward a happy growth equilibrium is recorded on Table 12.3 and plotted on Figure 12.6. Our emerging nation has a capital stock growing much more rapidly than output, which means that the capital/output ratio is on the rise and yields rising output per worker. While the process of converging to equilibrium can be quite slow, the end result is a country cruising along its

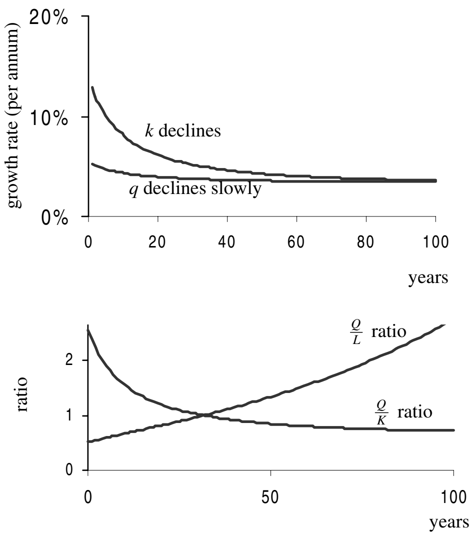


Fig. 12.6. Simulation #1: convergence to happy growth equilibrium
Top panel: The growth rates of output and of capital gradually decline toward the equilibrium growth rate.

Bottom panel: Output per worker, the Q/L ratio, grows rapidly as the economy moves into a more and more productive future. Since output per machine (Q/K) gradually declines while output per worker (Q/L) rises, the capital per worker ratio is increasing.

Parameter values: $\lambda = 0.65$, $\lambda' = 0.2$, $s = 5.0\%$, $n = 2\%$, $\rho = 1.4\%$,

Equilibrium values: $q \;\; = k \;\; = 3.3\%$; $(Q/K) \;\; = 0.67$; $q - n = 1.38\%$.

Table 12.3. Growth model simulation #1.

Parameter values: $\lambda = 0.65$, $\lambda' = 0.2$, $s = 5\%$, $n = 2\%$, $\rho = 1.4\%$, $K(0) = 20$, $N(0) = 100$, $R = 10$							
Year	N	K	Q	Q/K	q	k	Q/N
0	100	20	56	2.78			0.56
1	102	23	58	2.56	4.8%	13.9%	0.57
5	110	35	70	1.97	4.4%	10.4%	0.63
10	122	54	85	1.57	4.0%	8.2%	0.70
25	164	139	149	1.07	3.6%	5.5%	0.91
50	269	430	353	0.82	3.4%	4.1%	1.31
75	442	1,110	815	0.73	3.4%	3.7%	1.85
100	724	2,668	1,861	0.70	3.3%	3.5%	2.57
150	1,950	14,270	9,591	0.67	3.3%	3.4%	4.92
200	5,248	73,849	49,189	0.67	3.3%	3.3%	9.37

full-employment growth path with a constant rate of growth for both output and capital and a stable capital/output ratio, as specified by equations (21) and (24).

Figure 12.7 shows, for a different set of parameters, a most unhappy case in which the rate of growth of output is less than the rate of population growth, which means that the standard of living must inevitably decline! Comparison of the parameters with those of the earlier simulation reveals that this stagnant nation has a higher rate of population growth (n) coupled with a much less rapid pace of technological change (ρ).

Returning to equation (22), we find that improving this country's living standards would require a rate of technological change of at least 0.8% per annum, given the rapidly growing labor force coupled with the fact that certain resources R are in fixed supply means that the production function is subject to diminishing returns to scale in capital and labor,

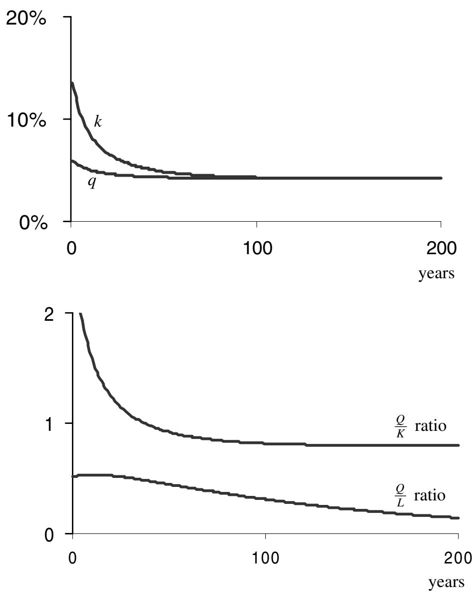


Fig. 12.7. Simulation #2: decline and fall
This nation enjoys a slight initial spurt of growth in output per capita, but after fewer then 20 years per capita income enters into a perpetual decline. The problem arises because the slow rate of technological advance is coupled with a high rate of population growth.

Parameter values: $\lambda = 0.65$, $\lambda' = 0.2$, $s = 5.25\%$, $n = 5\%$, $\rho = 0.1\%$,

Equilibrium values: $q = k = 4.2\%$; $(Q/K) = 0.8$; $q - n = -0.8\%$.

Table 12.4. Growth model simulation #2.

Parameter values: $\lambda = 0.65$, $\lambda' = 0.2$, $s = 5.25\%$, $n = 5\%$, $\rho = 0.1\%$, $K(0) = 20$, $N(0) = 100$, $R = 10$							
Year	N	K	Q	Q/K	q	k	Q/N
0	100	20	51	2.57			0.51
1	105	23	54	2.40	6.0%	13.5%	0.52
5	128	35	68	1.93	5.4%	10.6%	0.53
10	163	55	87	1.59	5.0%	8.6%	0.53
25	339	152	174	1.15	4.6%	6.1%	0.52
50	1,147	557	512	0.92	4.3%	4.9%	0.45
75	3,883	1,719	1,454	0.85	4.2%	4.5%	0.37
100	13,150	4,992	4,077	0.82	4.2%	4.3%	0.31
150	150,798	39,629	31,672	0.80	4.2%	4.2%	0.21

$\lambda + \lambda' = 0.85 < 1$. A slowing of the rate of population growth, as might be achieved through emigration or the encouragement of population control, might arrest the decline in living standards. Or more rapid technological advance might be achieved by borrowing state of the art techniques from more advanced nations or encouraged with government subsidies or tax breaks for research. If nothing is done, the grim predictions of Malthus will prove all too true, the decline in living standards continuing until the wage is driven below the subsistence level and the rate of population growth, g, is checked by starvation or disease.

It is intriguing to note from equation (22) that the equilibrium growth rate does *not* depend on s, which is the proportion of output that is saved for investment rather than consumed. However, s does affect the equilibrium capital/output ratio and the level of consumption at any particular point of time. Since $Q_t^{(1-\lambda')} = Q_t/Q_t^{\lambda'} = \alpha(1+\rho)^t L_t^{\lambda}(K_t/Q_t)^{\lambda'} R^{1-\lambda-\lambda'}$ from (12),

$$Q_t = (\alpha R^{1-\lambda-\lambda'})^{\gamma}(1+\rho)^{\gamma t} L_t^{\gamma\lambda}(K_t/Q_t)^{\gamma\lambda'}\,, \tag{25}$$

where $\gamma = 1/(1-\lambda')$. Substituting from (12) and (24) yields

$$Q_t^e = (\alpha R^{1-\lambda-\lambda'})^{\gamma}(1+\rho)^{\gamma t} L_o^{\gamma\lambda}(1+n)^{t\gamma\lambda}(s/q^e)^{\lambda\gamma'}\,. \tag{26}$$

Also, since $C_t = (1-s)Q_t$,

$$C_t^e = (1-s)Q_t^e\,. \tag{27}$$

Thus the height of the full employment growth path and the equilibrium consumption path are both affected by the savings ratio. In the longer run, other things being equal, two nations that are similar in terms of the pace of technological advance and the rate of population growth but with different saving ratios will end up growing at the same rate. But the saving ratio does matter, because one country may always enjoy a higher standard of living than the other at every point of time. The saving rate can be too high as well as too low. It can be shown that a country with a savings rate $s > \lambda'$ could enjoy a higher consumption path by reducing its savings rate.[9] A country can conceivably save too much, but $s > \lambda'$ means that saving is larger than capital's share in the nation's output!

12.5.4 *Why not growth?*

Because differences in living standards among nations are so huge, understanding why some nations prosper while others stagnate is one of the most pressing economic issues of our time. Many argue that secure property rights are a precondition for convergence of living standards among nations — who will invest if private property is not protected? Rapid development is said to be more likely when a country opens its doors to international trade, to foreign investment, and to the adoption of new technologies. Our growth model suggested that lagging nations will find it easier to catch up if more advanced nations are willing to share their advanced technology with less developed nations.

How willing countries are to import new technologies from more advanced nations may go a long way toward explaining why some nations experience development miracles while others stagnate. Indeed, Stephen L. Parente and Edward C. Prescott argue that countries stagnate because their governments discourage the adoption of new technologies. They explain that constraints on the adoption of modern techniques arise from the monopoly rights possessed by "industry insiders with vested interests tied to current production processes."[10] That is to say, the monopolists resist the adoption of new technologies that will undermine their monopoly position. According to their theory, Britain was the first to industrialize because the shift in power away from the crown to Parliament

---

[9] Edmund S. Phelps, "The Golden Rule of Accumulation: A Fable for Growthmen," *American Economic Review*, September, 1961.

[10] Stephen L. Parente and Edward C. Prescott, *Barriers to Riches*, MIT Press, 2000, p. 131.

led to a decline in regulation and meant that no group could successfully block the adoption of improved technologies. France, in contrast, was not hospitable to industrialization because the crown had sanctioned monopolies that were protected by elaborate regulations. Parente and Prescott argue that Japan's development miracle, a catching-up increase in the 1950s and 60s in their standard of living from only 20% to 75% of the U.S. standards, occurred because after World War II the U.S. occupying forces broke up much of Japan's industrial bureaucratic complex and succeeded in creating a more competitive environment.

Columbia University Professor Jeffrey Sachs argues that geography is a major determinant of economic growth and welfare.[11] Tropical countries tend to be underdeveloped, with the notable exceptions of Hong Kong and Singapore. Agricultural production is 50% less efficient in tropical countries, in part because of pests and parasites, soil erosion and problems of water availability. And the tropical countries have been falling further and further behind. Around 1820 GDP per capita in tropical regions was about 70% of GDP per capita in temperate zones, but by the 1990s GDP per capita in the tropics was only 25% of GDP per capita in temperate zone countries.

12.5.5 *Real business cycles*

The simple neoclassical growth model we have been discussing generates a smooth and steady path of economic development. This unrealistic result arises from the simplifying assumption that technological progress is a smooth and unbroken path and that the economy is not perturbed from its growth path by wars and other disturbances. More than a half century ago Harvard Professor Joseph Schumpeter [1883–1950] had argued that cyclical departures from the long run equilibrium growth path to which the economy naturally converges were an inherent part of the process of economic evolution. According to Schumpeter, the business cycle was part of the natural process by which the inherently stable economy responds to the shock of technological innovation. Downturns and recession are part of a process of "*creative destruction*" which contributes to economic development by weeding out the weak and unfit business enterprises and insuring the survival of the fittest.

[11] Jeffrey Sachs, "Tropical Underdevelopment," NBER Working Paper No 8119, February, 2001.

In 1982 Finn Kydland of Carnegie-Mellon University and Edward C. Prescott of the University of Minnesota published a pioneering article that led to the establishment of the Real Business Cycle school of macroeconomics.[12] Their analysis generated conclusions similar in a number of respects to Schumpeter's, but their analysis was based on a very sophisticated analytical foundation. They carefully developed the microfoundations of their model, invoking the Lucas supply function and rational expectations. They assumed that wages and prices are so flexible that they adjust promptly to balance the supply and demand for labor, arguing that fluctuations in the employment over the business cycle reflect voluntary adjustments in the labor supply to changes in real wages. Changes in the money supply, far from causing cyclical fluctuations, are endogenously generated when fluctuations in the pace of economic activity affect the demand for bank loans. According to the real business cycle theory, the economy is in continuous equilibrium, but equilibrium output fluctuates as a result of supply-side productivity shocks resulting from technological innovations and other disturbances, such as OPEC oil shocks and the aftermath of the September 11, 2001 terrorist attack. Arguing that the business cycle is the natural and efficient response of the economy to technological progress, real business cycle theorists conclude that attempts to smooth out the cycle, even if they worked, would be a mistake because they would generate harmful inefficiencies. Real Business Cycle theorists believe that recessions and unemployment are the price that must be paid for progress.

12.6 Population trends

The simplified neo-classical growth model presented in this chapter was optimistic about the future, provided that technological progress continues unabated. But admittedly, the model also involves a host of other major simplifications. In particular, it was assumed that the population grows at a constant rate, which is obviously far from the truth. Further, the model assumes that resources are never depleted! In fact, of course, some resources, such as oil, are subject to depletion while others, such as forests, are renewable. These issues deserve our attention. Let us start by looking at the population side of the Malthusian equation. What in fact has been the effect of rising worker productivity and higher living standards on the rate of population growth?

[12] "Time to build and aggregate fluctuations," *Econometrica*, 1982.

12.6.1 *The demographic transition*

Demographers analyze populations, gathering data, constructing models and interpreting population changes. Their studies of the *demographic transition*, the changes in the reproductive behavior of a country's population during the transformation from a traditional to a highly modernized state, reveal some surprising results. The demographers report that the classical assumption that the rate of population growth is an increasing function of the real wage could not be further from the truth. Quite the contrary, the transformation from a traditional pre-industrial society to a highly modernized state involves a dramatic rise in living standards coupled with a decisive *decline* in the birth rate. In the pre-industrialized state, high birthrates were balanced by low life expectancy. During the transition, mortality usually declined in advance of the decline in fertility, leading to a temporary spurt in the rate of population growth. It is generally true in most developed countries that women on average now bear only about half as many children as did their ancestors a century or two ago. But two centuries ago life expectancy was less than half of what it is today.

The decline in mortality in Europe, which began in the mid 18th century during the first phase of the demographic transition, resulted in large measure from dramatic improvements in health care, including improved sanitation, the pasteurization of milk, and vaccination for smallpox as well as tremendous advances in medical science. But what caused the decline in the birth rate? The customary explanation is that in earlier times children were a resource. They began work at an earlier age and were soon contributing more to the family than they consumed. Having a large family was also a means of providing for one's support if one were so fortunate as to live into old age. Contrast this with an advanced economy where child labor is generally outlawed, the costs of educating one's children can be substantial, and the emancipated young are said to make many trying demands on their parents. Further, lower child mortality rates reduce the number of births needed to achieve a family of targeted size. In addition, the development of pensions and social security provides an alternative to children as a source of support in one's old age.

According to projections by the United Nations Population Division, the population of the world is likely to grow from 5.7 billion in 1995 to about 9.4 billion in 2050 and 10.4 billion in 2100. The share of the world's population living in the currently more developed regions will decrease from

Table 12.5. Old age dependency ratios.

Year	Canada	France	Germany	Italy	United Kingdom	United States	Japan
	Retirement age (65 and older) as a percentage of the working age population (20–64 years)						
2000	20%	28%	25%	28%	27%	21%	27%
2025	36%	41%	36%	43%	36%	33%	37%

Source: United Nations.

19% to 10% in the next half century.[13] And declining fertility and mortality rates will lead to dramatic changes in the age-structure of the population. The share of the world population aged 60 and above will increase from 10% to 31% between 1995 and the middle of the 21st century. Table 12.5 shows how the old age *dependency ratio* is expected to increase in the next quarter century. For example, in Canada today there are about five people of working age for every senior citizen, but by 2025 there will be fewer than three working people for every senior citizen. No wonder many countries are worried about the financial viability of their social security systems. There will be major changes in career opportunities. One can expect to see a decline in the demand for teachers and an increase in demand for geriatric medical specialists and morticians.

12.6.2 *A simple overlapping generation model*

It is obvious that improved longevity will increase the average age of the population and may stress the financial viability of retirement programs. It is not so obvious that a reduction in the birth rate may have similar consequences. The hypothetical data presented on Table 12.6 are far from realistic, but they suffice for showing how changes in the rate of population growth can profoundly affect the age composition of the population, cause major shifts in the job market, influence the supply of aggregate savings and threaten the solvency of social security programs.

[13]The Population Division of the Department of Economics and Social Affairs at the United Nations Secretariat prepares population estimates and projections. Because long range projections are quite sensitive to changes in fertility rates, the demographers prepare low, medium and high estimates of likely population growth for alternative assumptions about fertility. Only the medium-fertility estimates are reported here. For more information, see
http://www.undp.org/popin/wdtrends/wdtrends.htm#World Population Estimates & Projections.

Table 12.6. Simple dynamics of population growth.

Panel 1: Population characteristics of Never-Never Land

Census year	0–20 youth	Age bracket 20–40 working	40–60 years	60–80 Retired	Total population	Adult (20+)	Average age	Retired/ working
1900	600	300	150	75	1,125	525	24.7	16.7%
1920	1,200	600	300	150	2,250	1,050	24.7	16.7%
1940	2,400	1,200	600	300	4,500	2,100	24.7	16.7%
1960	4,800	2,400	1,200	600	9,000	4,200	24.7	16.7%
1980	4,800	4,800	2,400	1,200	13,200	8,400	30.0	16.7%
2000	4,800	4,800	4,800	2,400	16,800	12,000	35.7	25.0%
2020	4,800	4,800	4,800	4,800	19,200	14,400	40.0	50.0%
2040	4,800	4,800	4,800	4,800	19,200	14,400	40.0	50.0%

Assumptions:

Everyone lives to be 80.

Until 1960 ever young working couple has four children.

After 1960 every young couple has two children (The pill or Roe vs Wade ???)

Panel 2: The teacher market in Never-Never Land

Census year	Youth	Students	Teachers	Teacher age 20–40	40–60	Average	Teacher/ Worker	Student/ Population
1900	600	300	15	10	5	36.7	3.3%	26.7%
1920	1,200	600	30	20	10	36.7	3.3%	26.7%
1940	2,400	1,200	60	40	20	36.7	3.3%	26.7%
1960	4,800	2,400	120	80	40	36.7	3.3%	26.7%
1980	4,800	2,400	120	40	80	43.3	1.7%	18.2%
2000	4,800	2,400	120	80	40	36.7	1.3%	14.3%
2020	4,800	2,400	120	40	80	43.3	1.3%	12.5%
2040	4,800	2,400	120	80	40	36.7	1.3%	12.5%

Assumptions:

50% of those in the 0–20 age bracket are students.

The student/teacher ratio is 20 to 1.

Panel 3: Consumption and saving in Never-Never Land

Census year	Adults	Workers	Income	Consumption Spending	Saving $S = Y - C$	Saving Ratio (S/Y)	Proportion Pop Retired
1900	525	450	1,350	1,050	300	22.2%	6.7%
1920	1,050	900	2,700	2,100	600	22.2%	6.7%
1940	2,100	1,800	5,400	4,200	1,200	22.2%	6.7%
1960	4,200	3,600	10,800	8,400	2,400	22.2%	6.7%
1980	8,400	7,200	21,600	16,800	4,800	22.2%	9.1%
2000	12,000	9,600	28,800	24,000	4,800	16.7%	14.3%
2020	14,400	9,600	28,800	28,800	0	0.0%	25.0%
2040	14,400	9,600	28,800	28,800	0	0.0%	25.0%

Assumptions:

Annual Wage Rate = \$3.00.

Each citizen's Lifetime Consumption = Lifetime Income (i.e., $C = \$2.00$ for both workers and retirees).

Panel 1 of Table 12.6 presents population data for Econoland, a mythical country where a census is taken every twenty years. In this grossly simplified economy, every family has four children and everyone lives for exactly 80 years. Everyone enters the work force at age 20 and works until age 60. Observe from the first several censuses that the population had been doubling every 20 years, thanks to the decision of every couple to have four children. Although everyone lives until age 80, the average age is not 40. The average age of the population is only 24.7 years because each successive generation is twice as large as its parents' generation.

Population momentum

After the 1960 census, perhaps as the result of the invention of a pill or the legalization of abortion, the number of children per family drops to two. As can be seen from the table, the number of citizens in the 0–20 age bracket stabilizes. But until 2020, the population continues to grow. *Zero population growth (ZPG)* is not reached until year 2020, when the cohort consisting of the 2,400 children born just in time to be measured in the 1940 census has died off and been replaced by a new cohort of 4,800 children. Thus the table illustrates the concept of *population momentum*: a long transitional period must pass before a change in child bearing behavior or mortality has its full impact on the rate of population growth and the age composition of the population. Adapting to the consequences of a change in the birthrate or mortality can require several generations.

Panel 1 also reveals several surprising demographic shifts. The average age increases from 24 to 40 as a result of the shift to zero population growth. And the ratio of retirees to workers rises from 1/8 to 1/2, which threatens the financial viability of the social security system. But while a higher percentage of the population is in retirement, the child proportion has shrunk and so the dependency ratio (the number of children plus retirees divided by the working population) is far below the rate prevailing in earlier times when the population was growing so rapidly.

Teacher job market

Panel 2 of Table 12.6 examines how the reduction in family size affects the job market for teachers. Under the assumption that half in the 0 to 20 age bracket are in school and that the average student teacher ratio is 20 to 1, 3.3% of the working population was in the teaching profession, with an average age of 36 years when the typical family had four children.

The population slowdown has a dramatic effect on the demand for new teachers. The 1980 census reports that there are only 40 teachers in the 20–40 age bracket — college graduates in the 80s or 90s who wanted to enter the teaching profession found that very few teachers were being hired — some would-be teachers joined the growing elder-care professions instead. Because the younger generation was unable to enter the teaching profession, the average teacher age climbed to 43 years. Once Zero Population Growth is reached, the proportion of workers who are employed as teachers is only 1.3%, less than half of the 3.3% in the days of larger families and steady population growth.

Saving ratio

Panel 3 of our table investigates the effect of zero population growth on the saving ratio. It is assumed that workers provide for their retirement by saving one-third of their income, which allows them to consume the same amount in retirement as when they were working.[14] With rapid population growth, the economy's savings rate was a high 22.2% because a very small proportion of the population was dissaving in retirement. With ZPG, 1/3rd of the adults are dissaving in retirement and the aggregate savings ratio is zero! This does not necessarily mean that the country will suffer from under-saving or over-consumption. A mature economy needs much less savings because it does not have to put as much aside for investment in the tools and equipment that were needed in the past for the growing generations of new workers. Thus the decline in the savings ratio may not be a bad thing!

12.7 Exhaustible resources

How soon will we run out of oil? Will the market mechanism appropriately allocate non-renewable resources (e.g., oil) over time? Many environmentalists answer such questions with a resounding no, arguing that government intervention is required to prevent the exploitation of our finite resources and protect the interests of future generations. The Solow style growth model we analyzed is not capable of analyzing these complications because

[14]In the Modigliani-Brumberg life cycle model of consumption, briefly discussed in Chapter 9.5.1, it is assumed that consumers put aside enough each year to be able to enjoy the same standard of living in retirement as in their working years but have no bequest motive.

it did not allow for exhaustible resources, such as coal and petroleum. A model appropriate for analyzing how the market allocates non-renewable resources, pioneered by Harold Hotelling, provides a surprisingly optimistic answer to these questions.[15]

12.7.1 *Two numerical examples*

We shall use a simple numerical example in analyzing exhaustible resources. Suppose there are 100,000 barrels of oil in Never-Never Land. The initial price of oil is $p_0 = \$5.00$ per barrel at the well-head and the rate of interest (borrowing or lending) is $i = 7\%$. How much oil should we pump next year? How much should we save for future generations? What will the market decide?

Suppose you own an oil well in Never-Never Land. Should you pump all your oil today for \$5.00 per barrel or should you leave it in the ground? The answer depends on what you think will happen to the price of oil. If you expect the price of oil to rise to less then \$5.35 in year two, you should pump all your oil today, sell it for \$5.00 per barrel and place the money in the bank earning 7% interest so as to get back \$5.35 in year two for each gallon of initial wealth. And other producers, if they have the same expectations you do, will also pump today, which will tend to push down today's price and raise tomorrow's. If, on the other hand, the price is expected to rise to more than \$5.35, you should leave your oil in the ground and pump tomorrow rather than hold money in the bank at only 7%. Further, if a speculator anticipates that the price of oil will rise by more than 7%, she can borrow from the bank at 7%, buy some oil, and inventory it for a year in order to turn a neat profit. All this means that competition among oil producers and speculators will tend to push next year's price to 1.07 times this year's price.

This argument underlies Professor Hotelling's proposition that in a competitive market with accurate expectations the price p_t of oil (net of extraction and refining costs) will increase at the same rate as the rate of interest:

$$p_{t+1} = (1 + i)p_t \,. \tag{28}$$

[15] Harold Hotelling "Economics of Exhaustible Resources," *Journal of Political Economy*, April 1931.

The price of gas at the pump might rise at a faster or lower rate, depending on the cost of extraction, transportation, and distribution, but the core price of petroleum rises at the same rate as money in the bank.[16]

It will be constructive to consider a numerical example. The initial stock of oil is 100,000 barrels and the demand function is $q = 50000p^{-1.1}$. Table 12.7 shows what happens. Whoops! The stock of oil was exhausted in year 25.

What went wrong? The problem is that the initial price of \$5 per barrel was too low. As a result there was excessive consumption in year 0, and since the price was rising at 7% per annum in accordance with equation (28), oil was under priced in future years as well! One suspects that long before the stock of oil was exhausted the rising consumption/stock ratio would signal that something was wrong. Speculators will sense an opportunity for profit. Rational speculators, anticipating the future shortage, will purchase

Table 12.7. Excessive depletion in Never-Never Land.

Initial oil stock = 100,000, Interest rate: 7.00%; Demand = $50000p^{-1\ 1}$				
Year	Price	Demand	Stock	Demand/Stock
0			100,000.00	
1	\$5.00	8,513.40	91,486.60	8.5%
2	\$5.35	7,902.80	83,583.80	8.6%
3	\$5.72	7,335.99	76,247.81	8.8%
4	\$6.13	6,809.83	69,437.98	8.9%
5	\$6.55	6,321.42	63,116.56	9.1%
6	\$7.01	5,868.03	57,248.54	9.3%
7	\$7.50	5,447.16	51,801.38	9.5%
8	\$8.03	5,056.48	46,744.90	9.8%
9	\$8.59	4,693.81	42,051.09	10.0%
10	\$9.19	4,357.16	37,693.93	10.4%
11	\$9.84	4,044.65	33,649.27	10.7%
12	\$10.52	3,754.56	29,894.71	11.2%
13	\$11.26	3,485.28	26,409.43	11.7%
14	\$12.05	3,235.30	23,174.13	12.3%
15	\$12.89	3,003.26	20,170.87	13.0%
23	\$22.15	1,655.83	2,731.58	37.7%
24	\$23.70	1,537.07	1,194.51	56.3%
25	\$25.36	1,426.83	0	119.4%

[16]The real rate of interest should be used if the price of oil is properly deflated by a general price index.

more oil and store it in anticipation of a tidy profit when the price rises more rapidly in response to the growing shortage. The higher price hike generated by spectator purchases will hold current consumption in check, thereby leaving more for future generations.

Table 12.8. Market equilibrium insures oil forever more.

Initial oil stock = 100,000; Interest rate: 7.00%; Demand = $50000p^{-1\ 1}$

Year	Price	Demand	Stock	Demand/Stock
0			100,000.00	
1	\$5.84	7,172.25	92,827.75	7.2%
2	\$6.25	6,657.83	86,169.92	7.2%
3	\$6.69	6,180.32	79,989.60	7.2%
4	\$7.16	5,737.05	74,252.55	7.2%
...				
19	\$19.75	1,878.70	24,315.27	7.2%
20	\$21.13	1,743.95	22,571.32	7.2%

Sum to Infinity: 100,000.00.

The next example is similar to that on Table 12.7, except that the initial price is \$5.84. At this price consumers will demand 7,172 barrels of oil in year 1. Thus the higher initial price is leading to more oil being put aside for future generations. And as is clear from the table, from this start 7.2% of the stock is consumed each year forever more. We never run out of oil! True, oil gets more and more expensive, and as its price rises consumers purchase less and less, perhaps by making use of alternative renewable resources, such as firewood, or hydroelectric power.

12.7.2 *Analysis*

Now we must examine how the size of the initial oil stock, the elasticity of demand for this exhaustible resource, and the rate of interest interact to determine the current price of petroleum. The elasticity of demand turns out to be critical, because it indicates how willing the public is to cut back on oil consumption when its price rises. The demand elasticity together with the rate of interest will determine how rapidly the public is weaned from petroleum.

Suppose that the demand for petroleum in year t is

$$q_t = \alpha p_t^{\eta}\,, \quad \eta < 0 \tag{29}$$

and that the price changes in accordance with Hotelling's principle at rate $i > 0$; i.e.,

$$p_t = (1+i)p_{t-1} = (1+i)^t p_0 \,. \tag{30}$$

Substituting into the demand equation reveals that the quantity of oil demanded in period t is

$$q_t = \alpha p_t^\eta = \alpha(1+i)^{\eta t} p_0^\eta \,. \tag{31}$$

To simplify notation, we let $\beta = (1+i)^\eta$ so that $p_t = \beta p_{t-1} = \beta^t p_0$ and equation (31) becomes

$$q_t = q_0 \beta^t \,. \tag{32}$$

If the initial stock of oil Q_0 is to be consumed over the infinite future with prices rising at rate i, we must have

$$Q_0 = q_0 + q_1 + q_2 + \cdots + q_t + \cdots = q_0(1 + \beta + \beta^2 + \ldots) \tag{33}$$

Now the expression in parentheses is the sum of a geometric series which must converge because $0 < \beta < 1$; therefore,

$$Q_0 = \frac{q_0}{(1-\beta)} \,. \tag{34}$$

Therefore, we must have first period consumption of

$$q_0 = (1-\beta)Q_0 = [1 - (1+i)^\eta]Q_0 \,, \tag{35}$$

which is achieved, given the demand function (31), with a first year price of

$$p_0 = \left(\frac{q_0}{\alpha}\right)^{1/\eta} = \left\{[1-(1+i)^\eta]\frac{Q_0}{\alpha}\right\}^{1/\eta} \tag{36}$$

Two points about this expression deserve special attention:

1. The smaller the price elasticity of demand, the lower the initial consumption, the higher the initial price and the less rapidly consumption is cut back with the passage of time, given the rate of interest.
   That is to say, the more vital the exhaustible resource is for our consumption, perhaps because there are few close substitutes, the less we should consume today and the more we ought to put aside for future generations.

2. The lower the interest rate the less we should consume today because the price will rise more slowly and the drop off in consumption will be less rapid, other things being equal.

This also makes sense. We provide for future generations in part by limiting our consumption of non-renewable resources and in part by accumulating physical capital, such as factories and equipment and housing. The interest rate, because it reflects the marginal productivity of funds invested in new capital equipment, balances these alternative means of providing for the future. If the marginal productivity of capital is high, investment in more productive capital equipment will provide for future generations more effectively than a larger petroleum stock. If the interest rate is low, investment in physical capital is less effective than providing for future generations by heightened conservation of non-renewable resources.

12.7.3 *Moral*

The moral of Hotelling's model is that competitive markets work to allocate non-renewable resources appropriately over time, giving proper weight to the interests of future generations. But the model is subject to serious question because it does not correctly predict the movements in the price of oil in recent decades. The price of petroleum products has increased at much too low a rate. Indeed, Table 8.8 revealed that the real price of gasoline in the United States was lower in 2000 than it had been in 1960. Hotelling's elegant principle is blatantly inconsistent with the facts. How can this be?

Part of the discrepancy may arise because the Hotelling principle applies only to the price of the raw material itself, while the price of gasoline at the pump reflects as well the costs of exploration, extraction, refining and distribution, which are not subject to the Hotelling principle. This means that the retail price of petroleum products should increase less rapidly than the rate of interest. Furthermore, rapid technological advance in exploration and drilling procedures have substantially reduced the cost of extraction.

While these two factors help to explain why petroleum prices have increased less rapidly than a naïve application of the Hotelling model would suggest, they are not the full story. Many of the richest sources of oil today are in countries that are politically unstable. Or in the jargon of economists, property rights are insecure. When revolution threatens, prudent oil barons will extract petroleum more rapidly because it is much safer to have your funds deposited in a Swiss bank account than to hold oil underground in the hope that you and your children will be able to get a better price in the future. The proper conclusion to be reached from Hotelling's model is that the market mechanism will work to allocate the use of non-renewable resources appropriately over time, but only if property rights are secure.

12.8 Renewable resources — Over-fishing

While some resources are exhaustible, so what is consumed today is not available for future generations to enjoy, others are renewable. Properly cared for, farmland, the forests and the oceans may be harvested in perpetuity. For such resources, we shall find, competition is not always for the best.

Generations of fisherman have assumed that the treasures of the sea are inexhaustible, but today it is all too clear that fishing resources are finite. Thanks in part to improved technology — electronic equipment for locating schools of fish, larger boats, and better nets — over-fishing has become a problem, and many fish stocks are collapsing. Here are some of the consequences:[17]

- In the past 20 years the average size of swordfish caught in the North Atlantic has dropped from about 265 pounds to 90 pounds. Most of the swordfish are being caught before they have had a chance to breed.
- With the passage in 1977 of the Magnuson Act, the U.S. took control of marine resources within 200 miles of the coast, driving out the factory ships from the Soviet Union, Japan, and Spain. But by 1980 the New England fishing fleet had expanded by 42% to take up the slack.
- After the Grand Banks stock of cod collapsed in the early 1990s, Canada closed its centuries old cod fishery, putting 30,000 people in Newfoundland out of work.
- The New England Groundfish Recovery Plan allows fishermen to spend only 88 days at sea each year catching groundfish.
- Georges Bank, 100 miles southeast of Cape Cod and for 500 years the worlds richest fishing ground, was closed to ground fishing when its stock of cod and haddock collapsed in the early 1990s, costing New England $350 million a year and 14,000 jobs.

[17] *Hartford Courant*, November 27, 1998.

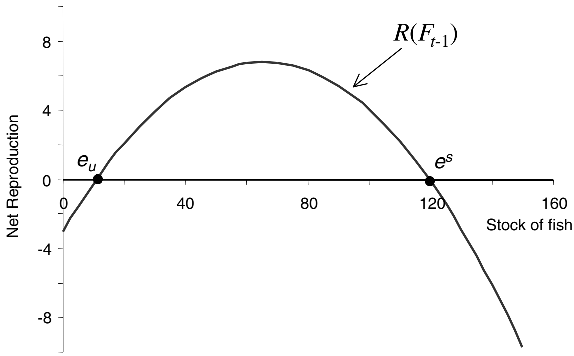


Fig. 12.8. Fish population and the balance of nature
The net-reproduction function $R(F_{\ -1})$ shows how the change in the stock of fish from one year to the next (the excess of births over deaths) depends on the number of fish already in the lake.

Point e is a stable equilibrium point characterized by zero population growth (ZPG). This point represents the balance of nature.

Point e is an unstable equilibrium. If the stock falls below e, the net-reproduction rate will be negative and extinction threatens.

Clearly, when the bounty of the sea is abused, the decline in the stock of fish is disruptive. It involves not only the loss of an important source of protein. It can also put thousands out of work. Fisherman must go further out to sea to make their catch. They now face the perils of collapsing fishing grounds as well as the traditional dangers of the sea. Restrictions on the number of days of fishing lead many fisherman to take on second jobs in the construction trades or elsewhere, but the boats are idle.

12.8.1 *Balance of nature*

In order to analyze some essential aspects of the problems of renewable resources, we shall construct a simple graphical model explaining the size of the population of a single species of fish.[18] The number of fish in the sea is plotted on the abscissa of Figure 12.8 while the change in the fish stock is on the ordinate. The curve labeled $R(F_{t-1})$ shows the net reproduction rate for the fish population. That is to say, $R(F_{t-1})$ is the excess of the

[18]For a clear mathematical analysis of renewable resources see David Levhari, Ron Michener and Leonard J. Mirman, "Dynamic Programming Models of Fishing: Competition," *American Economic Review*, September, 1981.

number of births over the number of deaths; it is the change in the number of fish from last year:

$$\Delta F_t = F_t - F_{t-1} = R(F_{t-1})\,. \tag{37}$$

If the population of fish is zero, it obviously remains at zero. Indeed a critical mass of fish (at least two) is required for reproduction to take place. Thus we must surely have $R(0) \leqslant 0$ and $R(1) \leqslant 0$. With more fish in the sea, more baby fish will be born and survive, and the faster the stock of fish will grow, but only up to a point. Once the fish population becomes large relative to the size of the sea, crowding and competition for food will reduce the number of young fish that survive and thrive. Therefore, for sufficiently large F_{t-1}, $dR/dF_{t-1} < 0$. More than this, if the fish population becomes much larger, the competition for food may be so intense that the number of deaths of older fish will exceed the number of replacement fish that hatch during the year; i.e., for F_{t-1} sufficiently large, $R(F_{t-1}) < 0$.

Point e_s on the graph represents zero population growth. If the fish population reaches magnitude e_s it will stay at that level indefinitely. Thus, e_s is an equilibrium fish stock. And it is a stable equilibrium: If there are fewer than e_s fish in the pond, the stock will expand; if there are more, the stock of fish will decline. Point e_u is also an equilibrium point, but it is unstable. If the number of fish is $e_u + \varepsilon$, $\varepsilon > 0$, $F_t > F_{t-1}$, and the population will continue to increase and we move further and further away from e_u; or if $\varepsilon < 0$, the population shrinks to zero (extinction). Point e_s may also be said to represent the *balance of nature*." Before fishermen began to harvest fish from the sea, the population will stabilize at e_s.

12.8.2 *Fishing*

Suppose that fishermen harvest Q fish from the sea each year. Let us see how this will upset the balance of nature. If Q fish are harvested each year, the fundamental equation describing how the population of fish changes is now

$$\Delta F_t = R(F_{t-1}) - Q\,. \tag{38}$$

What happens is shown on Figure 12.9. The ΔF_t function shifts down by the size of the annual catch. As a result, e_s is no longer viable. The population of fish will shrink toward the new equilibrium point e_s^*, and the fishermen can harvest Q fish from the sea every year forever more.

Over-fishing results if the fishermen try to harvest too many fish from the sea, as illustrated on Figure 12.10. The catch of magnitude Q is not sustainable, and the population of fish will "crash," threatening extinction! $Q^s = \max[R(F)]$ is the *maximum sustainable catch* that can be harvested from the sea for evermore.

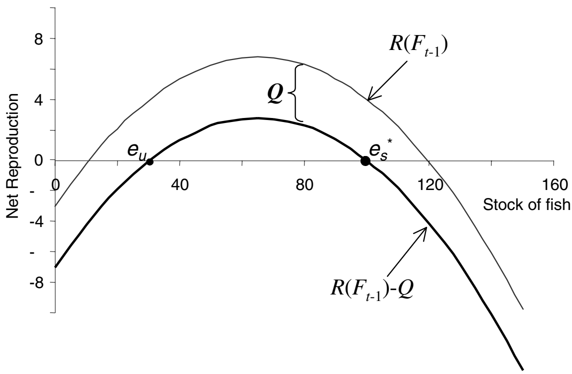


Fig. 12.9. Fishing
If fisherman harvest a catch of Q fish from the lake each year, the change in the number of fish from one year to the next will be $R(F_{-1}) - Q$. As a result the equilibrium stock of fish is smaller than it was in the state of nature.

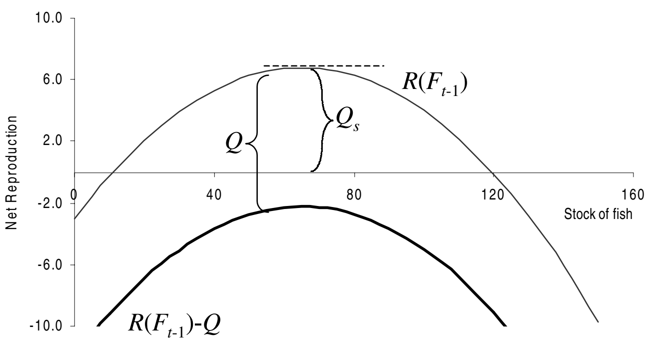


Fig. 12.10. Over fishing
Q is the maximum sustainable catch. When the catch Q exceeds the maximum sustainable catch, there is no equilibrium and the stock of fish shrinks. The fish stock may crash. Extinction threatens.

12.8.3 *Market equilibrium*

Assuming that fishing is a competitive industry with free entry, whether over-fishing occurs or not depends in part on the demand for fish and in part on the costs involved in harvesting the sea. Let $D(p)$ denote the demand function for fish and $S(p, F)$ the supply function. The stock of fish is included in the supply function because the more fish there are the easier they are to catch; which means that $\partial S/\partial F > 0$. On Figure 12.11 we have the demand curve and some representative supply curves. The more fish in the sea the larger the catch and the lower the price, as indicated by the succession of market equilibrium points on that graph. The short-run equilibrium price $p(F)$ and catch $Q(F)$ are determined by the market clearing requirement that $S(p, F) = D(p)$, This dependency of the size of the catch on the stock of fish, $Q(F)$, is plotted on Figure 12.12. The resulting long-run equilibrium stock of fish and the corresponding sustainable market determined catch is indicated by point e where $Q(F) = R(F_{t-1})$.

While the market determined catch displayed on Figure 12.12 is sustainable, this is not necessarily the case. The $Q(F)$ function will shift upwards, signifying that at any given stock of fish more will be harvested, if the demand curve for fish shifts upwards because of rising incomes or an expanding population. The $Q(F)$ function will also move upwards if the development of more efficient fishing techniques causes the supply function to shift downwards The result may be a crash in the stock of fish

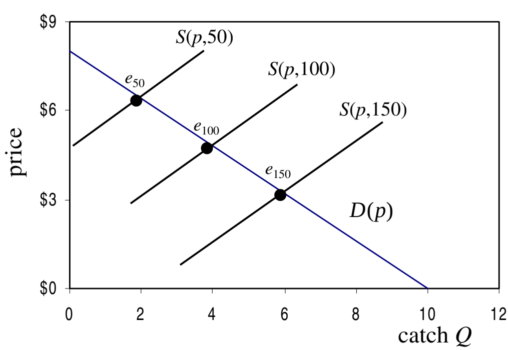


Fig. 12.11. Fish market equilibrium, short run
The more fish there are in the sea, the easier they are to catch. The family of short run supply curves, $S(p, F)$, reveal how the quantity of fish that will be brought to market as a function of the price depends on the number of fish in the sea. The short-run equilibrium points show how the price and quantity are simultaneously determined, given the stock of fish.

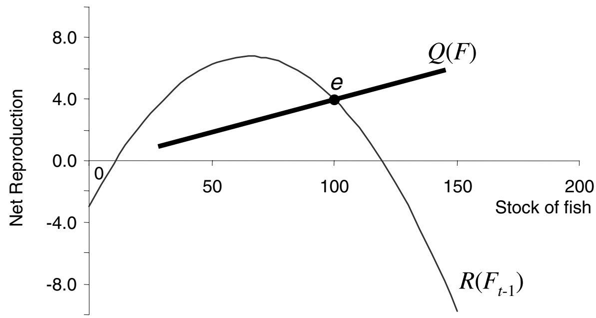


Fig. 12.12. Fish market equilibrium, long run
The $Q(F)$ function, derived from Figure 12.11, shows how the size of the catch depends on the number of fish in the sea. At long run equilibrium point e, the size of the catch just equals the reproduction rate; i.e., $Q(F\ \ ) = R(F\ \ )$.

A large upward shift in the demand for fish, function $D(p)$ on Figure 12.11, could push the $Q(F)$ function above $R(Ft-1)$. There would no longer be a fishing equilibrium point. Competition would generate over-fishing and cause the fish stock to crash. A tax on fish, a tax on fishing boats, quotas restricting the size of the catch, or a shortening of the fishing season might save the fish from extinction.

and disaster for the fishing industry. Measures that reduce $S(p, F)$, such as a tax on fishing or government regulation, such as quotas restricting the size of the catch or shortening the fishing season, may be the only way to prevent extinction.

Competition does not work to obtain economic efficiency in the case of fishing. The market mechanism fails to take into account an important cost of fishing activity. The external diseconomy in fishing arises from the fact that my fishing activity, by reducing the stock of fish, leads to an increase in your fishing costs. While all fishermen would gain from more limited fishing activity, each profit maximizing fisherman takes only his own fishing costs into account. A conservation aware fisherman who restricts production in the interest of future generations will suffer a loss in profits, unless all the other fisherman also sacrifice immediate profit by cutting back on their fishing activity.

12.9 Conclusions

This chapter has looked at a number of models. Some were failures. The classical model erroneously predicted stagnation at the subsistence level because the force of technological change and the contribution of capital accumulation were grossly underestimated. Hotelling's analysis of the way in which the market efficiently allocates petroleum and other resources in fixed supply erroneously predicted that the price of oil would increase overtime at a pace equal to the rate of interest. But even models that fail can be informative. The failure of the classical model teaches that the forces of technological change and capital accumulation must never be underestimated. The failure of Hotelling's model warns that when property rights are insecure the market cannot be relied upon to allocate our finite petroleum resources efficiently over time. The model used to analyze renewable resources — e.g., fish — did not fail. Its predictions are all too true. Once again we find that when property rights are insecure, competition can be extremely harmful. The neo-classical growth model provided a framework for analyzing what factors contribute to the growth of nations and the conditions necessary for the perpetual improvement in economic well-being.

Summary

1. Thomas Malthus argued that populations tend to grow geometrically while food production grows arithmetically, which means that the world will inevitably run out of food in the absence of population control. Adam Smith, David Ricardo and other members of the classical school worried that the economy would approach a stationary state characterized by a subsistence standard of living because the growing population would push the wage rate down to the subsistence level. The first growth model developed in this chapter approached the classical stationary state by assuming diminishing returns, by assuming that the rate of growth of population was proportional to the excess of the wage over the subsistence level, and by ignoring technological progress.
2. A neo-classical model of the growth process generated more optimistic results. This model assumed a constant rate of population growth and diminishing returns, but these factors could be offset by sufficiently rapid technological progress. It was shown that the model could approach an equilibrium characterized by a constant rate of output growth and a stable capital/output ratio. While the saving ratio does not affect the rate of growth of per capita income, it does affect the height of the equilibrium consumption growth path.
3. The neo-classical growth model predicts that poorer countries will tend to catch up with their wealthier neighbors if they have similar natural

resources, the same rate of population growth, and are able to adopt the technology enjoyed by their more advanced neighbors. But equation (22) revealed that decline rather than growth might be generated if the rate of technological progress is too low relative to the degree of diminishing returns. Parente and Prescott argue that countries fail to develop when vested interests protecting their investments in outdated technology prevent the adoption of more efficient production techniques. Jeffrey Sachs explains that tropical countries tend to be underdeveloped because agriculture is less efficient and disease is epidemic in tropical climates.

4. In contrast to classical theory, demographic studies reveal that populations grow less rapidly when wages move above the subsistence level. A simple overlapping generations model showed that it would take several generations for the population to stabilize after a change in the birth rate. A fall in the birth rate would lead to the aging of the population, subject the social security system to financial stress, adversely affect the market for teachers and cause the savings rate to decline.
5. The simple neo-classical growth model developed in this chapter did not allow for exhaustible resources, such as oil. The model of Hotelling predicts that the price of petroleum will rise at a rate equal to the rate of interest. As a result the consumption of oil will decline geometrically, but we will never run of out of oil. This has not happened to the price of oil, in part perhaps because the failure to secure property rights has encouraged excessively rapid depletion of petroleum resources.
6. Fish are a renewable resource, but they are subjected to over-fishing, resulting in the collapse of fishing stocks in many areas of the world and substantial economic loss to fisherman. The problem of over-fishing is said to arise because fish are a common resource unprotected by private property rights, which means that the market mechanism cannot function to allocate this scarce natural resource appropriately.

Key Concepts

balance of nature, 586
creative destruction, 572
demographic transition, 574
dependency ratio, 575
growth equilibrium, 566
growth model
  classical, 556
  neo-classical, 563
Malthus, Reverend Thomas R., 555
maximum sustainable catch, 587

Exercises

1. Suppose that output Q grows at rate $q = 4\%$, K grows at rate $k = 5\%$, and the population grows at rate 3%. How rapidly will Q/K change? How rapidly will per capita output grow?
   Hint: Consider footnote 3.
2. Never-Never Land has the following production function: $Q = (1.02)^t L^{2/3} K^{1/3}$.
   The savings rate is 10%. The labor force grows at 1% per annum.
   a. Derive the equilibrium growth rate of output, assuming that labor is always fully employed.
   b. Determine the equilibrium output/capital ratio.
   c. What is the equilibrium rate of growth of per capita income?
3. Congratulations, you have inherited an oil well from your late Uncle Rich. It is estimated to hold 500,000 barrels of oil. And the current price of oil is \$20 per barrel. You could pump the oil out of the well and put the money in a Swiss bank account, where it would earn 10% interest. Or you could leave the oil in the ground and pump it next year or ever further in the future if you like.
   a. How much oil will you pump this year *if* you think that next year the price of oil will be 12% higher than it is today? Explain why.
   b. Suppose that oil well owners pump some but not all of their oil out of the ground this year — some oil is left in the ground for another year. Suppose also that all oil well owners have the same expectations about future price increases. What must be the expected rate of increase in the price of oil?
4. The following function describes the net reproduction rate for fish in Lost Lake

$$R(F_{t-1}) = [25 - (F_{t-1} - 5)^2]^{1/2} - 1\,.$$

   a. Evaluate the net reproductive rate for $F_{t-1} = 0, 5, 8, 9$ and 10.

b. Plot the net reproduction rate as a function of F_{t-1} on a neat graph.
Hint: $(R+1)^2 + (F_{t-1} - 5)^2 = 25$ is the equation for a circle.

c. Determine the number of fish in Lost Lake in the state of nature (stable equilibrium).

d. Some fishermen find Lost Lake. Determine the equilibrium stock of fish if they catch only three fish per year.

e. Determine the maximum sustainable catch.

f. What will happen if our greedy fishermen catch six fish every year?

5.#* Solve the classical model summarized by equations (2) and (4) by finding the function $L_g = f(g, L_0)$.
Hint: Note from equation (4) that $\ln w_g = \ln(\lambda\rho R^{1-\lambda}) + (\lambda - 1)\ln L_g$. Substitute this expression into the ln transform of equation (2) in order to obtain a first order linear difference equation.

6.# Show that a savings rate equal to λ' will yield the maximum sustainable equilibrium consumption path characterized by a constant rate of growth.
Hint: Substitute equation 26 into 27 and differentiate with respect to s.

7.#* For the very first exercise in this book you were asked to write a summary of an article in a professional economics journal that was of particular interest to you. Go back and reread the article and your critique. You may be surprised at how much more sense that article makes to you now that you have finished this book. If you did not do this project when you finished Chapter 1, go back and reread the assignment (question 1.1) and do it now.

This page intentionally left blank

Appendix 1. Further Reading

Microeconomics (Chapters, 4–6)

Nicholson, Walter, *Microeconomic Theory: Basic Principles and Extensions*, 6th edition, 1995.

Varian, Hal, *Microeconomic Analysis*, Norton, 3rd edition, 1992.

Kreps, David M., *A Course in Microeconomic Theory*, Princeton University Press, 1990.

Tirole, Jean, *The Theory of Industrial Organization*, MIT Press, 1988.

More on Accounting (Chapter 6)

Davidson, Sidney and Roman Weil, *Accounting: the Language of Business*, Thomas Horton & Daughters, Inc.

More on Welfare (Chapters 4 and 7)

Rawls, John, *A Theory of Justice*, Harvard University Press, 1971.

Macroeconomics (Chapters 8–12)

Blanchard, Oliver, *Macroeconomics*, Prentice Hall, 2000.

Romer, David, *Advanced Macroeconomics*, McGraw Hill, 1996.

Diebold, Francis X., *Elements of Forecasting*, South Western, 2001.

More on Growth and Development (Chapter 12)

Barro, Robert J. and Xavier Sala-i-Martin, *Economic Growth*, MIT Press, 1999.

Parente, Stephen L. and Edward C. Prescott, *Barriers to Riches*, MIT Press, 2000.

More on Corporate Finance and the Stock Market

Ross, Stephen A., Randolph Westerfield and Jeffrey Jaffe, *Corporate Finance*, Irwin-McGraw-Hill, 1999.

Burton Malkiel, *A Random Walk Down Wall-Street*, a national best seller first published in 1973, presents the argument in favor of the hypothesis.

Andrew W. Lo and A. Craig MacKinlay present a number of empirical studies questioning the validity of the hypothesis in *A Non-Random Walk down Wall Street* by, Princeton University Press, 1999.

More on Game Theory

Avinash Dixit and Susan Skeath, *Games of Strategy*, Norton, 1999

Drew Fudenberg and Jean Tirole, *Game Theory*, MIT, 1991.

Statistics and Econometrics

Wonnacott, Thomas and Ronald J. Wonnacott, *Introductory Statistics for Business and Economics*, Wiley,4th edition, 1990.

Kennedy, Peter, *A Guide to Econometrics*, MIT Press, 1992.

Berndt, Ernst, R., *The Practice of Econometrics: Classic and Contemporary*, Addison Wesley, 1991.

More Mathematics

Chiang, Alpha C., *Fundamental Methods of Mathematical Economics*, Irwin/McGraw-Hill; 1984.

Michael Hoy, John Livernois, Chris McKenna, Ray Rees and Thanasis Stengos, *Mathematics for Economists*, MIT Press, 2001.

Information about graduate study in economics

(downloaded from the web pages of some leading graduate programs)

Harvard: Mathematics is an integral part of the preparation for graduate study in economics. Calculus through multivariate analysis (MATH 205) plus linear algebra (MATH 206) would provide adequate preparation, but you may benefit from differential equations (MATH 210) and real analysis (MATH 302). Thorough preparation in mathematics is as important as your economic courses.

Chicago: For admission to graduate study, a Bachelor's degree (or equivalent) is required; for some international students this may mean a degree beyond the baccalaureate. This degree need not be in economics. There are no formal course requirements for admission, but a strong background in mathematics is important. At the Ph.D. level, the study of economics requires a year of college calculus (at a minimum) and a quarter (or semester) each of both matrix algebra and mathematical statistics (that is, statistics using calculus, as distinct from introductory statistics for social science).

Berkeley: An undergraduate degree in economics is not required for admission to the Ph.D. program, provided that applicants have achieved adequate background in economics and mathematics at the undergraduate level. Applicants must have knowledge of multivariate calculus, basic matrix algebra, and differential equations; completion of a two-year math sequence that emphasizes proofs and derivations should provide adequate preparation.

This page intentionally left blank

Appendix 2. Web Resources

No printed list of web sites can possibly be up-to-date. A web search engine, such as Google, may provide instantaneous enlightenment. The first item on this list is a continuously updated guide to what is economics on the internet. I also maintain an updated copy of this list and other updates on my homepage:
http://mlovell.web.wesleyan.edu/EconCalc/MoreCalcIntro.html.

Resources for economists on the Internet:
http://www.aeaweb.org/RFE/~

> This wonderful guide to economic resources on the Internet, sponsored by the American Economic Association, will help you find whatever you want to know about the economy and economics. It catalogs sources of economic data for a large number of countries. Go to its Abridged Table of Contents for an overview of the resources that are available on the Internet.

Economists

Nobel Prize in Economics:
    http://castle.uvic.ca/econ/nobel.html.

Biographical Sketches of Economists:
    http://www.mtsu.edu/~tvs2/econ.html.

Economists on the World Wide Web:
    http://eclab.ch.pdx.edu/ecwww/.

American Economic Association Members:
    http://www.eco.utexas.edu/AEA/.

Simulations and Predictions

Iowa Electronic Market (Predictions of Presidential vote):
http://www.biz.uiowa.edu/iem.

Fair Model (Forecast the economy and the Presidential vote):
http://fairmodel.econ.yale.edu/.

Think Tanks

National Bureau for Economic Research (NBER):
http://www.nber.

Check out the NBER collection of Working Papers:
http://papers.nber.org/jel/.

Brookings Economic Studies:
http://www.brookings.org/es/es_hp.htm.

American Enterprise Institute:
http://www.aei.org/.

Sante Fe Institute Home Page:
http://www.santafe.edu/.

Miscellaneous

Economic Education Web Site:
http://ecedweb.unomaha.edu/teach-ec.htm.

History of Economic Thought Web Site:
http://cepa.newschool.edu/het/home.htm.

Eric Weinstein's World of Mathematics:
http://mathworld.wolfram.com/topics/.

Electronic Library Resources of Special Interest to students of economics (available only to users of subscribing libraries)

EconLIT: comprehensive bibliography with selected abstracts of the world's economic literature. It includes coverage of more than 400 major economic journals as well as articles in collective volumes (essays, proceedings, etc.), books, book reviews, dissertations, and working papers.

Social Science Citation Index: Once you have found an article in an economics journal that is of interest to you, the Social Science Citation Index can be used to find all the articles that have cited that paper. Check them out because they may be relevant to your research topic.

JSTOR: Contains the complete contents of many leading economic journals, including all the articles published in the American Economic Review from 1911–1996.

This page intentionally left blank

Index