AMBO UNIVERSITY WOLISO CAMPUS

SHOOL OF BUSINESS AND ECONOMICS

DEPARTMENT OF AGRICULTURAL ECONOMICS

COURSE TILTLE: ECONOMETRICS

COURSE CODE: AgEc2133

ECTS CREDITS (CP): 5

TARGET GROUP: II

INSTRUCTOR: TADESSE TOLERA (MSc.)

MAY, 2020

WOLISO, ETHIOPIA

1: INTRODUCTION

Definition: Econometrics deals with the measurement of economic relationships.

Econometrics is a combination of *economic theory*, *mathematical economics* and *statistics*, but it is completely distinct from each one of these three branches of science. The relationships and differences among these sciences are pointed out below.

A. **Economic theory** makes statements or hypotheses that are mostly *qualitative* in nature

Ex. Microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. But the theory itself does not provide any numerical measure of the relationship between the two: that is it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of econometrician to provide such numerical statements.

B. The main concern of *Mathematical economics* is to express *economic theory* in *mathematical form* (equations) without regard to measurability or empirical verification of the theory. Both economic theory and mathematical economics state the same relationships. Economic theory uses verbal exposition but mathematical economics employs mathematical symbolism. Neither of them allows for random elements which might affect the relationship and make it stochastic. Furthermore, they do not provide numerical values for the coefficients of the relationships.

Although econometrics presupposes the expression of economic relationships in mathematical form, like mathematical economics it does not assume that economic relationships are exact (deterministic).

- It assumes that relationships are not exact
- Econometric methods are designed to take in to account random disturbances which create deviations from the exact behavioral patterns suggested by economic theory and mathematical economics.
- Econometrics provides numerical values of the coefficients of economic phenomena.

C. **Economic Statistics** is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. It is mainly a descriptive aspect of economics. It does not provide explanations of the development of the various variables and it does not provide measurement of the parameters of economic relationships.

The econometrician often needs special methods since the data are not generated as the result of a controlled experiment. This creates special problems not normally dealt with in mathematical statistics. Moreover, such data are likely to contain errors of measurement, and the econometrician may be called up on to develop special methods of analysis to deal with such errors of measurement.

To Conclude: Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet, it is a subject that deserves to be studied in its own right for the above mentioned reasons.

1.2 GOALS OF ECONOMETRICS

Three main goals of econometrics

1. ***Analysis: - Testing Economic Theory***

Economists formulated the basic principles of the functioning of the economic system using verbal exposition and applying a deductive procedure. Economic theories thus developed in an abstract level were not tested against economic reality. Econometrics aims primarily at the verification of economic theories.

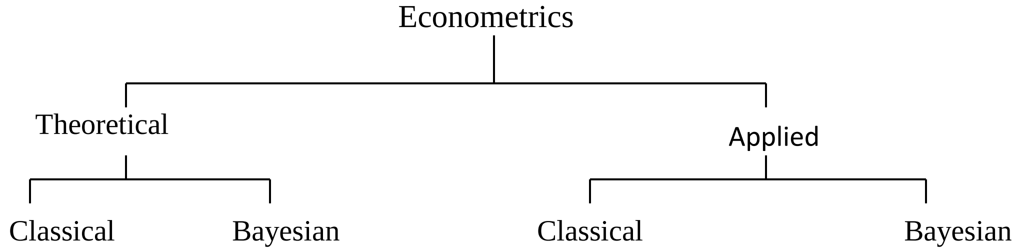
2. ***Policy-Making***

In many cases we apply the various econometric techniques in order to obtain reliable estimates of the individual coefficients of the economic relationships from which we may evaluate elasticities or other parameters of economic theory (multipliers, technical coefficients of production, marginal costs, marginal revenues, etc.) The knowledge of the numerical value of these coefficients is very important for the decisions of firms as well as for the formulation of the economic policy of the government. It helps to compare the effects of alternative policy decisions.

3. ***Forecasting***

In formulating policy decisions it is essential to be able to forecast the value of the economic magnitudes. Such forecasts will enable the policy-maker to judge whether it is necessary to take any measures in order to influence the relevant economic variables.

1.3 DIVISION OF ECONOMETRICS



Econometrics may be divided in to two broad categories

1. Theoretical Econometrics
2. Applied Econometrics

- Theoretical Econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. In this aspect, econometrics leans heavily on mathematical statistics. For example, one of the tools that is used extensively is the method of least squares. It is the concern of theoretical econometrics to spell out the assumptions of this method, its properties, and what happens to these properties when one or more of the assumptions of the method are not fulfilled.
- In applied Econometrics we use the tools of theoretical econometrics to study some special field(s) of economics, such as the production function, consumption function, investment function, demand and supply functions, etc.

Applied econometrics includes the applications of econometric methods to specific branches of economic theory. It involves the application of the tools of theoretical econometrics for the analysis of economic phenomena and forecasting economic behavior.

1.4 METHODOLOGY OF ECONOMETRICS

In any econometric research we may distinguish four stages:

A. Specification of the model

The first, and the most important, step the econometrician has to take in attempting the study of any relationship between variables is to express this relationship in mathematical form, that is to specify the model, with which the economic phenomenon will be explored empirically. This is called the specification of the model or formulation of the maintained hypothesis. It involves the determination of:

i) the dependent and explanatory variables which will be included in the model. The econometrician should be able to make a list of the variables that might influence the dependent variable.

. General economic theories,

. Previous studies in any particular field and

. Information about individual condition in a particular case, and the actual behavior of the economic agents may indicate the general factors that affect the dependent variable.

ii) the a priori theoretical expectations about the sign and the size of the parameters of the function. These a priori definitions will be the theoretical criteria on the basis of which the results of the estimation of the model will be evaluated

. Economic theory

. Other applied research

. Information about possible special features of the phenomena being studied will contain suggestions about the sign and size of the parameters.

Example: Consider the following simple consumption function:

$$C = B_0 + B_1Y + U$$

Where: C = Consumption function

Y = level of income

In this function the coefficient B_1 is the marginal propensity to consume (MPC) and should be positive with a value less than unity ($0<B_1<1$). The constant intercept, B_o of the function is expected to be positive. This is because when income is zero, consumption will assume a positive value; people will spend past savings, will borrow or find other means for covering their needs.

iii) the mathematical form of the model (number of equations liner or non-linear form of these equations, etc).

The specification of the econometric model will be based on economic theory and on any available information relating to the phenomenon being studied. The econometrics must know the general laws of economic theory, and furthermore, he must gather any other information relevant to the particular characteristics of the relationship as well as all studies already published on the subject by other research workers.

The most common errors of specification are:

- the omission of some variables from the functions
- the omission of some equations
- the mistaken mathematical form of the functions.

B. Estimation of the Model

Having specified the econometric model, the next task of the econometrician is to obtain estimates (numerical values) of the parameters of the model from the data available; consider the Keynesian consumption function.

$C = \beta_o + \beta_1 Y + U$

Where C is consumption

Y is income

If $\beta_1 = 0.8$ this value provides a numerical estimates of the marginal propensity to consume (MPC). If also supports Keynes' hypothesis that MPC is less than 1.

The stage of estimation includes the following steps.

- Gathering of statistical observations (data) on the variables included in the model
- Examination of the identification conditions of the function in which we are interested.
- Examination of the aggregation problems involved in the variables of the function.

- Examination of the degree of correlation between the explanatory variables.
- Choice of the appropriate econometric technique for the estimation of the function and critical examination of the assumptions of the chosen technique and of their economic implications for the estimates of the coefficients.

C. Evaluation of Estimates

After the estimation of the model the econometrician must proceed with the evaluation of the results of the calculations that is with the determination of the reliability of these results. The evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically satisfactory. Various criteria may be used.

- *Economic a prior criteria*: – These are determined by the principles of economic theory and refer to the sign and the size of the parameters of economic relationships. In econometric jargon we say that economic theory imposes restrictions on the signs and values of the parameters of economic relationships.
- *Statistical criteria*: – These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. The most widely used statistical criteria are the **correlation coefficient** and the **standard deviation (or the standard error)** of the estimates. These concepts will be discussed in the subsequent units. Note that the statistical criteria are secondary only to the a priori theoretical criteria. The estimates of the parameters should be rejected in general if they happen to have the wrong sign or size even though the pass the statistical criteria.
- *Econometric criteria*: – are determined by econometric theory. It aims at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. When the assumptions of an econometric technique are not satisfied it is customary to re specify the model.

D. Evaluation of the forecasting power of the estimated model

The final stage of any econometric research is concerned with the evaluation of the forecasting validity of the model. Estimates are useful because they help in decision-making. A model, after the estimation of its parameters, can be used in forecasting the values of economic variables. The econometrician must ascertain how good the forecasts are expected to be in other words he must test the forecasting power of the model.

It is conceivably possible that the model is economically meaningful and statistically and econometrically correct for the sample period for which the model has been estimated, yet it may very well not be suitable for forecasting due, for example, to rapid change in the structural parameters of the relationship in the real world.

Therefore, the final stage of any applied econometric research is the investigation of the stability of the estimates, their sensitivity to changes in the size of the sample.

One way of establishing the forecasting power of a model is to use the estimates of the model for a period not included in the sample. The estimated value (forecast value) is compared with the actual (realized) magnitude of the relevant dependent variable. Usually there will be a difference between the actual and the forecast value of the variable, which is tested with the aim of establishing whether it is (statistically) significant. If after conducting the relevant test of significance, we find that the difference between the realized value of the dependent variable and that estimated from the model is statistically significant, we conclude that the forecasting power of the model, its extra – sample performance, is poor.

Another way of establishing the stability of the estimates and the performance of the model outside the sample of data from which it has been estimated, is to re-estimate the function with an expanded sample, that is a sample including additional observations. The original estimates will normally differ from the new estimates. The difference is tested for statistical significance with appropriate methods.

Reasons for a model's poor forecasting performance

a) The values of the explanatory variables used in the forecast may not be accurate
b) The estimates of the coefficients $(\beta's)$ may be poor, due to deficiencies of the sample data.

c) The estimates are 'good' for the period of the sample, but the structural background conditions of the model may have changed from the period that was used as the basis for the estimation of the model, and therefore, the old estimates are not 'good' for forecasting. The whole model needs re-estimation before it can be used for prediction.

Example . Suppose that we estimate the demand function for a given commodity with a single equation model using time-series data for the period 1950 – 68 as follows

$$\hat{Q}_t = 100 + 5Y_t - 30P_t$$

This equation is then used for 'forecasting' the demand of the commodity in the year 1970, a period outside the sample data.

Given $Y_{1970} = 1000$ and $P_{1970} = 5$

$$\hat{Q}_t = 100 + 5(1000) - 30(5) = 4,950 \text{ units.}$$

If the actual demand for this commodity in 1970 is 4, 500 there is a difference of 450 between the estimated from the model and the actual market demand for the product. The difference can be tested for significance by various methods. If it is found significant, we try to find out what are the sources of the error in the forecast, in order to improve the forecasting power of our model.

1.5 THE NATURE AND SOURCES OF DATA FOR ECONOMETRIC ANALYSIS

The success of any econometric analysis ultimately depends on the availability of the appropriate data. Let us first discuss the types of data and then we will see the sources and limitations of the data.

1.5.1 Types of Data

There are three types of data

a) Time series data

This is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals: daily, weekly, monthly, quarterly, annually etc.

Example. data on stock prices, unemployment rate, GDP etc

Data may be qualitative or quantitative

Qualitative data are sometimes called dummy variables or categorical variable. These are variables that cannot be quantified.

Example: male or female, married or unmarried, religion, etc

Quantitative data are data that can be quantified

Example: income, prices, money etc.

b) Cross-Section data

These data give information on the variables concerning individual agents (consumers or producers) at a given point of time.

Example:

- the census of population conducted by CSA.

-survey of consumer expenditure conducted by Addis Ababa university

Note that due to heterogeneity, cross- sectional data have their own problems.

c) Pooled Data

These are repeated surveys of a single (cross-section) sample in different periods of time. They record the behavior of the same set of individual microeconomic units over time. There are elements of both time series and cross sectional data.

The panel or longitudinal data also called micro panel data, is a special type of pooled data in which the same cross-sectional unit is surveyed over time.

1.5.2 The Sources of Data

A governmental agency, an international agency, a private organization or an individual may collect the data used in empirical analysis.

Example. Governmental in Ethiopia: - MEDAC, MOF, CSA, NBE

International agencies: - International Monetary Fund (IMF), World Bank (WB)

The individual (researcher) himself may collect data through interviews or using questionnaire.
In the social sciences the data that one generally obtains is non-experimental in nature; that is not subject to the control of the researcher. For example, data on GNP, unemployment, stock prices

etc are not directly under the control of the investigator. This often creates special problems for the researcher in pinning down the exact cause or causes affecting a particular situation.

Limitations

Although there is plenty of data available for economic research, the quality of the data is often not that good. Reasons are:

- Since most social science data are not experimental in nature, there is the possibility of observational errors.
- Errors of measurement arising from approximations and round offs.
- In questionnaire type surveys, there is the problem of non-response
- Respondents may not answer all the questions correctly
- Sampling methods used in obtaining data
- Economic data is generally available at a highly aggregate level. For example most macro data like GNP, unemployment, inflation etc are available for the economy as a whole.
- Because of confidentiality, certain data can be published only in highly aggregate form For example, data on individual tax, production, employment etc at firm level are usually available in aggregate form.

Because of all these and many other problems, the researcher should always keep in mind that the results of research are only as good as the quality of the data. Therefore, the results of the research may be unsatisfactory due to the poor quality of the available data (may not be due to wrong model)

1.8 REFERENCES

Gujarati, D., Basic Econometrics.

Kmenta, J., Elements of Econometrics, Macmillan, New York, 1971

Koutsoyiannis, A., Theory of Econometrics, 2nd ed. Pal grave, 1977.

2. CORRELATION THEORY

2.1. Basic concepts of Correlation

Correlate has a relationship or connection in which one thing affects or depends on another. It implies establish a correlation between each of two or more related or complementary things.

Correlation is a mutual relationship or connection, the process of correlating two or more things. It is the extent to which two variables are interdependent. Unlike regression, this calculation is not used to predict the value of one variable from the other. **It is Statistics** interdependence of variable quantities.

2.2. Coefficient of Linear Correlation

❖ **Correlation Coefficient**

Correlation coefficient is a statistical number between +1 and -1 calculated so as to represent the linear interdependence of two variables or sets of data. (Symbol: **r**.)

✓ The quantity *r*, called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson.

✓ The mathematical formula for computing *r* is:

$$r \;=\; \frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{\mathrm{n}\sum X^2-(\sum X)^2}\sqrt{\mathrm{n}\sum Y^2-(\sum Y)^2}} \quad \text{or} \quad r \;=\; \frac{\sum xiyi}{\sqrt{\sum xi^2}\,\sqrt{\sum yi^2}}$$

where *x*i = Xi- $\overline{\mathrm{X}}$, yi = Yi-$\overline{\mathrm{Y}}$

where *n* is the number of pairs of data.

An increase in one variable may cause an increase in the other variable, or a decrease in one variable may cause decrease in the other variable. When the variables move in the same direction like this they are said to be positively correlated. The positive correlation may be termed as direct correlation. If a decrease in one variable causes an increase in the other variable or vice versa, the variables are said to be negatively correlated. The negative correlation may be termed as inverse correlation. In case the two variables are not at all related they are said to be independent or uncorrelated

✓ The value of r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

- ✓ Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase.
- ✓ Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- ✓ No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables
- ✓ Note that r is a dimensionless quantity; that is, it does not depend on the units employed.
- ✓ A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If *r* = -1, the slope of this line is negative.
- ✓ A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Properties of simple correlation coefficient

Coefficient of correlation lies between $-1 \leq r \leq 1$

If *r* = 0 indicate that there is no linear relationship between two variables.

If *r* = -1 or +1 indicate that there is perfect negative (inverse) or positive (direct) linear relationship between two variables respectively.

A coefficient of correlation(r) that is closes to zero shows the relationship is quite weak, whereas *r* is closest to +1 or -1, shows that the relationship is strong.

Note that

□□The strength of correlation does not depend on the positiveness and negativeness of *r* .

□□The slope of simple linear regression (coefficient of regression) and correlation coefficient should be the same in sign.

The correlation between two variables is linear if a unit changes in one variable result in a constant change in the other variable. Correlation can be studied through plotting scattered diagrams.

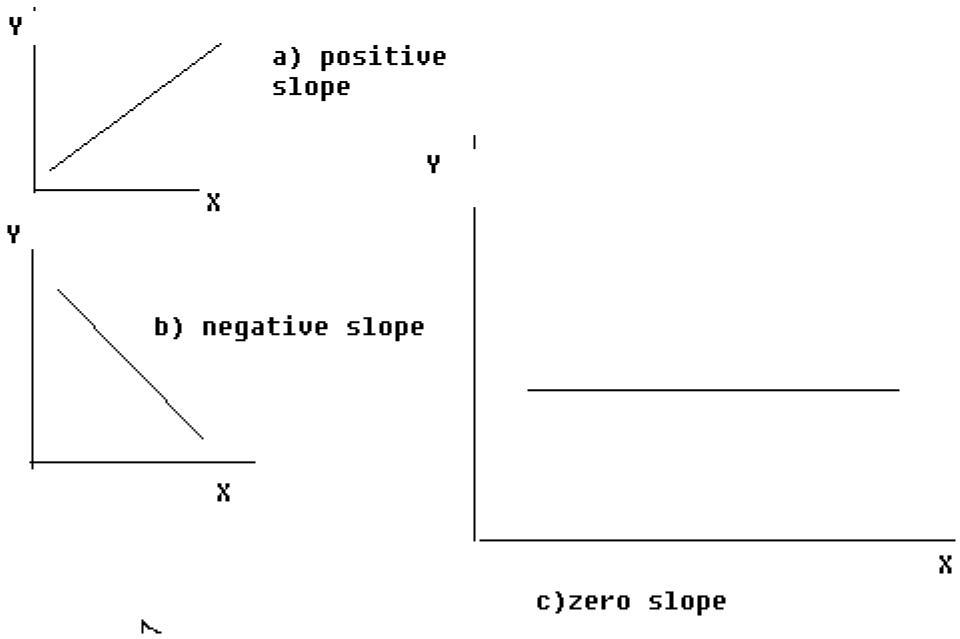


Figure: Slopes of simple linear regression lines

2.3. Types of Correlation Coefficient

Simple correlation coefficient: It is computed for continuous variables (interval ratio). It is developed by Karl Pearson and it is sometimes said to be Pearson correlation coefficient. It is computed based on data if Y is dependent variable on X as follows:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\mathrm{n}\sum X^2 - (\sum X)^2}\sqrt{\mathrm{n}\sum Y^2 - (\sum Y)^2}} \quad \text{or} \quad r = \frac{\sum xiyi}{\sqrt{\sum xi^2}\sqrt{\sum yi^2}}$$

where *x*i = Xi- $\overline{\mathrm{X}}$, yi = Yi-$\overline{\mathrm{Y}}$

where *n* is the number of pairs of data.

Example: If the following data is given for you as data collected from a given market on quantity and price find the simple correlation coefficient and discuss the type of relationship between the variables. Based on your knowledge of economics and simple correlation coefficient, what are the variables expressed as quantity?

Q	**P**	**Q^2**	**P^2**	**PiQi**	***qi***	***pi***	***qi^2***	***pi^2***	***piqi***
10	2	100	4	20	-51	-9	2601	81	459
20	4	400	16	80	-41	-7	1681	49	287
50	6	2500	36	300	-11	-5	121	25	55
40	8	1600	64	320	-21	-3	441	9	63
50	10	2500	100	500	-11	-1	121	1	11
60	12	3600	144	720	-1	1	1	1	-1
80	14	6400	196	1120	19	3	361	9	57
90	16	8100	256	1440	29	5	841	25	145
90	18	8100	324	1620	29	7	841	49	203
120	20	14400	400	2400	59	9	3481	81	531
SUM=610	**110**	**47700**	**1540**	**8520**	**0**	**0**	**10490**	**330**	**1810**

Rank Correlation

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement, but can be arranged in serial order. This happens when we dealing with qualitative characteristics (attributes) such as beauty, efficient, honest, intelligence ….etc. in such case one may rank the different items and apply the spearman method of rank difference for finding out the degree of relationship. The greatest use of this method (rank correlation) lies in the fact that one could use it to find correlation of qualitative variables, but since the method reduces the amount of labor of calculation, it is sometimes used also where quantitative data is available. It is used when statistical series are ranked according to their magnitude and the exact size of individual item is not known. Spearman's correlation coefficient is denoted by r'. **Steps** r'

i. Rank the different items in X and Y.

ii. Find the difference of the ranks in a pair, denote them by di

iii. Use the following formula

$$r \ = 1 - \frac{6\sum di^2}{n(n^2-1)}$$

Where di is the difference between ranks of corresponding pairs of X and Y

Example:

A market researcher asks two experts to express the preferences for 12 different brands of Soap

Brands of soap	X	Y	Di	Di^2
A	9	7	2	4
B	10	8	2	4
C	4	3	1	1
D	1	1	0	0
E	8	10	-2	4
F	11	12	-1	1
G	3	2	-1	1
H	2	6	-4	16
I	5	5	0	0
J	7	4	3	9
K	12	11	1	1
L	6	9	-3	9
				50

$$r' = 1 - \frac{6\sum di^2}{n(n^2-1)}$$

$r' = 1 - \frac{6(50)}{12(144-1)} = 1 - 300/1716$ = 1416/1716 = **0.825**

3. SIMPLE LINEAR REGRESSION MODELS

3.1. Basic Concepts and Assumptions

3.1.1 The Modern Interpretation of Regression

Broadly speaking, we may say Regression analysis is concerned with the study of the dependence of one variable, the *dependent variable*, on one or more other variables, the *explanatory variables*, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

Example

Dependent Variable Y; Explanatory Variable Xs

1. Y = Personal Consumption Expenditure X = Personal Disposable Income

2. Y = Demand; X = Price

3. Y = % Change in Demand; X = % Change in the advertising budget

4. Y = Crop yield; Xs = temperature, rainfall, sunshine, fertilizer

Statistical versus Deterministic Relationships

In statistical relationships among variables we essentially deal with **random** or **stochastic** variables, that is, variables that have probability distributions. In functional or deterministic dependency, on the other hand, we also deal with variables, but these variables are not random or stochastic. In regression analysis, we are concerned with STATISTICAL DEPENDENCE among variables (not Functional or Deterministic), we essentially deal with RANDOM or STOCHASTIC variables (with the probability distributions).

Regression versus Causation

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. In the words of Kendall and Stuart, “A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other.”

Regression versus Correlation

Closely related to but conceptually very much different from regression analysis is **correlation analysis,** where the primary objective is to measure the *strength* or *degree* of *linear association* between two variables. In regression analysis, however, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables.

3.1.2 TERMINOLOGY AND NOTATION

In the literature the terms *dependent variable* and *explanatory variable* are described variously. A representative list is:

Dependent Variable

↑↓

Explained Variable

↑↓

Predictand

↑↓

Regressand

↑↓

Response

↑↓

Endogenous

Explanatory Variable(s)

↑↓

Independent Variable(s)

↑↓

Predictor(s)

↑↓

Regressor(s)

↑↓

Stimulus or control variable(s)

↑↓

Exogenous (es)

If we are studying the dependence of a variable on only a single explanatory variable, such as that of consumption expenditure on real income, such a study is known as ***simple***, or **two-variable, regression analysis.** However, if we are studying the dependence of one variable on more than one explanatory variable, as in the crop-yield, rainfall, temperature, sunshine, and fertilizer examples, it is known as **multiple regression analysis.**

3.1.3 Types of Data Required for Economic Analysis

The success of any econometric analysis ultimately depends on the availability of the appropriate data. Three types of data may be available for empirical analysis: **time series, cross-section,** and **pooled** (i.e., combination of time series and cross section) data.

Time Series Data

A *time series* is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as **daily** (e.g., stock prices, weather reports), **weekly** (e.g., money supply figures), **monthly** [e.g., the unemployment rate, the Consumer Price Index (CPI)], **quarterly** (e.g., GDP), **annually** (e.g., government budgets).

Cross-Section Data Cross-section data are data on one or more variables collected *at the same point in time.*

Pooled Data In pooled, or combined, data are elements of both time series and cross-section data.

3.1.4 A Note on the Measurement Scales of Variables

The variables that we will generally encounter fall into four broad categories: *ratio scale, interval scale, ordinal scale, and nominal scale.* It is important that we understand each.

Ratio Scale For a variable X, taking two values, X_1 and X_2, the ratio X_1/X_2 and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a natural ordering (ascending or descending) of the values along the scale. Therefore, comparisons such as $X_2 \leq X_1$ or $X_2 \geq X_1$ are meaningful. Most economic variables belong to this category. E.g., it is meaningful to ask how big this year's GDP is compared with the previous year's GDP.

Interval Scale An interval scale variable satisfies the last two properties of the ratio scale variable but not the first. Thus, the distance between two time periods, say (2000–1995) is meaningful, but not the ratio of two time periods (2000/1995).

Ordinal Scale A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering). Examples are grading systems (A, B, C grades) or income class (upper, middle, lower). For these variables the ordering exists but the distances between the categories cannot be quantified.

Nominal Scale Variables in this category have none of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories.

3.1.5. Two-Variable Regression Analysis: Some Basic Ideas

A HYPOTHETICAL EXAMPLE

As noted in Section 2.1.1, regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable(s). To understand this, consider the data given in Table 2.1.

Table 2-1: **Weekly family income X ($), and consumption Y ($)**

X Y	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y ($)	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	--	88	--	113	125	140	--	160	189	185
	--	--	--	115	--	--	--	162	--	191
Total	325	462	445	707	678	750	685	1043	966	1211
Mean	65	77	89	101	113	125	137	149	161	173

The data in the table refer to a total **population** of 60 families in a hypothetical community and their weekly income (*X*) and weekly consumption expenditure (*Y*), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 *fixed* values of *X* and the corresponding *Y* values against each of the *X* values.

There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly from Figure 2.1. But the general picture that one gets is that, despite the variability of weekly consumption expenditure within each income bracket, *on the average*, weekly consumption expenditure increases as income increases.

Fig 3.1.5 Conditional distribution of expenditure for various level of income (data of table 2.1)

To see this clearly, in Table 2.1 we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of \$80, the mean consumption expenditure is \$65, while corresponding to the income level of \$200, it is \$137. In all we have 10 mean values for the 10 subpopulations of *Y*. We call these mean values **conditional expected values,** as they depend on the given values of the (conditioning) variable *X*. Symbolically, we denote them as E(Y |X), which is read as the expected value of Y given the value of X.

It is important to distinguish these conditional expected values from the **unconditional expected value of** weekly consumption expenditure, *E*(*Y*). If we add the weekly consumption expenditures for all the 60 families in the *population* and divide this number by 60, we get the number \$121.20 (\$7272/60), which is the unconditional mean, or expected, value of weekly consumption expenditure, *E*(*Y*); it is unconditional in the sense that in arriving at this number we have disregarded the income levels of the various families. Obviously, the various conditional expected values of *Y* given in Table 2.1 are different from the unconditional expected value of *Y* of \$121.20. When we ask the question, "What is the *expected value* of weekly consumption expenditure of a family," we get the answer \$121.20 (the unconditional mean). But if we ask the question, "What is the *expected value* of weekly consumption expenditure of a family whose monthly income is, say, \$140," we get the answer \$101 (the conditional mean).

Geometrically, a population regression curve (line) is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s).

3.1.5.1 The Concept Of Population Regression Function (PRF)

From the preceding discussion, it is clear that each conditional mean $E(Y \mid X_i)$ is a function of X_i, where X_i is a given value of *X*. Symbolically,

$E(Y \mid X_i) = f\,(X_i)$ -- **(3.1.1)**

Where $f\,(X_i)$ denotes some function of the explanatory variable *X*. In the above example, $E(Y \mid X_i)$ is a linear function of X_i. Equation (2.2.1) is known as the **conditional expectation function (CEF)** or **population regression function (PRF)** or **population regression (PR)** for short. It states merely that the *expected value* of the distribution of *Y* given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of *Y* varies with *X*.

As a first approximation or a working hypothesis, we may assume that the PRF $E(Y \mid X_i)$ is a linear function of X_i, say, of the type

$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$ --- **(3.1.2)**

Where β_1 and β_2 are unknown but fixed parameters known as the **regression coefficients**

It is clear from Figure 2.1 that, as family income increases, family consumption expenditure on the average increases, too. But, what about the consumption expenditure of an individual family in relation to its (fixed) level of income? It is obvious from Table 2.1 and Figure 2.1 that an individual family's consumption expenditure does not necessarily increase as the income level increases. For example, from Table 2.1 we observe that corresponding to the income level of \$100 there is one family whose consumption expenditure of \$65 is less than the consumption expenditures of two families whose weekly income is only \$80. But notice that the *average* consumption expenditure of families with a weekly income of \$100 is greater than the average consumption expenditure of families with a weekly income of \$80 (\$77 versus \$65).

We see from Figure 2.1 that, given the income level of X_i, an individual family's consumption expenditure is clustered around the average consumption of all families at that X_i, that is, around its conditional expectation. Therefore, we can express the *deviation* of an individual Y_i around its expected value as follows:

$$u_i = Y_i - E(Y \mid X_i)$$

or

$$Y_i = E(Y \mid X_i) + u_i \text{ -- } \mathbf{(3.1.3)}$$

where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is known as the **stochastic disturbance** or **stochastic error term.**

How do we interpret (2.2.3)? We can say that the expenditure of an individual family, given its income level, can be expressed as the sum of two components: (1) $E(Y \mid X_i)$, which is simply the mean consumption expenditure of all the families with the same level of income. This component is known as the **systematic,** or **deterministic,** component, and (2) u_i, which is the random, or **nonsystematic,** component. Stochastic disturbance term is a *surrogate or proxy* for all the omitted or neglected variables that may affect Y but are not (or cannot be) included in the regression model.

If $E(Y \mid X_i)$ is assumed to be linear in X_i, as in Eq. (2.2.2), Eq. (2.2.3) may be written as

$$Y_i = E(Y \mid X_i) + u_i$$

$$= \beta_1 + \beta_2 X_i + u_i \text{ -- } \mathbf{(3.1.4)}$$

We call this equation stochastic specification of the PRF (true PRF)

3.1.6. The Sample Regression Function (SRF)

It is about time to face up to the sampling problems, for in most practical situations what we have is but a sample of Y values corresponding to some fixed X's. Therefore, the task now is to estimate the PRF on the basis of the sample information.

As an illustration, pretend that the population of Table 2.1 was not known to us and the only information we had was a randomly selected sample of Y values for the fixed X's as given in Table 2.2.

The question is: From the sample of Table 2.2 can we predict the average weekly consumption expenditure Y in the population as a whole corresponding to the chosen X's? In other words, can we estimate the PRF from the sample data? As one surely suspects, we may not be able to estimate the PRF "accurately" because of sampling fluctuations.

Plotting the data of Tables 2.2 a and 2.2 b, we obtain the scatter diagram given in Figure 2.2. In the scatter diagram two samples regression lines are drawn so as to “fit” the scatters reasonably well: SRF1 is based on the first sample, and SRF2 is based on the second sample. Which of the two regression lines represents the “true” population regression line? There is no way we can be absolutely sure that either of the regression lines shown in Figure 2.2 represents the true population regression line (or curve).

Table 2-2 a: A random sample from the population

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Table 2-2 b: Another random sample from the population

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

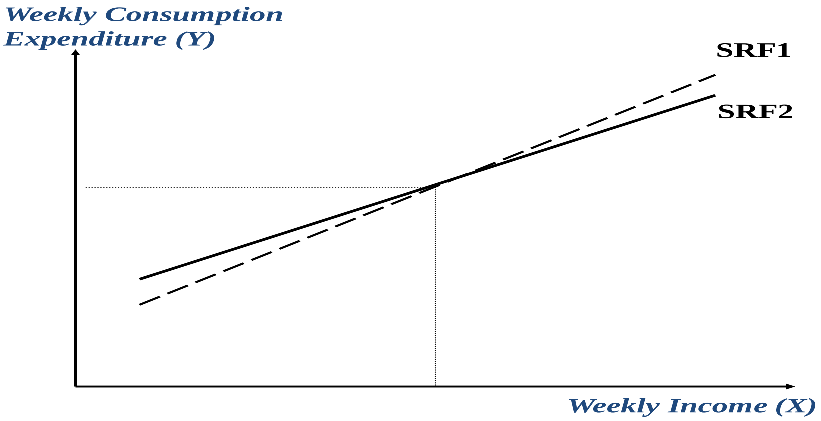


Fig 2.2

The regression lines in Figure 2.2 are known as the sample regression lines. They represent the population regression line, but because of sampling fluctuations they are at best an approximation of the true PR.

Now, analogously to the PRF that underlies the population regression line, we can develop the concept of the **sample regression function** (SRF) to represent the sample regression line. The sample counterpart of (2.2.2) may be written as

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ -- **(3.1.5)**

where $\hat{Y}_i$ is read as “*Y*-hat’’ or “*Y*-cap’’

$\hat{Y}_i$ = estimator of $E(Y \mid X_i)$

$\hat{\beta}_1$ = estimator of β_1 $\hat{\beta}_2$ = estimator of β_2

Note that an **estimator,** also known as a (sample) **statistic,** is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator in an application is known as an **estimate.**

Now just as we expressed the PRF in two equivalent forms, (2.2.2) and (2.2.4), we can express the SRF (2.2.5) in its stochastic form as follows:

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ -- **(3.1.6)**

Where, in addition to the symbols already defined, $\hat{u}_i$ denotes the (sample) **residual** term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an *estimate* of u_i.

To sum up, our primary objective in regression analysis is to estimate the PRF

$Y_i = \beta_1 + \beta_2 X_i + u_i$ -- **(3.1.4)**

On the basis of the SRF

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ -- **(3.1.6)**

because more often than not our analysis is based upon a single sample from some population:

☞ The deviations of the observations from the line may be attributed to several factors.

(1) *Omission of variables from the function*
In economic reality each variable is influenced by a very large number of factors. However, not all the factors influencing a certain variable can be included in the function for various reasons.

(2) *Random behavior of the human beings*
The scatter of points around the line may be attributed to an erratic element which is inherent in human behavior. Human reactions are to a certain extent unpredictable and may cause deviations from the normal behavioral pattern depicted by the line.

(3) *Imperfect specification of the mathematical form of the model*
We may have linearised a possibly nonlinear relationship. Or we may have left out of the model some equations.

(4) *Errors of aggregation*
We often use aggregate data (aggregate consumption, aggregate income), in which we add magnitudes referring to individuals whose behavior is dissimilar. In this case we say that variables expressing individual peculiarities are missing.

(5) *Errors of measurement*
This refers to errors of measurement of the variables, which are inevitable due to the methods of collecting and processing statistical information.

The first four sources of error render the form of the equation wrong, and they are usually referred to as error in the equation or error of omission. The fifth source of error is called error of measurement or error of observation.

In order to take in to account the above sources of error we introduce in econometric functions a random variable u called random disturbance term of the function, so called because u is supposed to disturb the exact linear relationship which is assumed to exist between X and Y.

3.1.7 The Meaning of the Term Linear

Linearity can be interpreted in two different ways.

Linearity in the Variables

The first and perhaps more "natural" meaning of linearity is that the conditional expectation of *Y* is a linear function of X_i, such as, for example, (2.2.2). Geometrically, the regression curve in this case is a straight line. In this interpretation, a regression function such as $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is not a linear function because the variable *X* appears with a power or index of 2.

Linearity in the Parameters

The second interpretation of linearity is that the conditional expectation of *Y*, $E(Y \mid X_i)$, is a linear function of the parameters, the β's; it may or may not be linear in the variable X. In this interpretation $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is a linear (in the parameter) regression model.

Of the two interpretations of linearity, linearity in the parameters is relevant for the development of the regression theory to be presented shortly. Therefore, from now on the term "linear" regression will always mean a regression that is linear in the parameters; the β's (that is, the parameters are raised to the first power only). It may or may not be linear in the explanatory variables, the X's.

3.1.8. The Ordinary Least Squares Methods (OLS)

To estimate the coefficients β_1 and β_2 we need observations on *X*, *Y* and *u*. yet *u* is never observed like the other explanatory variables, and therefore in order to estimate the function $Y_i = \beta_1 + \beta_2 X_i + u_i$, we should guess the values of *u*, that is we should make some reasonable assumptions about the shape of the distribution of each u_i (its means, variance and covariance with other *u*'s). These assumptions are guesses about the true, but unobservable, value of u_i.

THE ASSUMPTIONS UNDERLYING THE METHOD OF LEAST SQUARES

The linear regression model is based on certain assumptions, some of which refers to the distribution of the random variable *u*, some to the relationship between *u* and the explanatory variables, and some refers to the relationship between the explanatory variables themselves.

1. u_i is a random real variable and has zero mean value: $E(u_i) = 0$ (or $E(u_i|X_i) = 0$)
    - This implies that for each value of X, u may assume various values, some positive, and some negative but on average zero.
    - Further $E(Y_i) = \beta_1 + \beta_2 X_i$ gives the relationship between X and Y on the average, i.e. when X takes on value X_i , then Y will on the average take on $E(Y_i)$ (or $E(Y_i|X_i)$)

2. The variance of u_i is constant for all i, i.e., $var(u_i|X_i) = E(u_i^2|X_i) = \sigma^2$, and is called the assumptions of common variance or homoscedasticity.
    - The implication is that for all values of X, the values of u show the same dispersion around their mean.
    - The consequence of this assumption is that $var(y_i|X_i) = \sigma^2$
    - If on the other hand the variance of Y population varies as X changes, a situation of non-constancy of the variance of Y, called heteroscedasticity arises.
3. u_i has a normal distribution, i.e., $u_i \sim N(0, \sigma^2)$, which also implies $Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$.
4. The random terms of different observations are independent, $cov(u_i u_j) = E(u_i u_j) = 0$ for $i \neq j$ where i and j run from 1 to n. This is called the assumption of no autocorrelation (serial) among the error terms.
    - The consequence of this assumption is that $cov(Y_i Y_j) = 0$, for $i \neq j$ i.e. no autocorrelation among the Y's.
5. X_i's are a set of fixed values in the process of repeated sampling which underlies the linear regression model, i.e. they are non-stochastic.
6. u is independent of the explanatory variables, i.e., $cov(u_i X_i) = E(u_i X_i) = 0$.
7. Variability in X values. The X values in a given sample must not all be the same. Technically, var(X) must be a finite positive number.
8. The regression model is correctly specified.

3.2. THE LEAST SQAURE CRITERION AND NORMAL EQUATIONS OF OLS

Thus far we have completed the work involved in the first stage of any econometric application, namely we have specified the model and stated explicitly its assumptions. The next step is the estimation of the model, that is, the computation of the numerical values of its parameters.

The linear relationship $Y_i = \beta_1 + \beta_2 X_i + u_i$ holds for the population of the values of *X* and *Y*, so that we could obtain the numerical values of β_1 and β_2 only if we could have all the possible values of *X*, *Y* and *u* which form the population of these variables. Since this is impossible in practice, we get a sample of observed values of *Y* and *X*, specify the distribution of the u's and try to get satisfactory estimates of the true parameters of the relationship. This is done by fitting a regression line through the observations of the sample, which we consider as an approximation to the true line.

The method of ordinary least squares is one of the econometric methods which enable us to find the estimate of the true parameter and is attributed to Carl Friedrich Gauss, a German mathematician. To understand this method, we first explain the least squares principle.

Recall the two-variable PRF:

$Y_i = \beta_1 + \beta_2 X_i + u_i$ -- (3.1.4)

However, as noted in earlier, the PRF is not directly observable. We estimate it from the SRF:

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ -- (3.1.6)

$= \hat{Y}_i + \hat{u}_i$ -- (*)

Where $\hat{Y}_i$ is the estimated (conditional mean) value of Y_i

But how is the SRF itself determined? To see this, let us proceed as follows. First, express (*) as

$\hat{u}_i = Y_i - \hat{Y}_i$

$= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$ -- (3.2.1)

which shows that the $\hat{u}_i$ (the residuals) are simply the differences between the actual and estimated Y values.

Now given *n* pairs of observations on *Y* and *X*, we would determine the SRF in such a manner that it is as close as possible to the actual *Y*. To this end, we adopt the *least-squares criterion,* which states that the SRF can be fixed in such a way that

$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$

$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$ -- (3.2.2)

is as small as possible, where $\hat{u}_i^2$ are the squared residuals.

It is obvious from (3.2.2) that $\sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2)$ that is, the sum of the squared residuals is some function of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. For any given set of data, choosing different values for $\hat{\beta}_1$and $\hat{\beta}_2$will give different $\hat{u}$'s and hence different values of $\sum \hat{u}_i^2$.

The principle or the method of least squares chooses $\hat{\beta}_1$ and $\hat{\beta}_2$ in such a manner that, for a given sample or set of data, $\sum \hat{u}_i^2$ is as small as possible. In other words, for a given sample, the method of least squares provides us with unique estimates of β_1 and β_2 that give the smallest possible value of $\sum \hat{u}_i^2$.

The process of differentiation yields the following equations for estimating β_1 and β_2. Differentiating Eq. (3.2.2) partially with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$, we obtain

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_1} = -2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_2} = -2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\, X_i = 0$$

Setting these equations to zero gives, the normal equations below

$\sum Y_i = n\,\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$ -- (3.2.3)

$$\sum X_i\, Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \text{ --- (3.2.4)}$$

where *n* is the sample size. These simultaneous equations are known as the **normal equations.**

Solving the normal equations simultaneously, we obtain

$$\hat{\beta}_2 = \frac{n\sum Y_i\, X_i - \sum X_i \sum Y_i}{n\Sigma X_{\mathrm{i}}^2 - (\sum X_i)^2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\Sigma X_i^2 \sum Y_i\, - \sum X_i \sum X_i Y_i}{n\Sigma X_{\mathrm{i}}^2 - (\sum X_i)^2}$$

$$= \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2} \text{ ----- (3.2.5)} \qquad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2\overline{X} \text{ ---------------- (3.2.6)}$$

where $\overline{X}$ and $\overline{Y}$are the sample means of X and Y and where we define $x_i = (X_i - \overline{X})$ *and* $y_i = (Y_i - \overline{Y})$. The above lowercase letters in the formula denote deviations from mean values. Equation (3.2.6) can be obtained directly from (3.2.3) by simply dividing both sides of the equation by n.

Note that, by making use of simple algebraic identities, formula (3.2.5) for estimating β_2 can be alternatively expressed as

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \text{ --- (3.2.7)}$$

The estimators obtained previously are known as the **least-squares estimators,** for they are derived from the least-squares principle. We finally write the regression line equation as $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$.

☞ Interpretation of estimates

- Estimated intercept, $\hat{\beta}_1$: The estimated average value of the dependent variable when the independent variable takes on the value zero
- Estimated slope, $\hat{\beta}_2$: The estimated change in the average value of the dependent variable when the independent variable increases by one unit.
- $\hat{\mathrm{Y}}_{\mathrm{i}}$ gives average relationship between Y and X. i.e. $\hat{\mathrm{Y}}_{\mathrm{i}}$ is average value of Y given X_i.

Example 1

A random sample of ten families had the following income and food expenditure (in $ per week)

Families	A	B	C	D	E	F	G	H	I	J
Family income	20	30	33	40	15	13	26	38	35	43
Family expenditure	7	9	8	11	5	4	8	10	9	10

Estimate the regression line of food expenditure on income and interpret your results.

Note the following **numerical properties** of estimators obtained by the method of OLS.

1. The OLS estimators are expressed solely in terms of the observable (i.e., sample) quantities (i.e., *X* and *Y*). Therefore, they can be easily computed.
2. They are **point estimators.**
3. Once the OLS estimates are obtained from the sample data, the sample regression line can be easily obtained. The regression line thus obtained has the following properties:
   3.1. It passes through the sample means of *Y* and *X*.
   3.2. The mean value of estimated Y = $\hat{Y}_i$ is equal to the actual value of Y. i.e. $\bar{\hat{Y}} = \bar{Y}$
   3.3. The mean value of the residuals $\hat{u}_i$ is zero.
   3.4. As a result of the preceding property, the sample regression $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ can be expressed in an alternative form where both *Y* and *X* are expressed as deviations from their mean values. i.e. $y_i = \hat{\beta}_2 x_i + \hat{u}_i$. The SRF can also be written as $\hat{y}_i = \hat{\beta}_2 x_i$, whereas in the original units of measurement it was $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$. The above equations are called the deviation form.
   3.5. The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i.
   3.6. The residuals $\hat{u}_i$ are uncorrelated with X_i.

3.3 PRECISION OR STANDARD ERRORS OF LEAST-SQUARES ESTIMATES

It is evident that least-squares estimates are a function of the sample data. But since the data are likely to change from sample to sample, the estimates will change ipso facto. Therefore, what is needed is some measure of "reliability" or precision of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. In statistics the precision of an estimate is measured by its standard error (se). The standard errors of the OLS estimates can be obtained as follows:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \text{ -- } (3.3.1)$$

$$\text{se}(\hat{\beta}_2) = \sigma_{(\hat{\beta}_{2i})} = \frac{\sigma}{\sqrt{\sum x_i^2}} \text{ -- } (3.3.2)$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2 \, \sigma^2}{n \sum x_i^2} \text{ -- } (3.3.3)$$

Where var = variance and se = standard error and where σ^2 is the constant or homoscedastic variance of u_i of Assumption 2.

All the quantities entering into the preceding equations except σ^2 can be estimated from the data. σ^2 itself is estimated by the following formula:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

Where $\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2 and where the expression $n - 2$ is known as the **number of degrees of freedom (df),** $\sum \hat{u}_i^2$ being the sum of the residuals squared or the **residual sum of squares (RSS).**[1]

Once $\sum \hat{u}_i^2$ is known, $\hat{\sigma}^2$can be easily computed. $\sum \hat{u}_i^2$ itself can be computed either from (3.2.2) or from the following expression.

$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2$ -- (3.3.5)

Compared with Eq. (3.2.2), Eq. (3.3.5.) is easy to use, for it does not require computing $\hat{u}_i$ for each observation.

Since $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$, an alternative expression for computing $\sum \hat{u}_i^2$ is

$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$ --- (3.3.6)

[1] Note: The term number of degrees of freedom means the total number of observations in the sample (= n) less the number of independent (linear) constraints or restrictions put on them. In other words, it is the number of independent observations out of a total of n observations. **The general rule is this:** df = (n − number of parameters estimated).

Note that the positive square root of $\hat{\sigma}^2$

$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$ --- (3.3.7)

is known as the **standard error of estimate** or **the standard error of the regression (se).** It is simply the standard deviation of the *Y* values about the estimated regression line and is often used as a summary measure of the "goodness of fit" of the estimated regression line.

Note the following features of the variances (and therefore the standard errors) of $\hat{\beta}_1$ and $\hat{\beta}_2$.

1 The variance of $\hat{\beta}_2$ is directly proportional to σ^2 but inversely proportional to$\sum x_i^2$.
2 The variance of $\hat{\beta}_1$ is directly proportional to σ^2 and $\sum X_i^2$ but inversely proportional to$\sum x_i^2$ and the sample size *n*.

A NUMERICAL EXAMPLE

We illustrate the econometric theory developed so far by considering the Keynesian consumption function discussed in the Introduction. As a test of the Keynesian consumption function, we use the sample data of Table 2.2a, which for convenience is reproduced as Table 3.2.

Table 3.2: hypothetical data on weekly family consumption expenditure Y and weekly family income X

Y($)	X($)
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Table 3.3 raw data based on table 3.2

	Y_i (1)	X_i (2)	Y_iX_i (3)	X_i^2 (4)	$x_i = X_i - \bar{X}$ (5)	$y_i = Y_i - \bar{Y}$ (6)	x_i^2 (7)	x_iy_i (8)	$\hat{Y}_i$ (9)	$\hat{u}_i = Y_i - \hat{Y}_i$ (10)
	70	80	5600	6400	-90	-41	8100	3690	65.1818	4.8181
	65	100	6500	10000	-70	-46	4900	3220	75.3636	-10.3636
	90	120	10800	14400	-50	-21	2500	1050	85.5454	4.4545
	95	140	13300	19600	-30	-16	900	480	95.7272	-0.7272
	110	160	17600	25600	-10	-1	100	10	105.9090	4.0909
	115	180	20700	32400	10	4	100	40	116.0909	-1.0909
	120	200	24000	40000	30	9	900	270	125.2727	-6.2727
	140	220	30800	48400	50	29	2500	1450	136.4545	3.5454
	155	240	37200	57600	70	44	4900	3080	145.6363	8.3636
	150	260	39000	67600	90	39	8100	3510	156.8181	-6.8181
Sum	1110	1700	205500	322000	0	0	33000	16800	1109.9995 ≈ 1110.0	0
Mean	111	170	nc	nc	0	0	nc	nc	111	0

$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$ $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$

$= 111 - 0.5091(170)$

Notes: ≈ symbolizes "approximately equal to"; nc means "not computed."

The raw data required to obtain the estimates of the regression coefficients, their standard errors, etc., are given in Table 3.3. From these raw data, the following calculations are obtained.

$\hat{\beta}_1 = 24.4545$ var $(\hat{\beta}_1) = 41.1370$ and se $(\hat{\beta}_1) = 6.4138$

$\hat{\beta}_2 = 0.5091$ var $(\hat{\beta}_2) = 0.0013$ and se $(\hat{\beta}_2) = 0.0357$

cov $(\hat{\beta}_1, \hat{\beta}_2) = -0.2172$ $\hat{\sigma}^2 = 42.1591$

$r^2 = 0.9621$ $r = 0.9809$ df = 8

The estimated regression line therefore is

$$\hat{Y}_i = 24.4545 + 0.5091X_i$$

The associated regression line are interpreted as follows: Each point on the regression line gives an estimate of the expected or mean value of Y corresponding to the chosen X value; that is, $\hat{Y}_i$ is an estimate of $E(Y \mid X_i)$. The value of $\hat{\beta}_2 = 0.5091$, which measures the slope of the line, shows that, within the sample range of X between $80 and $260 per week, as X increases, say, by $1, the estimated increase in the mean or average weekly consumption expenditure amounts to about 51 cents. The value of $\hat{\beta}_1 =$ 24.4545, which is the intercept of the line, indicates the average level of weekly consumption expenditure when weekly income is zero.

3.3.1 PROPERTIES OF LEAST-SQUARES ESTIMATORS: THE GAUSS–MARKOV THEOREM

Given the assumptions of the classical linear regression model, the least-squares estimates possess some ideal or optimum properties. These properties are contained in the well-known Gauss–Markov theorem. An estimator, say the OLS estimators $\hat{\beta}_2$, is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:

1 It is linear, that is, a linear function of a random variable, such as the dependent variable Y in the regression model.
2 It is unbiased, that is, its average or expected value, $E(\hat{\beta}_2)$, is equal to the true value, β_2.
3 It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

In the regression context it can be proved that the OLS estimators $(\hat{\beta}_1, \hat{\beta}_2)$ are BLUE.

3.4. REGRESSION THROUGH THE ORIGIN

There are occasions when the two-variable PRF assumes the following form:

$Y_i = \beta_2 X_i + u_i$ --- (3.3.8)

In this model the intercept term is absent or zero, hence the name regression through the origin. How do we estimate models like (3.1.5)? To answer these questions, let us first write the SRF of (3.1.5), namely, $Y_i = \hat{\beta}_2 X_i + \hat{u}_i$ -- (3.3.9)

Now applying the OLS method to (3.3.9), we obtain the following formulas for $\hat{\beta}_2$ and its variance

$\hat{\beta}_2 = \frac{\sum Y_i X_i}{\Sigma X_i^2}$ and $\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma X_i^2}$ where σ^2 is estimated by $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-1}$

The differences between the two sets of formulas should be obvious: In the model with the intercept term absent, we use **raw** sums of squares and cross products but in the intercept-present model, we use adjusted (from mean) sums of squares and cross products. Second, the df for computing $\hat{\sigma}^2$ is $(n - 1)$ in the model without intercept and $(n - 2)$ in the model with intercept.

3.5: STATISTICAL TESTS OF SIGNIFICANCE OF THE LEAST SQUARE ESTIMATES

After estimation of the parameters, the next stage is to establish the criteria for judging the goodness of the parameter estimates. As indicated in chapter one, we divide the available criteria in to three groups: theoretical a prior criteria, statistical criteria and econometric criteria. The theoretical criteria (sign and size of the coefficients) are set by economic criteria and defined in the stage of the specification of the model. In this chapter, we develop the statistical criteria for the evaluation of the parameter estimates.

The two most commonly used tests in econometrics are the following:

1. The square of the correlation coefficient, r^2, which is used to judge the explanatory power of the linear regression of Y on X.
2. The standard error of the parameter estimated and is applied for judging the statistical reliability of the estimates of the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$. It provides a measure of the degree of confidence we attribute to the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$

3.5.1 THE COEFFICIENT OF DETERMINATION r^2: A MEASURE OF "GOODNESS OF FIT"

After the estimation of the parameters and the determination of the least squares regression line, we need to know how good is the fit of this line to the sample observations of Y and X, i.e. we need to measure the dispersion of the observations around the regression line. If all the observations were to lie on the regression line, we would obtain a "perfect" fit, but this is rarely the case. Generally, there will be some positive $\hat{u}_i$ and some negative $\hat{u}_i$. What is needed is that these residuals around the regression line are as small as possible.

The coefficient of determination r^2 (two-variable case) or R^2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data. *r^2 shows the percentage of the total variation of the dependent variable that can be explained by the independent variable X.*

To compute this r^2, we proceed as follows: Recall that

$Y_i = \hat{Y}_i + \hat{u}_i$

Or in the deviation form

$y_i = \hat{y}_i + \hat{u}_i$ -- (3.4.1)

Squaring (4.1.1) on both sides and summing over the sample, we obtain

$$\left.\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i\hat{u}_i \\ \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \end{aligned}\right\} \text{-- (3.4.2)}$$

$$\sum y_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2$$

Since $\sum \hat{y}_i\hat{u}_i$ = 0 (why?) and $\hat{y}_i = \hat{\beta}_2 x_i$

The various sums of squares appearing in (4.1.2) can be described as follows:

- $\sum y_i^2 = \sum(Y_i - \bar{Y})^2$ = total variation of the actual Y values about their sample mean, which may be called the total sum of squares (TSS).
- $\sum \hat{y}_i^2 = \sum(\hat{Y}_i - \bar{\hat{Y}}_i)^2 = \sum(\hat{Y}_i - \bar{Y}_i)^2 = \hat{\beta}_2^2 \sum x_i^2$ = variation of the estimated *Y* values about their mean which is called sum of squares due to regression [i.e., due to the explanatory variable(s)], or explained by regression, or simply the **explained sum of squares (ESS).**
- $\sum \hat{u}_i^2$= residual or **unexplained** variation of the *Y* values about the regression line, or simply the **residual sum of squares (RSS).**

Thus, (4.1.2) is

TSS = ESS + RSS -- (3.4.3)

and shows that the total variation in the observed *Y* values about their mean value can be partitioned into two parts, one attributable to the regression line and the other to random forces because not all actual *Y* observations lie on the fitted line. Geometrically,

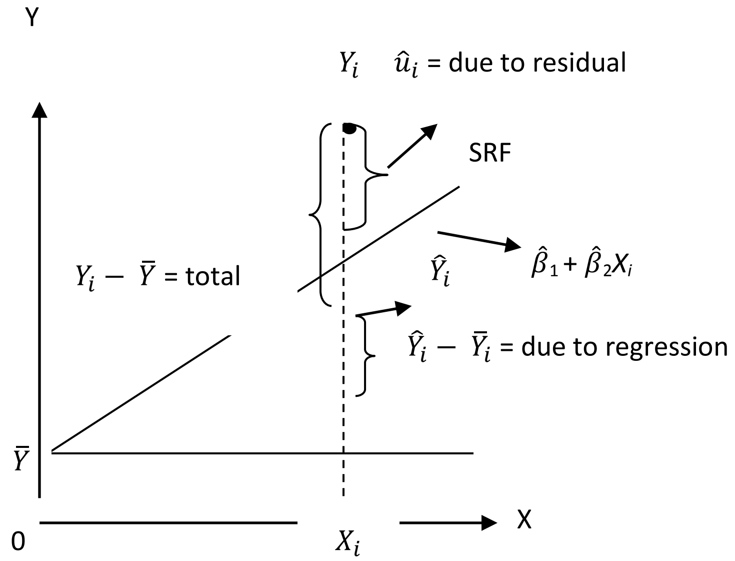

Y
Y_i
$\hat{u}_i$ = due to residual
SRF
$Y_i - \bar{Y}$ = total
$\hat{Y}_i$
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$\hat{Y}_i - \bar{Y}_i$ = due to regression
$\bar{Y}$
0
X_i
X

Fig. 4.1 Breakdown of the variation of Y_i into two components

Now dividing (3.4.3) by TSS on both sides, we obtain

$$\left.\begin{aligned} 1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \end{aligned}\right\} \text{-------------------------------------- (3.4.4)}$$

We now define r^2 as

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}} \text{ -- (3.4.5)}$$

or, alternatively, as

$$\begin{aligned} r^2 &= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \qquad \text{-- (3.4.6)} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \end{aligned}$$

r^2 can be computed more quickly from the following formula:

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2}\right) \text{ -- (3.4.7)}$$

If we divide the numerator and the denominator of (4.1.7) by the sample size *n* (or *n − 1* if the sample size is small), we obtain

$r^2 = \hat{\beta}_2^2 \left(\frac{S_X^2}{S_Y^2}\right)$ --- (3.4.8)

where S_X^2 and S_Y^2 are the sample variances of *Y* and *X*, respectively.

Since $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$ Eq. (4.1.7) can also be expressed as

$r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$ --- (3.4.9)

Two properties of r^2 may be noted:

1. It is a nonnegative quantity. (Why?)
2. Its limits are $0 \leq r^2 \leq 1$. An r^2 of 1 means a perfect fit, that is, $\hat{Y}_i = Y_i$ for each *i*. On the other hand, an r^2 of zero means that there is no relationship between the regressand and the regressor whatsoever (i.e., $\hat{\beta}_2 = 0$).

A quantity closely related to but conceptually very much different from r^2 is the **coefficient of correlation,** which is a measure of the degree of association between two variables. It can be computed either from

$r = \pm\sqrt{r^2}$ --- (3.4.10)

or from its definition

$$\left.\begin{aligned} r &= \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \\ r &= \frac{n\sum Y_i X_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}} \end{aligned}\right\} \text{-- (3.4.11)}$$

Some of the properties of *r* are as follows:

1. It can be positive or negative, the sign depending on the sign of the term in the numerator
2. It lies between the limits of −1 and +1; that is, $-1 \leq r \leq 1$.
3. It is symmetrical in nature; that is, the coefficient of correlation between *X* and *Y* (r_{XY}) is the same as that between *Y* and *X* (r_{YX}).
4. It is a measure of *linear association* or *linear dependence* only; it has no meaning for describing nonlinear relations like $Y = X^2$

<u>Example: 1</u> Find the value of r^2 for the numerical example given 0n table 3.2 and interpret it?

- ☞ $r^2 = 0.9621$ and $r = 0.9809$

The value of r^2 of 0.9621 means that about 96 percent of the variation in the weekly consumption expenditure is explained by income. The coefficient of correlation of 0.9809 shows that the two variables, consumption expenditure and income, are highly positively correlated.

Reporting the results of regression analysis

It has become customary to write all the results of regression analysis by writing out the estimated regression equation with all the values estimated (the standard errors, r^2, …) with the standard errors of the regression parameters put in parenthesis just under the respective values as shown below.

$$\hat{Y}_i = \underset{(se(\hat{\beta}_1))}{\hat{\beta}_1} + \underset{(se(\hat{\beta}_2))}{\hat{\beta}_2 X_i}, \ r^2 = \ldots, \hat{\sigma}^2 = \ldots, n = \ldots$$

3.5.2. TEST OF SIGNIFICANCE OF THE PARAMETER ESTIMATES

The **classical theory of statistical inference** consists of two branches, namely, **estimation** and **hypothesis testing.** We have thus far covered the topic of estimation of the parameters of the (two variable) linear regression model. Note that, since these are estimators, their values will change from sample to sample. Therefore, these estimators are random variables. Since $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$ are random variables, we need to find out their probability distributions, for without that knowledge we will not be able to relate them to their true values.

To find out the probability distributions of the OLS estimators, we proceed as follows. Specifically, consider $\hat{\beta}_2$,. It can be shown as

$\hat{\beta}_2 = \sum k_i Y_i$ -- (3.5.1)

where $k_i = \frac{x_i}{\sum x_i^2}$. But since the X's are assumed fixed, or nonstochastic, Eq. (3.5.1) shows that $\hat{\beta}_2$ is a linear function of Y_i, which is random by assumption. But since $Y_i = \beta_1 + \beta_2 X_i + u_i$, we can write (3.5.1) as

$\hat{\beta}_2 = \sum k_i(\beta_1 + \beta_2 X_i + u_i)$ --- (3.5.2)

Because k_i, the betas, and X_i are all fixed, $\hat{\beta}_2$ is ultimately a *linear* function of the random variable u_i, which is random by assumption. Therefore, the probability distribution of $\hat{\beta}_2$ (and also of $\hat{\beta}_1$) will depend on the assumption made about the probability distribution of u_i.

With the assumption that u_i follow the normal distribution, the OLS estimators have the following properties;

- $\hat{\beta}_1$ (being a linear function of u_i) is *normally distributed* with

Mean: $E(\hat{\beta}_1) = \beta_1$ -- (3.5.3)
var($\hat{\beta}_1$): $\sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$ -- (3.5.4)
Or more compactly,
$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$

- Then by the properties of the normal distribution the variable Z, which is defined as
  Z = $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ --- (3.5.5)
  follows the standard normal distribution, that is, a normal distribution with zero mean and unit (= 1) variance, or Z ∼ N(0, 1)
- $\hat{\beta}_2$ (being a linear function of u_i) is *normally* distributed with
  Mean: $E(\hat{\beta}_2) = \beta_2$ -- (3.5.6)
  var($\hat{\beta}_2$): $\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$ -- (3.5.7)

  Or, more compactly,
  $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$
- Then Z = $\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}}$ --- (3.5.8)
  also follows the standard normal distribution.
- $(n - 2)(\hat{\sigma}^2/\sigma^2)$ is distributed as the χ^2 (chi-square) distribution with (n − 2)df.
  - ☞ The important point to note is that the normality assumption enables us to derive the probability, or sampling, distributions of $\hat{\beta}_1$and $\hat{\beta}_2$ (both normal) and $\hat{\sigma}^2$ (related to the chi square). This simplifies the task of establishing confidence intervals and testing (statistical) hypotheses.

3.5.3 CONFIDENCE INTERVALS FOR REGRESSION ESTIMATES

The theory of estimation consists of two parts: point estimation and interval estimation. We have discussed point estimation thoroughly previously where we introduced the OLS method of point estimation. In this section we first consider interval estimation and then take up the topic of hypothesis testing, a topic intimately related to interval estimation.

In order to define how close to the estimate the true parameter lies, we must construct confidence intervals for the true parameter, in other words we must establish limiting values around the estimates within which the true parameter is expected to lie with a certain degree of confidence. In this respect we say that with a given probability the population parameter will be within the defined confidence interval or confidence limits.

How are the confidence intervals constructed? If the **sampling or probability distributions** of the estimators are known, one can make confidence interval statements. We choose a probability in advance and refer to it as the confidence level (confidence coefficient). It is customary in econometrics to choose the 95 percent confidence level. This means that in repeated sampling the confidence limits, computed from the sample, would include the true population parameter in 95 percent of the cases. In other 5 percent of the cases the population parameter will fall outside the confidence limits.

Confidence interval from the standard normal distribution

Z distribution will be employed either if we know the true standard deviation $\sigma_{(\hat{\beta}_{\mathrm{i}})}$, or when we have a large sample (n > 30), because, for large samples, the sample standard deviation, se, is a reasonably good estimate of the unknown population standard deviation.

The Z statistic for $\hat{\beta}_{\mathrm{i}}$ is

$$Z = \frac{\hat{\beta}_{\mathrm{i}} - \beta_{\mathrm{i}}}{\sigma_{(\hat{\beta}_{\mathrm{i}})}}$$

Our first task is to choose a confidence coefficient designated by α. We next look at the standard normal table and find out the probability of the value of Z lying between $-\mathrm{Z}_{\frac{\alpha}{2}}$ and $\mathrm{Z}_{\frac{\alpha}{2}}$ is 1- α. This may be written as follows

$$\mathrm{P}\left\{-\mathrm{Z}_{\frac{\alpha}{2}} < Z < \mathrm{Z}_{\frac{\alpha}{2}}\right\} = 1\text{- }\alpha$$

Substituting Z = $\left(\hat{\beta}_{\mathrm{i}} - \beta_{\mathrm{i}}\right)/\sigma_{(\hat{\beta}_{\mathrm{i}})}$ and rearranging slightly, we get

$$\mathrm{P}\left\{-\mathrm{Z}_{\frac{\alpha}{2}} < \frac{\hat{\beta}_{\mathrm{i}} - \beta_{\mathrm{i}}}{\sigma_{(\hat{\beta}_{\mathrm{i}})}} < \mathrm{Z}_{\frac{\alpha}{2}}\right\} = 1\text{- }\alpha$$

$$\mathrm{P}\left\{\hat{\beta}_{\mathrm{i}} - \mathrm{Z}_{\frac{\alpha}{2}}\sigma_{(\hat{\beta}_{\mathrm{i}})} < \beta_{\mathrm{i}} < \hat{\beta}_{\mathrm{i}} + \mathrm{Z}_{\frac{\alpha}{2}}\sigma_{(\hat{\beta}_{\mathrm{i}})}\right\} = 1\text{- }\alpha$$

Thus the (1- α)100 percent confidence interval for β_{i} is

$$\hat{\beta}_{\mathrm{i}} - \mathrm{Z}_{\frac{\alpha}{2}}\sigma_{(\beta_{\mathrm{i}})} < \beta_{\mathrm{i}} < \hat{\beta}_{\mathrm{i}} + \mathrm{Z}_{\frac{\alpha}{2}}\sigma_{(\beta_{\mathrm{i}})}$$

or

$$\beta_{\mathrm{i}} = \hat{\beta}_{\mathrm{i}} \pm \mathrm{Z}_{\frac{\alpha}{2}}\sigma_{(\hat{\beta}_{\mathrm{i}})}$$

The meaning of the confidence interval is that the unknown population parameter, β_{i}, will lie within the defined limits (1- α)100 times out of 100.

<u>Example: 2</u> If $\hat{\beta}_2$ = 8.4 and $\sigma_{(\hat{\beta}_2)}$ = 2.2, choosing a value of 95 percent for the confidence coefficient, construct confidence interval for β_2 and interpret it?

Confidence interval from the student's t distribution

The student's t distribution is applicable when

- The population variance is unknown, and the sample with which we work is small (n < 30) provided that the population of the parameters is normal

In econometric applications the true variances of the estimates, $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_2}$, are unknown, because they involve the true variance of the random term, σ^2, which is unknown. We may, however, use the unbiased estimate $\hat{\sigma}^2 = \sum u_i^2/n - K$ (K is number of parameters estimated) and obtain estimates of the variances of the coefficients, $se^2_{\hat{\beta}_1}$ and $se^2_{\hat{\beta}_2}$.

The *t* distribution is always symmetric, with mean equal to zero and variance $n - 1/n - 3$, which approaches unit when *n* is large. Clearly as *n* increases, the t distribution approaches the Standard Normal distribution Z ~ N(0,1).

The procedure for constructing a confidence interval with the t distribution is similar to the one outlined earlier with the main difference that in this case we must take in to account the degrees of freedom.

The *t* statistic for $\hat{\beta}_i$ is

$$t = \frac{\hat{\beta}_i - \beta_i}{se_{(\hat{\beta}_i)}}.$$

We first choose a confidence coefficient α. We next look at the *t* table and find out the probability of the value of *t* lying between $-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$ with n – K degrees of freedom is 1- α. This may be written as follows

$$P\left\{-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}\right\} = 1 - \alpha$$

Substituting t = $(\hat{\beta}_i - \beta_i)/se_{(\hat{\beta}_i)}$ and rearranging slightly, we get

$$P\left\{-t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_i - \beta_i}{se_{(\hat{\beta}_i)}} < t_{\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$P\left\{\hat{\beta}_i - t_{\frac{\alpha}{2}}se_{(\hat{\beta}_i)} < \beta_i < \hat{\beta}_i + t_{\frac{\alpha}{2}}se_{(\hat{\beta}_i)}\right\} = 1 - \alpha$$

Thus the (1- α)100 percent confidence interval for β_i is

$\hat{\beta}_i - t_{\frac{\alpha}{2}} se_{(\hat{\beta}_i)} < \beta_i < \hat{\beta}_i + t_{\frac{\alpha}{2}} se_{(\hat{\beta}_i)}$ with n – K degrees of freedom

or

$\beta_i = \hat{\beta}_i \pm t_{\frac{\alpha}{2}} se_{(\hat{\beta}_i)}$ with n – K degrees of freedom

The meaning of the confidence interval is that the unknown population parameter, β_i, will lie within the defined limits $\hat{\beta}_i \pm t_{\frac{\alpha}{2}} se_{(\hat{\beta}_i)}$ with n – K degrees of freedom (1- α)100 times out of 100.

<u>Example: 3</u> Suppose we have estimated the following regression line from a sample of 20 observations.

$$\hat{Y} = \underset{(38.2)}{128.5} + \underset{(0.85)}{2.88X}$$

Construct confidence interval for $\hat{\beta}_1$ and $\hat{\beta}_2$ and give interpretation for each?

CONFIDENCE INTERVAL FOR σ^2

As pointed out in Section 3.5.2 under the normality assumption, the variable

$$\chi^2 = (n\text{-}2)\frac{\hat{\sigma}^2}{\sigma^2}$$

follows the χ^2 distribution with $n - 2$ df. Therefore, we can use the χ^2 distribution to establish a confidence interval for σ^2

$$P\{\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\} = 1\text{-}\alpha$$

where the χ^2 value in the middle of this double inequality is as given by the above equation and where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are two values of χ^2 (the **critical** χ^2 values) obtained from the chi-square table for $n - 2$ df.

Substituting $\chi^2 = (n\text{-}2)\frac{\hat{\sigma}^2}{\sigma^2}$ and rearranging the terms, we obtain

$$P\left\{(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} < \sigma^2 < (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right\} = 1\text{-}\alpha$$

which gives the 100(1 − α)% confidence interval for σ^2.

Example: 4 If $\hat{\sigma}^2$ = 42.1591 and df = 8, then construct confidence interval for σ^2 taking α =5% and interpret it?

3.7. HYPOTHESIS TESTING

The concept of statistical hypothesis testing may be stated simply as follows: *Is a given observation or finding compatible with some stated hypothesis or not?* The word "compatible," as used here, means "sufficiently" close to the hypothesized value so that we do not reject the stated hypothesis.

There are two *mutually complementary* approaches for devising such rules, namely, **confidence interval** and **test of significance.** In the confidence-interval procedure we try to establish a range or an interval that has a certain probability of including the true but unknown β_i, whereas in the test-of-significance approach we hypothesize some value for β_i and try to see whether the computed $\hat{\beta}_i$ lies within reasonable (confidence) limits around the hypothesized value.

In general, the following are the steps involved in testing a statistical hypothesis:

Step 1. State the null hypothesis H_0 and the alternative hypothesis H_1

(e.g., H_0: $\mu = 69$ and H_1: $\mu \neq 69$).

Step 2. Select the test statistic (e.g., $\bar{X}$).

Step 3. Determine the probability distribution of the test statistic (e.g., $\bar{X} \sim N(\mu, \sigma^2/n)$.

Step 4. Choose the level of significance (i.e., the probability of committing a type I error) α.[2]

Step 5. Using the probability distribution of the test statistic, establish a 100(1 − α)% confidence interval. If the value of the parameter under the null hypothesis (e.g., $\mu = \mu^* = 69$) lies in this confidence region, the region of acceptance, do not reject the null hypothesis. But if it falls outside this interval (i.e., it falls into the region of rejection), you may reject the null hypothesis. Keep in mind that in not rejecting or rejecting a null hypothesis you are taking a chance of being wrong α percent of the time.

In practice, there is no need to estimate the confidence interval explicitly. One can compute the *test statistic* and see whether it lies within the acceptance or rejection (critical) region. We can summarize the *t* test of significance approach to hypothesis testing as shown in Table 4.1.

- Table 4-1: Decision Rule for t-test of significance

Type of Hypothesis	H_0	H_1	Reject H_0 if
Two-tail	$\beta_2 = \beta_2{*}$	$\beta_2 \# \beta_2{*}$	$\lvert t\rvert > t_{\alpha/2,df}$
Right-tail	$\beta_2 \leq \beta_2{*}$	$\beta_2 > \beta_2{*}$	$t > t_{\alpha,df}$
Left-tail	$\beta_2 \geq \beta_2{*}$	$\beta_2 < \beta_2{*}$	$t < - t_{\alpha,df}$

Notes: β_2^* is the hypothesized numerical value of β_2.

$|t|$ means the absolute value of t.

t_α or $t_{\alpha/2}$ means the critical t value at the α or $\alpha/2$ level of significance.

df: degrees of freedom, $(n - 2)$ for the two-variable model, $(n - 3)$ for the three variable model, and so on.

The same procedure holds to test hypotheses about β_1 and to undertake Z-test.

Example: 5 Suppose that from a sample of size n = 20, we estimate the following consumption function

$$\hat{C} = \underset{(75.5)}{100} + \underset{(0.21)}{0.70Y}$$

Test the hypothesis H_0: $\beta_2 = 0$.

Example: 6 Suppose $\hat{\beta}_2 = 29.48$ and $\sigma_{(\hat{\beta}_2)} = 36.0$. Test the hypothesis H_0: $\beta_2 = 25.0$.

The χ^2 test of significance approach to hypothesis testing is summarized in Table 4.2.

Table 4.2: A summary of the χ^2 test

Types of hypothesis	H_1: the alternative hypothesis	Critical region: reject H_o if
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{df(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha,df}$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{df(\hat{\sigma}^2)}{\sigma_0^2} < \chi^2_{1-\alpha,df}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{df(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha/2,df}$ or $< \chi^2_{(1-\alpha/2),df}$

Note: σ_0^2 is the value of σ^2 under the null hypothesis. The first subscript on χ^2 in the last column is the level of significance, and the second subscript is the degrees of freedom. These are critical chi-square values.

In statistics, when we reject the null hypothesis, we say that our finding is **statistically significant.** On the other hand, when we do not reject the null hypothesis, we say that our finding is **not statistically significant.**

We use two-sided hypothesis test when we do not have a strong a priori or theoretical expectation about the direction in which the alternative hypothesis should move from the null hypothesis. Sometimes we have a strong a priori or theoretical expectation (or expectations based on some previous empirical work) that the alternative hypothesis is one-sided or unidirectional rather than two-sided, and we use in this case one sided hypothesis test.

The "Zero" Null Hypothesis and the "2-*t*" Rule of Thumb

A null hypothesis that is commonly tested in empirical work is H_0: $\beta_2 = 0$, that is, the slope coefficient is zero. The objective of "zero" null hypothesis is to find out whether Y is related at all to X, the explanatory variable. If there is no relationship between Y and X to begin with, then testing a hypothesis such as $\beta_2 = 0.3$ or any other value is meaningless. This null hypothesis can be easily tested by the confidence interval or the *t*-test approach discussed in the preceding sections. But very often such formal testing can be shortcut by adopting the "2-*t*" rule of significance, which may be stated as

"2-*t*" Rule of Thumb. If the number of degrees of freedom is 20 or more and if α, the level of significance, is set at 0.05, then the null hypothesis $\beta_i = 0$ can be rejected if the t value

$[= \frac{\hat{\beta}_i}{se_{(\hat{\beta}_i)}})]$ computed from $t = \frac{\hat{\beta}_i - \beta_i}{se_{(\hat{\beta}_i)}}$ exceeds 2 in absolute value.

This statement assumes a two-tail test conducted at 5 percent level of significance.

__**Example: 7**__ Suppose that we have estimated the following supply function from a sample of 700 observations (n = 700)

$$Y = \underset{(20)}{100} + \underset{(1.5)}{4.00X}$$

Conduct the Z-test for the hypothesis H_0: $\beta_2 = 0$?

4. MULTIPLE REGRESSIONS

The two-variable model studied extensively in the previous chapter is often inadequate in practice. In our consumption–income example, for instance, it was assumed implicitly that only income X affects consumption Y. But economic theory is seldom so simple for, besides income, a number of other variables are also likely to affect consumption expenditure. An obvious example is wealth of the consumer. Therefore, we need to extend our simple two-variable regression model to cover models involving more than two variables. Adding more variables leads us to the discussion of multiple regression models, that is, models in which the dependent variable, or regressand, Y depends on two or more explanatory variables, or regressors. The simplest possible multiple regression model is three-variable regression, with one dependent variable and two explanatory variables.

In this chapter we shall extend the simple linear regression model to relationships with two explanatory variables and consequently to relationships with any number of explanatory variables.

4.1 MODELS WITH TWO EXPLANATORY VARIABLES

4.1.1 The normal equations

The population regression model with two explanatory variables is given as

$$Y_i = \underbrace{\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}}_{\text{Systematic component}} + \underbrace{u_i}_{\text{Random component}}, \quad \text{i = 1, 2, …, n} \text{ ------------------------------- } (4.1.1)$$

- β_1 is the intercept term which gives the average values of Y when $X_2$and $X_3$are zero.
- β_2 and β_3 are called the partial slope coefficient, or partial regression coefficients.
- β_2 measures the change in the mean value of Y resulting from a unit change in the X_2 given X_3(i.e. holding the value of X_3constant). Or equivalently $\beta_2$measures the direct or net effect of a unit change in X_2on the mean value of Y(net of any effect that X_3may have on the mean of Y). The interpretation of β_3 is also similar.

To complete the specification of our simple model we need some assumptions about the random variable u. These assumptions are the same as in the single explanatory variable model developed in chapter 3. That is:

- Zero mean value of u_i, or E($u_i|X_{2i}, X_{3i}$) = 0 for each i
- No serial correlation, or cov(u_i, u_j) = 0 where i $\neq$ j
- Homoscedasticity, or var(u_i) = σ^2
- Normality of u_i i.e $u_i \sim$ N(0, σ^2)

- Zero covariance between u_i and each X variable, or cov(u_i, X_{2i}) = cov(u_i, X_{3i}) = 0
- No specification bias, or the model is correctly specified
- No exact collinearity between the X variables, or no **exact linear relationship** between $X_2$and X_3

For notational symmetry, Eq. (4.1.1) can also be written as $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ with the provision that X_{1i} = 1 for all i.

The assumption of no collinearity is a new one and means the absence of possibility of one of the explanatory variables being expressed as a linear combination of the other. Existence of exact linear dependence between $X_{2i}$and X_{3i}would mean that we have only one independent variable in our model than two. If such a regression is estimated there is no way to estimate the separate influence of $X_2(\beta_2)$ and $X_3(\beta_3)$ on Y, since such a regression gives us only the combined influence of $X_2$and X_3on Y.

To see this suppose $X_3 = 2X_2$ then

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i$$

$$Y_i = \beta_1 + (\beta_2 + 2\beta_3) X_{2i} + u_i$$

$$Y_i = \beta_1 + \alpha X_{2i} + u_i \text{ , where } \alpha = (\beta_2 + 2\beta_3)$$

Estimating the above regression yields the combined effect of X_2 and X_3 as represented by $\alpha = (\beta_2 + 2\beta_3)$ where there is no possibility of separating their individual effects which are represented by β_2 and β_3.

This assumption does not guarantee there will not be correlations among the explanatory variables; it only means that the correlations are not exact or perfect, as it is not impossible to find two or more (economic) variables that may not be correlated to some extent. Likewise the assumption does not guarantee absence of non-linear relationships among X's either.

Having specified our model we next use sample observations on Y, X_2 and X_3 and obtain estimates of the true parameters β_1, β_2 and β_3:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

where $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ are estimates of the true parameters β_1, β_2 and β_3 of the relationship.

As before, the estimates will be obtained by minimizing the sum of squared residuals

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}))^2$$

A necessary condition for this expression to assume a minimum value is that its partial derivatives with respect to $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ be equal to zero:

$$\frac{\partial \sum(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}))^2}{\partial \hat{\beta}_1}$$

$$\frac{\partial \sum(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}))^2}{\partial \hat{\beta}_2}$$

$$\frac{\partial \sum(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}))^2}{\partial \hat{\beta}_3}$$

Performing the partial differentiations we get the following system of three normal equations in three unknown parameters $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$

$$\left.\begin{aligned} \sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} \\ \sum X_{2i} Y_i &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} \\ \sum X_{3i} Y_i &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \end{aligned}\right\} \text{----------- (4.1.2)}$$

From the solution of this system (by any method, for example using determinants) we obtain values for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$.

Also by solving the system of normal equations

$$\left.\begin{aligned} \sum x_{2i} y_i &= \hat{\beta}_2 \sum x_{2i}^2 + \hat{\beta}_3 \sum x_{2i} x_{3i} \\ \sum x_{3i} y_i &= \hat{\beta}_2 \sum x_{2i} x_{3i} + \hat{\beta}_3 \sum x_{3i}^2 \end{aligned}\right\} \text{--- (4.1.3)}$$

The following formulae, in which the variables are expressed in deviations from their mean, may be obtained for estimating the values of the parameter estimates.

$$\left.\begin{aligned} \hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \\ \hat{\beta}_2 &= \frac{(\sum x_{2i} y_i)(\sum x_{3i}^2) - (\sum x_{3i} y_i)(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \\ \hat{\beta}_3 &= \frac{(\sum x_{3i} y_i)(\sum x_{2i}^2) - (\sum x_{2i} y_i)(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \end{aligned}\right\} \text{------------------------------------ (4.1.4)}$$

where $y_i = Y_i - \bar{Y}$, $x_{2i} = X_{2i} - \bar{X}_2$ and $x_{3i} = X_{3i} - \bar{X}_3$

4.1.2 The coefficient of multiple determination (or the squared multiple correlation coefficient) R^2

In the two-variable case we saw that r^2 measures the goodness of fit of the regression equation; that is, it gives the proportion or percentage of the total variation in the dependent variable Y explained by the (single) explanatory variable X. This notation of r^2 can be easily extended to regression models containing more than two variables. Thus, in the three variable model we would like to know the proportion of the variation in Y explained by the variables X2 and X3 jointly. The quantity that gives this information is known as the **multiple coefficient of determination** and is denoted by R^2; conceptually it is akin to r^2.

$$R^2 = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum y^2}$$

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} - - - - - - - - - - - -(4.1.5)$$

The value of R^2 lies between 0 and 1. The higher R^2 the greater the percentage of the variation of Y explained by the regression plane, that is, the better the 'goodness of fit' of the regression plane to the sample observations. The closer R^2 to zero, the worse the fit.

4.1.3 The mean and variance of the parameter estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$

The mean of the estimates of the parameters in the three-variable model is derived in the same way as in the two-variable model. The estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are unbiased estimates of the true parameters of the relationship between Y, X2 and X3: their mean expected value is the true parameter itself.

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_2) = \beta_2 \qquad E(\hat{\beta}_3) = \beta_3$$

The variance of the parameter estimates are obtained by the following formulae

$$var(\hat{\beta}_1) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2i}x_{3i}}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\right]$$

$$var(\hat{\beta}_2) = \hat{\sigma}^2 \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$$

$$var(\hat{\beta}_3) = \hat{\sigma}^2 \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$$

where $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-K)$, K being the total number of parameters which are estimated. In the three-variable model K = 3.

4.2 THE GENERAL LINEAR REGRESSION MODEL

In this section we will extend the method of least squares to models including any number k of explanatory variables. There are some rule of thumb by which we can derive (a) the normal equations, (b) the coefficients of multiple determination, (c) the variances of the coefficients, for relationships including any number of explanatory variables.

4.2.1 Derivations of the normal equations

The general linear regression model with k explanatory variables is of the form

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i$$

There are K parameters to be estimated (K = k+1). Clearly the system of normal equations will consist of K equations, in which the unknowns are the parameters $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, …, $\hat{\beta}_K$, and the known terms will be the sums of squares and the sums of products of all the variables in the structural equation.

In order to derive the K normal equations without the formal differentiation procedure, we start from the equation of the estimated relationship

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} + \hat{u}_i$$

and we make use of the assumptions

$\sum \hat{u}_i = 0$ and $\sum u_i X_j = 0$ where (j = 1, 2, 3, …, K)

The normal equations for a model with any number of explanatory variables may be derived in a mechanical way, without recourse to differentiation. We will introduce a practical rule of thumb, derived by inspection of the normal equations of the two-variable and the three-variable models. We begin by rewriting these normal equations.

1. Model with one explanatory variables
   Structural form $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$
   Estimated form $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{u}_i$

$$\text{Normal equations} \begin{cases} \sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} \\ \sum X_{2i} Y_i = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \end{cases}$$

2. Models with two explanatory variables

Structural form $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

Estimated form $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$

Normal equations
$$\begin{cases} \sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} \\ \sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} \end{cases}$$

Comparing the normal equations of the above models, we can generalize the procedure to find the Kth equation of the normal equations for the K-variable model which may be obtained by multiplying the estimated form of the K-variable model by X_{Ki} and then summing over all sample observations. The estimated form of the model is

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_K X_{Ki} + \hat{u}_i$$

Multiplication through by X_{Ki} yields

$$Y_i X_{Ki} = \hat{\beta}_1 X_{Ki} + \hat{\beta}_2 X_{2i} X_{Ki} + \hat{\beta}_3 X_{3i} X_{Ki} + \cdots + \hat{\beta}_K X_{Ki}^2 + \hat{u}_i X_{Ki}$$

and summation over the n sample observation gives the required Kth equation

$$\sum Y_i X_{Ki} = \ \hat{\beta}_1 \sum X_{Ki} + \hat{\beta}_{2i} \sum X_{2i} X_{Ki} + \hat{\beta}_3 \sum X_{3i} X_{Ki} + \cdots + \hat{\beta}_K \sum X_{Ki}^2$$

given that by assumption $\sum \hat{u}_i X_{Ki} = 0$

The generalization of the linear regression model with the variables expressed in deviations from their means is the same. Thus the estimated form of the K-variable model in deviation form is

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_K x_{Ki} + \hat{u}_i$$

The Kth equation is derived by multiplying through the estimated form by x_{Ki} and summing over all the sample observations

$$\sum y_i x_{Ki} = \hat{\beta}_2 \sum x_{2i} x_{Ki} + \hat{\beta}_3 \sum x_{3i} x_{Ki} + \cdots + \hat{\beta}_K \sum x_{Ki}^2$$

4.2.2 Generalization of the formula for R^2

The generalization of the formula of the coefficient of multiple determination may be derived by inspection of the formulae of R^2 for the two-variable and three-variable models.

1. Model with one explanatory variable

$$R_{Y.X_2}^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i}}{\sum y_i^2}$$

2. Model with two explanatory variables

$$R^2_{Y.X_2X_3} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$$

By inspection we see that for each additional explanatory variable the formula of the squared multiple correlation coefficients includes an additional term in the numerator, formed by the estimate of the parameter corresponding to the new variable multiplied by the sum of products of the deviations of the new variable and the dependent one. For example, the formula of the coefficient of multiple determinations for the K-variable model is

$$R^2_{Y.X_2\dots X_K} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \cdots + \hat{\beta}_K \sum y_i x_K}{\sum y_i^2}$$

4.2.3 The adjusted coefficient of determination: $\bar{R}^2$

The inclusion of additional explanatory variables in the function can never reduce the coefficients of multiple determination and will usually raise it. By introducing a new regressor we increase the value of the numerator of the expression for R^2, while the denominator remain the same ($\sum y_i^2$ the total variations of Yi is given in any particular sample).

To correct for this defect we adjust R^2 by taking into account the degrees of freedom, which clearly decrease as new regressors are introduced in the function. The expression for the adjusted coefficient of multiple determination is:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-K}$$

or

$$\bar{R}^2 = 1 - \left[\frac{\sum u_i^2/(n-K)}{\sum y_i^2/(n-1)}\right]$$

where R^2 is the unadjusted multiple correlation coefficient, n is the number of sample observations and K is the number of parameters estimated from the sample. If n is large $\bar{R}^2$ and R^2 will not differ much. But with small samples, if the number of regressors (X's) is large in relation to the sample observations, $\bar{R}^2$ will be much smaller than R^2 and can even assume negative values, in which case $\bar{R}^2$ should be interpreted as being equal to zero.

4.2.4 Generalization of the formulae of the variances of the parameter estimates

The generalization of the formulae of the variances of the parameter estimates is facilitated by the use of determinants. In the preceding sections we have developed the formulae of the variances of the estimates for models with one and two explanatory variables.

1. Model with one explanatory variable

$$var(\hat{\beta}_2) = \sigma^2 \frac{1}{\sum x_{2i}^2}$$

2. Model with two explanatory variables

$$var(\hat{\beta}_2) = \sigma^2 \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$$

$$var(\hat{\beta}_3) = \sigma^2 \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$$

The above expressions may be written in the form of determinants as follows. The normal equations of the model with two explanatory variables, written in deviation form, are

$$(\sum x_{2i}y_i) = \hat{\beta}_2(\sum x_{2i}^2) + \hat{\beta}_3(\sum x_{2i}x_{3i})$$

$$(\sum x_{3i}y_i) = \hat{\beta}_2(\sum x_{2i}x_{3i}) + \hat{\beta}_3(\sum x_{3i}^2)$$

The terms in the parentheses are the 'knowns' which are computed from the sample observations, while $\hat{\beta}_2$ and $\hat{\beta}_3$ are the unknowns. The known terms appearing on the right-hand side may be written in the form of a determinant

$$\begin{vmatrix} \sum x_{2i}^2 & \sum x_{2i}x_{3i} \\ \sum x_{2i}x_{3i} & \sum x_{3i}^2 \end{vmatrix} = |A|$$

The variance of each parameter is the product of σ^2 multiplied by the ratio of the minor determinant[1] associated with this parameters divided by the (complete) determinant.

Thus

$$var(\hat{\beta}_2)=\sigma^2.\frac{\begin{vmatrix}\cancel{\sum x_{2i}^2} & \cancel{\sum x_{2i}x_{3i}}\\ \cancel{\sum x_{2i}x_{3i}} & \sum x_{3i}^2\end{vmatrix}}{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2\end{vmatrix}}=\sigma^2.\frac{\sum x_{3i}^2}{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2\end{vmatrix}}=\sigma^2.\frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2)-(\sum x_{2i}x_{3i})^2}$$

$$var(\hat{\beta}_3)=\sigma^2.\frac{\begin{vmatrix}\sum x_{2i}^2 & \cancel{\sum x_{2i}x_{3i}}\\ \cancel{\sum x_{2i}x_{3i}} & \cancel{\sum x_{3i}^2}\end{vmatrix}}{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2\end{vmatrix}}=\sigma^2.\frac{\sum x_{2i}^2}{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2\end{vmatrix}}=\sigma^2.\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2)-(\sum x_{2i}x_{3i})^2}$$

Examing the above expressions of the variances of the coefficient estimates we may generalize as follows. The variances of the estimates of the model including k-explanatory variables can be computed by the ratio of two determinants: the determinant appearing in the numerator is the minor formed after striking out the row and column of the terms corresponding to the coefficient whose variance is being computed; the determinant appearing in the denominator is the complete determint of the known terms appearing on the rihgt-hand side of the normal equations. For example the variance of $\hat{\beta}_K$ is given by the following expression.

$$var(\hat{\beta}_K)=\sigma^2.\frac{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\cdots & \cancel{\sum x_{2i}x_{Ki}}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2 \quad \cdots & \cancel{\sum x_{3i}x_{Ki}}\\ \vdots & \vdots & \vdots\\ \cancel{\sum x_{2i}x_{Ki}} & \cancel{\sum x_{3i}x_{Ki}} & \cancel{\sum x_{Ki}^2}\end{vmatrix}}{\begin{vmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\cdots & \sum x_{2i}x_{Ki}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2 \quad \cdots & \sum x_{3i}x_{Ki}\\ \vdots & \vdots & \vdots\\ \sum x_{2i}x_{Ki} & \sum x_{3i}x_{Ki} & \sum x_{Ki}^2\end{vmatrix}}$$

Example 1

The table below contains observations on the quantity demanded (Y) of a certain commodity, its price (X2) and consumers' income (X3). Fit a linear regression to these observations and test the overall goodness of fit (with R^2) as well as the statistical reliability of the estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.

Quantity demanded	100	75	80	70	50	65	90	100	110	60
Price	5	7	6	6	8	7	5	4	3	9
Income	1,000	600	1,200	500	300	400	1,300	1,100	1,300	300

[1] The minor determinant for each parameter is formed by the elements of the determinant left after striking out the row and column including the parameter

5.3 HYPOTHESIS TESTING IN MULTIPLE REGRESSION:

Once we go beyond the simple world of the two-variable linear regression model, hypothesis testing assumes several interesting forms, such as the following:

1. Testing hypotheses about an individual partial regression coefficient
2. Testing the overall significance of the estimated multiple regression model, that is, finding out if all the partial slope coefficients are simultaneously equal to zero

Hypothesis testing about individual regression coefficients

The procedure for testing the significance of the partial regression coefficients is the same as that discussed for the two-variable case, i.e. we just use the t-test (or Z-test) to test a hypothesis about any individual partial regression coefficient.

Assuming $u_i \sim N(0, \sigma^2)$, the estimators $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_1$ are BLUE and normally distributed with means equal to true β_2, β_3, and β_1 and the variances given in section 5.2.4. Furthermore, $(n-3)\,\hat{\sigma}^2/\sigma^2$ follows the χ^2 distribution with n − 3 df.

Upon replacing σ^2 by its unbiased estimator $\hat{\sigma}^2$ in the computation of the standard errors, each of the following variable

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)}$$

follows the t distribution with n− 3 df.

Therefore, the t distribution can be used to establish confidence intervals as well as test statistical hypotheses about the true population partial regression coefficients. Similarly, the χ^2 distribution can be used to test hypotheses about the true σ^2.

To illustrate the procedure consider the following test

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0 \qquad \text{i = 1, 2, …, K}$$

$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-K}$, K = k + 1 = number of variables. Similarly the 100(1-α) percent level of confidence interval for β_i will be given by $\beta_i = \hat{\beta}_i \pm t_{\alpha/2,n-K} se\big(\hat{\beta}_i\big)\ \forall_i \in [1, K]$

Example 2

A production function is estimated as

$$\hat{Y} = \underset{(0.78)}{4.0} + \underset{(0.102)}{0.7X_2} + \underset{(0.102)}{0.2X_3} \qquad \begin{matrix} R^2 = 0.86 \\ n = 23 \end{matrix}$$

where X_2= labor, X_3= capital, and Y = output

Test the hypothesis $\beta_2 = 0$, $\beta_3 = 0$ at α = 5% using the test of significance and confidence interval approach

Testing the overall significance of the sample regression

Throughout the previous section we were concerned with testing the significance of the estimated partial regression coefficients individually, that is, under the separate hypothesis that each true population partial regression coefficient was zero. But now consider the following hypothesis:

$$H_0: \beta_2 = \beta_3 = 0$$

This null hypothesis is a joint hypothesis that β_2 and β_3 are jointly or simultaneously equal to zero. A test of such a hypothesis is called a test of the overall significance of the observed or estimated regression line, that is, whether Y is linearly related to both X_2 and X_3. Can the joint hypothesis given above be tested by testing the significance of $\hat{\beta}_2$ and $\hat{\beta}_3$ individually as in the previous section? The answer is no, and the reasoning is as follows.

In testing the individual significance of an observed partial regression coefficient, we assume implicitly that each test of significance was based on a different (i.e., independent) sample. But to test a joint hypothesis, if we use the same sample data, we shall be violating the assumption underlying the test procedure.

In other words, although the statements

$$P\big[\hat{\beta}_2 - t_{\alpha/2} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} se(\hat{\beta}_2)\big] = 1 - \alpha$$

$$P\big[\hat{\beta}_3 - t_{\alpha/2} se(\hat{\beta}_3) \le \beta_3 \le \hat{\beta}_3 + t_{\alpha/2} se(\hat{\beta}_3)\big] = 1 - \alpha$$

are individually true, it is not true that the probability that the intervals

$$[\hat{\beta}_2 \pm t_{\alpha/2} se(\hat{\beta}_2), \hat{\beta}_3 \pm t_{\alpha/2} se(\hat{\beta}_3)]$$

simultaneously include β_2 and β_3 is$(1-\alpha)^2$, because the intervals may not be independent when the same data are used to derive them. To state the matter differently, testing a series of single [individual] hypotheses is not equivalent to testing those same hypotheses jointly. The intuitive reason for this is that in a joint test of several hypotheses any single hypothesis is "affected'' by the information in the other hypotheses.

The upshot of the preceding argument is that for a given example (sample) only one confidence interval or only one test of significance can be obtained. How, then, does one test the simultaneous null hypothesis that β_2 = β_3 = 0? The answer follows.

The analysis of variance approach to testing the overall significance of an observed multiple regression: the F test

For reasons just explained, we cannot use the usual t test to test the joint hypothesis that the true partial slope coefficients are zero simultaneously. However, this joint hypothesis can be tested by the **analysis of variance** (ANOVA) technique.

In Chapter 3, we developed the following identity:

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2$$

that is, TSS = ESS + RSS, which decomposed the total sum of squares (TSS) into two components: explained sum of squares (ESS) and residual sum of squares (RSS). A study of these components of TSS is known as the **analysis of variance** (ANOVA) from the regression viewpoint. In this technique, we test the significance of the ESS or the null hypothesis $H_0: \beta_i = 0$.

Under the assumptions of the regression model $u_i \sim N(0, \sigma^2)$

$\frac{RSS}{\sigma^2} = \frac{\sum \hat{u}_i^2}{\sigma^2} \sim \chi^2_{(n-K)}$ and

$\frac{ESS}{\sigma^2} = \frac{\sum \hat{y}_i^2}{\sigma^2} \sim \chi^2_{(K-1)}$

Further the two chi-square distributions are independent and thus under the null hypothesis $H_0: \beta_i = 0$

$$F = \frac{\chi^2_{(K-1)}/K-1}{\chi^2_{(n-K)}/n-K} = \frac{ESS/K-1}{RSS/n-K} \sim F_{(K-1,n-K)}$$

What use can be made of the preceding F ratio? Let us take the two variable case

$$F = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)}$$

$$F = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$$

It can be shown that

$$E\left(\hat{\beta}_2^2 \sum x_i^2\right) = \sigma^2 + \beta_2^2 \sum x_i^2 \text{ and}$$

$$E\left(\frac{\sum \hat{u}_i^2}{n-2}\right) = E(\hat{\sigma}^2) = \sigma^2$$

(Note that β_2 and σ^2appearing on the right sides of these equations are the true parameters.) Therefore, if β_2 is in fact zero, both the above equations provide us with identical estimates of true σ^2. In this situation, the explanatory variable X has no linear influence on Y whatsoever and the entire variation in Y is explained by the random disturbances ui. If, on the other hand, β_2 is not zero, the two equations will be different and part of the variation in Y will be ascribable to X. Therefore, the F ratio provides a test of the null hypothesis $H_0\colon \beta_2 = 0$. Since all the quantities entering into this equation can be obtained from the available sample, this F ratio provides a test statistic to test the null hypothesis that true β_2 is zero. All that needs to be done is to compute the F ratio and compare it with the critical F value obtained from the F tables at the chosen level of significance.

Next the ANOVA table will be prepared as follows

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean sum of squares (MSS)
Explained sum of squares	$\sum \hat{y}_i^2$	$\hbar$ = K-1	$\sum \hat{y}_i^2 \,/(K-1)$
Residual sum of squares	$\sum \hat{u}_i^2$	n-K	$\sum \hat{u}_i^2/(n-K) = \hat{\sigma}^2$
Total sum of squares	$\sum \boldsymbol{y_i^2}$	n-1	Fratio = ratio of MSS

Associated with any sum of squares is its df, the number of independent observations on which it is based. TSS has n-1df because we lose 1 df in computing the sample mean $\bar{Y}$. RSS has n−K df. (Why?) ESS has $\hbar$ = K-1df. Mean sum of squares is obtained by dividing SS by their df.

☞ We can generalize the F-testing procedure as follows.
Given the K-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i$$

To test the hypothesis

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

(i.e., all slope coefficients are simultaneously zero) versus

H1: Not all slope coefficients are simultaneously zero

Compute

$$F = \frac{ESS/df}{RSS/df} = \frac{ESS/(K-1)}{RSS/(n-K)}$$

If F > $F_\alpha(K-1, n-K)$, reject H0; otherwise you do not reject it, where $F_\alpha(K-1, n-K)$ is the critical F value at the α level of significance and (K − 1) numerator df and (n − K) denominator df.

A summary of the F-statistic

Null hypothesis	Alternative hypothesis	Critical region
$\boldsymbol{H_0}$	H1	Reject $\boldsymbol{H_0}$ if
$\boldsymbol{\sigma_1^2 = \sigma_2^2}$	$\sigma_1^2 > \sigma_2^2$	$\frac{S_1^2}{S_2^2} > F_{\alpha,ndf,ddf}$
$\boldsymbol{\sigma_1^2 = \sigma_2^2}$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{S_1^2}{S_2^2} > F_{\alpha/2,ndf,ddf}$
		Or $\frac{S_1^2}{S_2^2} < F_{(1-\alpha/2),ndf,ddf}$

Notes:

1. σ_1^2 and σ_2^2 are the two population variances.

2. S_1^2 and S_2^2 are the two sample variances.

3. ndf and ddf denote, respectively, the numerator and denominator df.

4. In computing the F ratio, put the larger S^2value in the numerator.

5. The critical F values are given in the last column. The first subscript of F is the level of significance and the second subscript is the numerator and denominator df.

Example 3

With reference to the production function regression in the previous example suppose you are given with the following intermediary results

$$\text{Normal equations} \begin{cases} 12\hat{\beta}_2 + 8\hat{\beta}_3 = 10 \\ 8\hat{\beta}_2 + 12\hat{\beta}_3 = 8 \end{cases}$$

Test the joint hypothesis $H_0\colon \beta_2 = \beta_3 = 0$

An important relationship between R^2 and F

There is an intimate relationship between the coefficient of determination R^2 and the F test used in the analysis of variance. More generally, in the K-variable case, if we assume that the disturbances are normally distributed and that the null hypothesis is

$$H_0\colon \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

then it follows that

$$F = \frac{ESS/(K-1)}{RSS/(n-K)}$$

follows the F distribution with K − 1 and n − K df. (Note: The total number of parameters to be estimated is K, of which one is the intercept term.)

Let us manipulate the above equation as follows:

$$F = \frac{n-K}{K-1}\frac{ESS}{RSS}$$

$$F = \frac{n-K}{K-1}\frac{ESS}{TSS-ESS}$$

$$F = \frac{n-K}{K-1}\frac{ESS/TSS}{1-(ESS/TSS)}$$

$$F = \frac{n-K}{K-1}\frac{R^2}{1-R^2}$$

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

where use is made of the definition R^2 = ESS/TSS. The above shaded equation shows how F and R^2 are related. These two vary directly. When R^2 = 0, F is also zero. The larger the R^2, the greater the F value. In the limit, when R^2 = 1, F is infinite. Thus the F test, which is a measure of the overall significance of the estimated regression, is also a test of significance of R^2. In other words, testing the null hypothesis $H_0: \beta_2 = \beta_3 = \cdots = \beta_K = 0$ is equivalent to testing the null hypothesis that (the population) R^2 is zero.

One advantage of the F test expressed in terms of R^2 is its ease of computation: All that one needs to know is the R^2 value. Therefore, the overall F test of significance can be recast in terms of R^2 as shown in the table below

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean sum of squares (MSS)
Explained sum of squares	R^2.TSS	$\cancel{k}$ = K-1	$R^2.TSS/K-1$
Residual sum of squares	(1- R^2).TSS	n-K	$(1-R^2).TSS/(n-K)$
Total sum of squares	TSS	n-1	

$$F = \frac{R^2.\cancel{TSS}/K-1}{(1-R^2).\cancel{TSS}/(n-K)} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

5. DUMMY VARIABLE REGRESSION ANALYSIS

In Chapter 3 we discussed briefly the four types of variables that one generally encounters in empirical analysis: These are: **ratio scale, interval scale, ordinal scale,** and **nominal scale.** The types of variables that we have encountered in the preceding chapters were essentially *ratio scale.* But this should not give the impression that regression models can deal only with ratio scale variables. Regression models can also handle other types of variables mentioned previously. In this chapter, we consider models that may involve not only ratio scale variables but also **nominal scale** variables. Such variables are also known as **indicator variables, categorical variables, qualitative variables,** or **dummy variables.**

5.1 DEFINITIONS AND THE NATURE OF DUMMY VARIABLES

In regression analysis the dependent variable, or regressand, is frequently influenced not only by ratio scale variables (e.g., income, output, prices, costs, height, temperature) but also by variables that are essentially qualitative, or nominal scale, in nature, such as sex, race, color, religion, nationality, geographical region, political upheavals, and party affiliation. For example, holding all other factors constant, female workers are found to earn less than their male counterparts or nonwhite workers are found to earn less than whites. This pattern may result from sex or racial discrimination, but whatever the reason, qualitative variables such as sex and race seem to influence the regressand and clearly should be included among the explanatory variables, or the regressors.

Since such variables usually indicate the presence or absence of a “quality” or an attribute, such as male or female, black or white, Catholic or non-Catholic, Democrat or Republican, they are essentially *nominal scale* variables. One way we could “quantify” such attributes is by constructing artificial variables that take on values of 1 or 0, 1 indicating the presence (or

possession) of that attribute and 0 indicating the absence of that attribute. For example 1 may indicate that a person is a female and 0 may designate a male; or 1 may indicate that a person is a college graduate, and 0 that the person is not, and so on. Variables that assume such 0 and 1 values are called **dummy variables.** *Such variables are thus essentially a device to classify*

data into mutually exclusive categories such as male or female.

Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called **Analysis of Variance (ANOVA) models**

5.2 ANOVA MODELS

It is not absolutely essential that dummy variables take the values of 0 and 1. The pair (0,1) can be transformed into any other pair by a linear function such that $Z = a + bD(b \neq 0)$, where a and b are constants and where $D = 1$ or 0. When $D = 1$, we have $Z = a + b$, and when $D = 0$, we have $Z = a$. Thus the pair (0, 1) becomes $(a, a + b)$. For example, if $a = 1$ and $b = 2$, the dummy variables will be (1, 3). *This expression shows that qualitative, or dummy, variables do not have a natural scale of measurement.* That is why they are described as nominal scale variables.

ANOVA models are used to assess the statistical significance of the relationship between a quantitative regressand and qualitative or dummy regressors. They are often used to compare the differences in the mean values of two or more groups or categories, and are therefore more general than the *t* test which can be used to compare the means of two groups or categories only.

To illustrate the ANOVA models, consider the following example.

Example 5.1

Public school teachers' salaries by geographical region

Table 5.1 gives data on average salary (in dollars) of public school teachers in 51 states. These 51 areas are classified into three geographical regions: (1) Northeast and North Central (21 states in all), (2) South (17 states in all), and (3) West (13 states in all). For the time being, do not worry about the format of the table and the other data given in the table.

Suppose we want to find out if the average annual salary (AAS) of public school teachers differs among the three geographical regions of the country. If you take the simple arithmetic average of the average salaries of the teachers in the three regions, you will find that these averages for the three regions are as follows: $24,424.14 (Northeast and North Central), $22,894 (South), and $26,158.62 (West). These numbers look different, but are they statistically different from one another? There are various statistical techniques to compare two or more mean values, which generally go by the name of **analysis of variance.** But the same objective can be accomplished within the framework of regression analysis.

To see this, consider the following model:

$Yi = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + ui$ ---**(5.2.1)**

where Yi = (average) salary of public school teacher in state i

D_{2i} = 1 if the state is in the Northeast or North Central

= 0 otherwise (i.e., in other regions of the country)

$D3i$ = 1 if the state is in the South = 0 otherwise (i.e., in other regions of the country)

Example 5.1

Table 5.1 Average Salary Of Public School Teachers, By State

Salary	Spending	*D*2	*D*3	Salary	Spending	*D*2	*D*3
19583	3346	1	0	22795	3366	0	1
20263	3114	1	0	21570	2920	0	1
20325	3554	1	0	22080	2980	0	1
26800	4642	1	0	22250	3731	0	1
29470	4669	1	0	20940	2853	0	1
26610	4888	1	0	21800	2533	0	1
30678	5710	1	0	22934	2729	0	1
27170	5536	1	0	18443	2305	0	1
25853	4168	1	0	19538	2642	0	1
24500	3547	1	0	20460	3124	0	1
24274	3159	1	0	21419	2752	0	1
27170	3621	1	0	25160	3429	0	1
30168	3782	1	0	22482	3947	0	0
26525	4247	1	0	20969	2509	0	0
27360	3982	1	0	27224	5440	0	0
21690	3568	1	0	25892	4042	0	0
21974	3155	1	0	22644	3402	0	0
20816	3059	1	0	24640	2829	0	0
18095	2967	1	0	22341	2297	0	0
20939	3285	1	0	25610	2932	0	0
22644	3914	1	0	26015	3705	0	0
24624	4517	0	1	25788	4123	0	0
27186	4349	0	1	29132	3608	0	0
33990	5020	0	1	41480	8349	0	0
23382	3594	0	1	25845	3766	0	0
20627	2821	0	1				

***Note**: D_2 = 1 for states in the Northeast and North Central; 0 otherwise.*

D_3 = 1 for states in the South; 0 otherwise.

Note that (5.2.1) is like any multiple regression model considered previously, except that, instead of quantitative regressors, we have only qualitative, or dummy, regressors, taking the value of 1 if the observation belongs to a particular category and 0 if it does not belong to that category or group. *Hereafter, we shall designate all dummy variables by the letter D.*

Table 5.1 shows the dummy variables thus constructed.

What does the model (5.2.1) tell us? Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of (5.2.1) on both sides, we obtain:

Mean salary of public school teachers in the Northeast and North Central:

$E(Yi\ |D2i = 1, D3i = 0) = \beta 1 + \beta 2$--- **(5.2.2)**

Mean salary of public school teachers in the South:

$E(Yi\,|D2i = 0, D3i = 1) = \beta1 + \beta3$-- **(5.2.3)**

You might wonder how we find out the mean salary of teachers in the West. If you guessed

that this is equal to β1, you would be absolutely right, for

Mean salary of public school teachers in the West:

$E(Yi\,|D2i = 0, D3i = 0) = \beta1$-- **(5.2.4)**

In other words, the mean salary of public school teachers in the West is given by the intercept, β1, in the multiple regression (5.2.1), and the “slope” coefficients β2 and β3 tell by how much the mean salaries of teachers in the Northeast and North Central and in the South differ from the mean salary of teachers in the West. But how do we know if these differences are statistically significant? Before we answer this question, let us present the results based on the regression (5.2.1). Using the data given in Table 5.1, we obtain the following results:

$$\begin{aligned}
&\hat{Y}i = 26{,}158.62 - 1734.473D2i - 3264.615D3i \\
&\text{se} = (1128.523) \quad (1435.953) \quad (1499.615) \\
&t = (23.1759) \quad (-1.2078) \quad (-2.1776) \qquad\qquad \mathbf{(5.2.5)} \\
&\quad (0.0000)^* \quad (0.2330)^* \quad (0.0349)^* \qquad R2 = 0.0901
\end{aligned}$$

where * indicates the *p* values.

As these regression results show, the mean salary of teachers in the West is about $26,158, that of teachers in the Northeast and North Central is lower by about $1734, and that of teachers in the South is lower by about $3265. The actual mean salaries in the last two regions can be easily obtained by adding these differential salaries to the mean salary of teachers in the West, as shown in Eqs. (5.2.3) and (5.2.4). Doing this, we will find that the mean salaries in the latter two regions are about $24,424 and $22,894. But how do we know that these mean salaries are statistically different from the mean salary of teachers in the West, the comparison category? That is easy enough. All we have to do is to find out if each of the “slope” coefficients in (5.2.5) is statistically significant. As can be seen from this regression, the estimated slope coefficient for Northeast and North Central is not statistically significant, as its *p* value is 23 percent, whereas that of the South is statistically significant, as the *p* value is only about 3.5 percent. Therefore, the overall conclusion is that statistically the mean salaries of public school teachers in the West and the Northeast and North Central are about the same but the mean salary of teachers in the South is statistically significantly lower by about $3265.

A caution is in order in interpreting these differences. The dummy variables will simply point out the differences, if they exist, but they do not suggest the reasons for the differences. Differences in educational levels, in cost of living indexes, in gender and race may all have some effect on the observed differences. Therefore, unless we take into account all the other variables that may affect a teacher’s salary, we will not be able to pin down the cause(s) of the differences.

From the preceding discussion, it is clear that all one has to do is see if the coefficients attached to the various dummy variables are individually statistically significant. This example also shows how easy it is to incorporate qualitative, or dummy, regressors in the regression models.

Caution in the Use of Dummy Variables

Although they are easy to incorporate in the regression models, one must use the dummy variables carefully. In particular, consider the following aspects:

1. In Example 5.1, to distinguish the three regions, we used only two dummy variables, *D*2 and *D*3. Why did we not use three dummies to distinguish the three regions? Suppose we do that and write the model (5.2.1) as:

$Yi = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + ui$-- **(5.2.6)**

where D_{1i} takes a value of 1 for states in the West and 0 otherwise. Thus, we now have a dummy variable for each of the three geographical regions.

Using the data in Table 5.1, if you were to run the regression (5.2.6), the computer will “refuse” to run the regression (try it). Why? The reason is that in the setup of (5.2.6) where you have a dummy variable for each category or group and also an intercept, you have a case of **perfect collinearity,** that is, exact linear relationships among the variables. Why? Refer to Table 5.1.

Imagine that now we add the *D*1 column, taking the value of 1 whenever a state is in the West and 0 otherwise. Now if you add the three *D* columns horizontally, you will obtain a column that has 51 ones in it. But since the value of the intercept " is (implicitly) 1 for each observation, you will have a column that also contains 51 ones. In other words, the sum of the three *D* columns will simply reproduce the intercept column, thus leading to perfect collinearity. In this case, estimation of the model (5.2.6) is impossible.

The message here is: **If a qualitative variable has m categories, introduce only (*m* − 1) dummy variables.** In our example, since the qualitative variable “region” has three categories, we introduced only two dummies. If you do not follow this rule, you will fall into what is called the **dummy variable trap,** that is, the situation of perfect collinearity or perfect multicollinearity, if there is more than one exact relationship among the variables.

This rule also applies if we have more than one qualitative variable in the model, an example of which is presented later. Thus we should restate the preceding rule as: **For each qualitative regressor the number of dummy variables introduced must be one less than the categories of that variable.** Thus, if in Example 5.1 we had information about the gender of the teacher, we would use an additional dummy variable (but not two) taking a value of 1 for female and 0 for male or vice versa.

2. The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category.** And all comparisons are made in relation to the benchmark category.

3. The intercept value (β1) represents the *mean value* of the benchmark category. In Example 5.1, the benchmark category is the Western region. Hence, in the regression (5.2.5) the intercept value of about 26,159 represents the mean salary of teachers in the Western states.

4. The coefficients attached to the dummy variables in (5.2.1) are known as the **differential intercept coefficients** because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the benchmark category. For example, in (5.2.5), the value of about −1734 tells us that the mean salary of teachers in the Northeast or North Central is smaller by about $1734 than the mean salary of about $26,159 for the benchmark category, the West.

5. If a qualitative variable has more than one category, as in our illustrative example, the choice of the benchmark category is strictly up to the researcher. Sometimes the choice of the benchmark is dictated by the particular problem at hand. In our illustrative example, we could have chosen the South as the benchmark category. In that case the regression results given in (5.2.5) will change, because now all comparisons are made in relation to the South. Of course, this will not change the overall conclusion of our example (why?). In this case, the intercept value will be about $22,894, which is the mean salary of teachers in the South.

6. We warned above about the dummy variable trap. There is a way to circumvent this trap by introducing as many dummy variables as the number of categories of that variable, *provided we do not introduce the intercept in such a model.* Thus, if we drop the intercept term from (5.2.6), and consider the following model,

$Yi = \beta 1 D1i + \beta 2 D2i + \beta 3 D3i + ui$ -- **(5.2.7)**

we do not fall into the dummy variable trap, as there is no longer perfect collinearity. *But make sure that when you run this regression, you use the no intercept option in your regression package.*

How do we interpret regression (5.2.7)? If you take the expectation of (5.2.7), you will find that:

β_1 = mean salary of teachers in the West

β_2 = mean salary of teachers in the Northeast and North Central.

β_3 = mean salary of teachers in the South.

In other words, *with the intercept suppressed, and allowing a dummy variable for each category, we obtain directly the mean values of the various categories.*

The results of (5.2.7) for our illustrative example are as follows:

$\hat{Y}i = 26{,}158.62D_{1i} + 24{,}424.14D_{2i} + 22{,}894D_{3i}$

se = (1128.523) (887.9170) (986.8645) **(5.2.8)**

t = (23.1795)* (27.5072)* (23.1987)*

$R^2 = 0.0901$

where * indicates that the p values of these t ratios are very small.

As you can see, the dummy coefficients give directly the mean (salary) values in the three regions, West, Northeast and North Central, and South.

7. Which is a better method of introducing a dummy variable: (1) introduce a dummy for each category and omit the intercept term or (2) include the intercept term and introduce only ($m - 1$) dummies, where m is the number of categories of the dummy variable? As Kennedy notes: Most researchers find the equation with an intercept more convenient because it allows them to address more easily the questions in which they usually have the most interest, namely, whether or not the categorization makes a difference, and if so, by how much. If the categorization does make a difference, by how much is measured directly by the dummy variable coefficient estimates. Testing whether or not the categorization is relevant can be done by running a t test of a dummy variable coefficient against zero (or, to be more general, an F test on the appropriate set of dummy variable coefficient estimates)

5.2.1 ANOVA MODELS WITH TWO QUALITATIVE VARIABLES

In the previous section we considered an ANOVA model with one qualitative variable with three categories. In this section we consider another ANOVA model, but with two qualitative variables, and bring out some additional points about dummy variables.

Example 5.2
Hourly wages in relation to marital status and region of residence
From a sample of 528 persons, the following regression results were obtained:

$\hat{}Yi = 8.8148 + 1.0997D2i - 1.6729D3i$

se = (0.4015) (0.4642) (0.4854)

t = (21.9528) (2.3688) (−3.4462) **(5.2.9)**

(0.0000)* (0.0182)* (0.0006)*

$R^2 = 0.0322$

where *Y* = hourly wage ($)

*D*2 = married status, 1 = married, 0 = otherwise

*D*3 = region of residence; 1 = South, 0 = otherwise

and * denotes the *p* values.

In this example we have two qualitative regressors, each with two categories. Hence we have assigned a single dummy variable for each category Which is the benchmark category here? Obviously, it is unmarried, non-South residence. In other words, unmarried persons who do not live in the South are the omitted category. Therefore, all comparisons are made in relation to this group. The mean hourly wage in this benchmark is about $8.81. Compared with this, the average hourly wage of those who are married is higher by about $1.10, for an actual average wage of $9.9 (= 8.81 + 1.10). By contrast, for those who live in the South, the average hourly wage is lower by about $1.67, for an actual average hourly wage of $7.14.

Are the preceding average hourly wages statistically different compared to the base category? They are, for all the differential intercepts are statistically significant, as their *p* values are quite low.

The point to note about this example is this: *Once you go beyond one qualitative variable, you have to pay close attention to the category that is treated as the base category, since all comparisons are made in relation to that category. This is especially important when you have several qualitative regressors, each with several categories.* But the mechanics of introducing several qualitative variables should be clear by now.

5.3 The ANCOVA Models: Regression With A Mixture Of Quantitative And Qualitative Regressors

ANOVA models of the type discussed in the preceding sections, although common in fields such as sociology, psychology, education, and market research, are not that common in economics. Typically, in most economic research a regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing an admixture of quantitative and qualitative variables are called **analysis of covariance (ANCOVA) models.** ANCOVA models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called **covariates** or **control variables,** in a model that includes both quantitative and qualitative, or dummy, regressors. We now illustrate the ANCOVA models.

To motivate the analysis, let us reconsider Example 5.1 by maintaining that the average salary of public school teachers may not be different in the three regions if we take into account any variables that cannot be standardized across the regions. Consider, for example, the variable *expenditure on public schools by local authorities,* as public education is primarily a local and state question. To see if this is the case, we develop the following model:

$Yi = \beta 1 + \beta 2D2i + \beta 3D3i + \beta 4Xi + ui$-- **(5.3.1)**

where Yi = average annual salary of public school teachers in state ($)

Xi = spending on public school per pupil ($)

$D2i$ = 1, if the state is in the Northeast or North Central

= 0, otherwise

$D3i$ = 1, if the state is in the South

= 0, otherwise

The data on X are given in Table 5.1. Keep in mind that we are treating the West as the benchmark category. Also, note that besides the two qualitative regressors, we have a quantitative variable, X, which in the context of the ANCOVA models is known as a **covariate,** as noted earlier.

Example 5.3
Teacher’s salary in relation to region and spending on public school per pupil
From the data in Table 5.1, the results of the model (5.3.1) are as follows:

$$\hat{} Yi = 13{,}269.11 - 1673.514D2i - 1144.157D3i + 3.2889Xi$$

$$\text{se} = (1395.056) \quad (801.1703) \quad (861.1182) \quad (0.3176) \qquad \textbf{(5.3.2)}$$

$$t = (9.5115)^* \quad (-2.0889)^* \quad (-1.3286)^{**} \quad (10.3539)^*$$

$$R^2 = 0.7266$$

where * indicates p values less than 5 percent, and ** indicates p values greater than 5 percent. As these results suggest, *ceteris paribus:* as public expenditure goes up by a dollar, on average, a public school teacher’s salary goes up by about $3.29. Controlling for spending on education, we now see that the differential intercept coefficient is significant for the Northeast and North-Central region, but not for the South. These results are different from those of (5.2.5). But this should not be surprising, for in (5.2.5) we did not account for the covariate, differences in per pupil public spending on education.

6. ECONOMETRIC PROBLEMS

6.1. MULTICOLLINREARITY

THE NATURE OF MULTICOLLINEARITY

Originally, the term multicollinearity meant the existence of a “perfect,” or exact, linear relationship among some or all explanatory variables of a regression model. For the K-variable regression involving explanatory variable X_1, X_2, . . . , X_K (where X_1 = 1 for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_K X_K = 0 \text{ .... } 6.1.1$$

Where λ_1, λ_2, . . . , λ_K are constants such that not all of them are zero simultaneously.

Today, however, the term multicollinearity is used in a broader sense to include the case of perfect multicollinearity, as shown by (6.1.1), as well as the case where the X variables are intercorrelated but not perfectly so, as follows:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_K X_K + v_i = 0 \text{ ……………………..………… } 6.1.2$$

where v_i is a stochastic error term.

To see the difference between perfect and less than perfect multicollinearity, assume, for example, that

$\lambda_2 \neq 0$. Then, (6.1.1) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_K}{\lambda_2} X_{Ki} \text{ ……………………..………… } 6.1.3$$

which shows how X_2 is exactly linearly related to other variables or how it can be derived from a linear combination of other X variables. In this situation, the coefficient of correlation between the variable X_2 and the linear combination on the right side of (6.1.3) is bound to be unity.

Similarly, if $\lambda_2 \neq 0$, Eq. (6.1.2) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_K}{\lambda_2} X_{Ki} - \frac{1}{\lambda_2} v_i \text{ ………………………} 6.1.4$$

which shows that X_2 is not an exact linear combination of other X’s because it is also determined by the stochastic error term v_i.

Why does the classical linear regression model assume that there is no multicollinearity among the X’s? The reasoning is this: **If multicollinearity is perfect in the sense of (6.1.1), the regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, as in (6.1.2), the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy.** The proofs of these statements are given as follows.

Estimation in the presence of perfect multicollinearity

The fact that in the case of perfect multicollinearity the regression coefficients remain indeterminate and their standard errors are infinite can be demonstrated readily in terms of the three-variable regression model. Using the deviation form, where all the variables are expressed as deviations from their sample means, we can write the three-variable regression model as

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.5$$

Now from Chapter 4 we obtain

$$\hat{\beta}_2 = \frac{(\sum x_{2i} y_i)\left(\sum x_{3i}^2\right) - (\sum x_{3i} y_i)(\sum x_{2i} x_{3i})}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum x_{3i} y_i)\left(\sum x_{2i}^2\right) - (\sum x_{2i} y_i)(\sum x_{2i} x_{3i})}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - (\sum x_{2i} x_{3i})^2}$$

Assume that X_{3i} = λX_{2i} , where λ is a nonzero constant. Substituting this into the formula for $\hat{\beta}_2$, we obtain

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum \lambda^2 x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.6$$

$$= \frac{0}{0}$$

which is an indeterminate expression. Verify that $\hat{\beta}_3$ is also indeterminate.

Why do we obtain the result shown in (6.1.6)? Recall the meaning of $\hat{\beta}_2$: It gives the rate of change in the average value of Y as X2 changes by a unit, holding X3 constant. But if X3 and X2 are perfectly collinear, there is no way X3 can be kept constant: As X2 changes, so does X3 by the factor λ. This means, there is no way of disentangling the separate influences of X2 and X3 from the given sample: For practical purposes X2 and X3 are indistinguishable.

Estimation in the presence of "high" but "imperfect" multicollinearity

Generally, there is no exact linear relationship among the X variables, especially in data involving economic time series. Thus, turning to the three-variable model in the deviation form given in (6.1.5), instead of exact multicollinearity, we may have

$$x_{3i} = \lambda x_{2i} + v_i \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.7$$

where λ ≠ 0 and where vi is a stochastic error term such that $\sum x_{2i} v_i = 0$. (Why?)

In this case, estimation of regression coefficients β2 and β3 may be possible. For example, substituting (6.1.7) into the formula for $\hat{\beta}_2$, we obtain

$$\hat{\beta}_2 = \frac{\sum (y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{\sum x_{2i}^2 (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.8$$

where use is made of $\sum x_{2i}v_i = 0$. A similar expression can be derived for $\hat{\beta}_3$.

Now, unlike (6.1.6), there is no reason to believe a priori that (6.1.8) cannot be estimated. Of course, if vi is sufficiently small, say, very close to zero, (6.1.8) will indicate almost perfect collinearity and we shall be back to the indeterminate case of (6.1.6).

SOURCES OF MULTICOLLINEARITY

Multicollinearity may be due to the following factors:

1. The data collection method employed, for example, sampling over a limited range of the values taken by the regressors in the population.
2. An overdetermined model. This happens when the model has more explanatory variables than the number of observations.
3. Inherent nature of the data. Especially in time series data, where the regressors included in the model share a common trend, that is, they all increase or decrease over time. For example, in the regression of consumption expenditure on income, wealth, and population, the regressors income, wealth, and population may all be growing over time at more or less the same rate, leading to collinearity among these variables.

CONSEQUENCES OF MULTICOLLINEARITY

In cases of near or high multicollinearity, one is likely to encounter the following consequences:

1. Although BLUE, the OLS estimators have large variances, making precise estimation difficult.
2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” (i.e., the true population coefficient is zero) more readily.
3. Also because of consequence 1, the t ratio of one or more coefficients tends to be statistically insignificant.
4. Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit, can be very high.
5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

The preceding consequences can be demonstrated as follows.

Large Variances of OLS Estimators

To see large variances, it is necessary and can be shown for the model (6.1.5) the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are given by

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)} \text{ ……………………………..} \quad 6.1.9$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)} \text{ ……………………………..} \quad 6.1.10$$

where r_{23} is the coefficient of correlation between X2 and X3.

It is apparent from (6.1.9) and (6.1.10) that as r_{23} tends toward 1, that is, as collinearity increases, the variances of the two estimators increase and in the limit when r_{23} = 1, they are infinite.

The speed with which variances and covariances increase can be seen with the variance-inflating factor (VIF), which is defined as

$$VIF = \frac{1}{(1-r_{23}^2)} \text{ ……………………………..…….. 6.1.11}$$

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. That is, as the extent of collinearity increases, the variance of an estimator increases, and in the limit it can become infinite. As can be readily seen, if there is no collinearity between X2 and X3, VIF will be 1.

Using this definition, we can express (6.1.9) and (6.1.10) as

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} VIF$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} VIF$$

which show that the variances of $\hat{\beta}_2$and $\hat{\beta}_3$are directly proportional to the VIF.

The results just discussed can be easily extended to the k-variable model. In such a model, the variance of the Kth coefficient can be expressed as:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2 (1-R_j^2)} \text{……………………………… 6.1.12}$$

where $\hat{\beta}_j$= (estimated) partial regression coefficient of regressor Xj

R_j^2 = R^2 in the regression of Xj on the remaining (K − 2) regressions

[Note: There are (K − 1) regressors in the K-variable regression model.]

$\sum x_j^2 = \sum (X_j - \bar{X}_j)^2$

We can also write (6.1.12) as

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} VIF_j \text{ ………………….………. (6.1.13)}$$

As you can see from this expression, $var(\hat{\beta}_j)$ is proportional to σ^2 and VIF but inversely proportional to $\sum x_j^2$. The last one states that the larger the variability in a regressor, the smaller the variance of the coefficient of that regressor, assuming the other two ingredients are constant, and therefore the greater the precision with which that coefficient can be estimated.

Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger.

“Insignificant” t Ratios

Recall that to test the null hypothesis that, say, β2 = 0, we use the t ratio, that is, $\hat{\beta}_2/se(\hat{\beta}_2)$, and compare the estimated t value with the critical t value from the t table. But as we have seen, in cases of high collinearity the estimated standard errors increase dramatically, thereby making the t values smaller. Therefore, in such cases, one will increasingly accept the null hypothesis that the relevant true population value is zero.

DETECTION OF MULTICOLLINEARITY

Multicollinearity is essentially a sample phenomenon, arising out of the largely non-experimental data collected in most social sciences. Multicollinearity is also a question of degree and not of kind. Some rules of thumb for detecting it or measuring its strength are as follows.

1. **High R^2 but few significant t ratios.** If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero.
2. **High pair-wise correlations among regressors.** Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem. High zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low (say, less than 0.50).
3. **High variance inflation factor.** The larger the value of VIF_j, the more “troublesome” or collinear the variable Xj. As a rule of thumb**,** if the VIF of a variable exceeds 10, which will happen if R_j^2 exceeds 0.90, that variable is said be highly collinear.

REMEDIAL MEASURES

What can be done if multicollinearity is serious? We have two choices:

(1) do nothing or (2) follow some rules of thumb.

Rule-of-Thumb Procedures

1. **Combining cross-sectional and time series data** (pooling the data)

2. **Dropping a variable(s).** When faced with severe multicollinearity, one of the “simplest” things to do is to drop one of the collinear variables. But in dropping a variable from the model we may be committing a **specification bias** or **specification error.**
3. **Transformation of variables.** Suppose we have time series data on consumption expenditure, income, and wealth. One reason for high multicollinearity between income and wealth in such data is that over time both the variables tend to move in the same direction. One way of minimizing this dependence is to proceed as follows.

If the relation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \text{ ………………………………… } 6.1.14$$

holds at time t, it must also hold at time t − 1 because the origin of time is arbitrary anyway. Therefore, we have

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \text{ ……………...……… } 6.1.15$$

If we subtract (6.5.2) from (6.5.1), we obtain

$$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t \text{ …….………… } 6.1.16$$

Where $v_t = u_t - u_{t-1}$. Equation (6.1.16) is known as the **first difference form** because we run the regression, not on the original variables, but on the differences of successive values of the variables.

The first difference regression model often reduces the severity of multicollinearity because, although the levels of X2 and X3 may be highly correlated, there is no a priori reason to believe that their differences will also be highly correlated.

Another commonly used transformation in practice is the **ratio transformation.** Consider the model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \text{ ………………………………… } 6.1.17$$

where Y is consumption expenditure in real dollars, X2 is GDP, and X3 is total population. Since GDP and population grow over time, they are likely to be correlated. One “solution” to this problem is to express the model on a per capita basis, that is, by dividing (6.1.17) by X3, to obtain:

$$\frac{Y_t}{X_{3t}} = \beta_1\left(\frac{1}{X_{3t}}\right) + \beta_2\left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \text{ ……………………...… } 6.1.18$$

Such a transformation may reduce collinearity in the original variables.

4. **Additional or new data.** Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be as serious as in the first sample. Sometimes simply increasing the size of the sample may attenuate the collinearity problem. For example, in the three-variable model we saw that

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)}$$

Now as the sample size increases, $\sum x_{2i}^2$ will generally increase. (Why?) Therefore, for any given r_{23}, the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β2 more precisely.

IS MULTICOLLINEARITY NECESSARILY BAD?

If the sole purpose of regression analysis is prediction or forecasting, then multicollinearity is not a serious problem because the higher the R^2, the better the prediction. But, if the objective of the analysis is not only prediction but also reliable estimation of the parameters, serious multicollinearity will be a problem because we have seen that it leads to large standard errors of the estimators.

6.2 HETEROSCEDASTICITY

6.2.1 THE NATURE OF HETEROSCEDASTICITY

As noted in Chapter 3, one of the important assumptions of the classical linear regression model is that the variance of each disturbance term u_i, conditional on the chosen values of the explanatory variables, is some constant number equal to σ^2. This is the assumption of **homoscedasticity**, or equal (homo) spread (scedasticity), that is, equal variance. Symbolically,

$$E\left(u_i^2\right) = \sigma^2 \quad i = 1, 2, \dots, n \text{ ………………………………. } 6.2.1$$

Diagrammatically, in the two-variable regression model homoscedasticity can be shown as in Figure 6.2.1

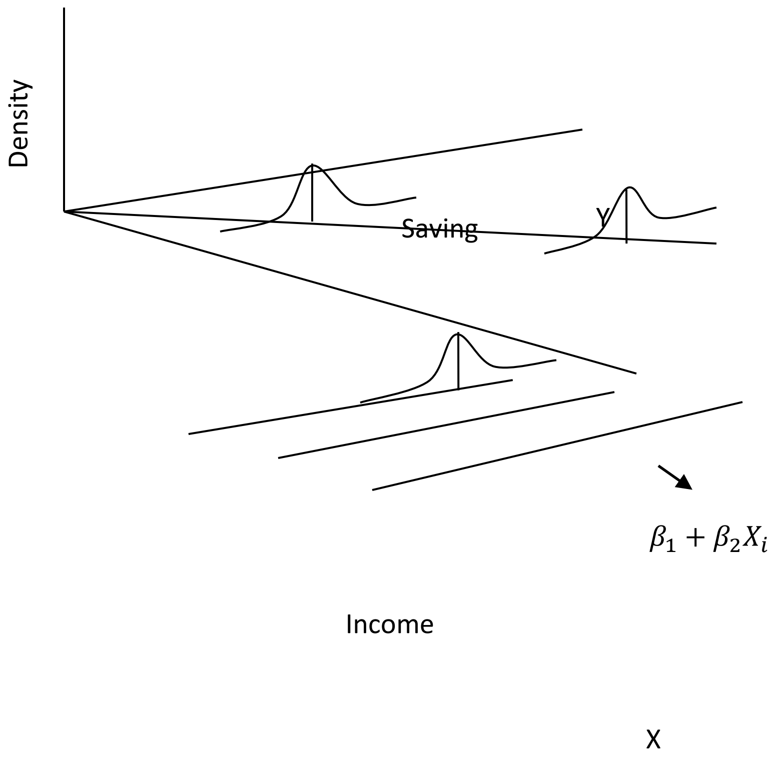


Fig 6.2.1 Homoscedastic disturbances

As Figure 6.2.1shows, the conditional variance of Y_i (which is equal to that of u_i), conditional upon the given X_i, remains the same regardless of the values taken by the variable X.

In contrast, consider Figure 6.2.2 below, which shows that the conditional variance of Y_i increases as X increases. Here, the variances of Y_i are not the same. Hence, there is **heteroscedasticity**. Symbolically,

$$E(u_i^2) = \sigma_i^2 \text{ .... } 6.2.1'$$

Notice the subscript of σ^2, which reminds us that the conditional variances of u_i (= conditional variances of Y_i) are no longer constant.

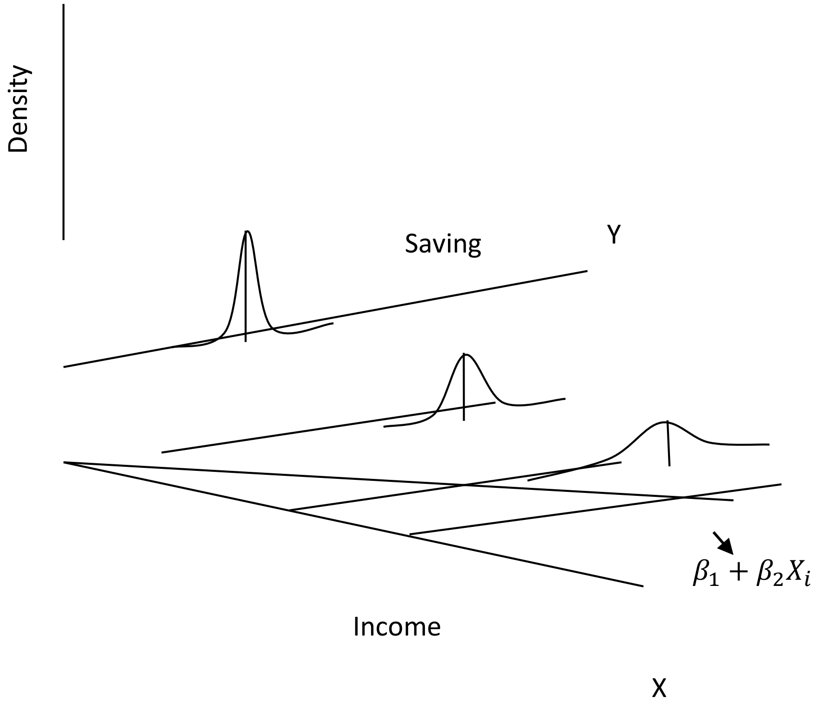


Fig 6.2.2 Heteroscedastic disturbances

To make the difference between homoscedasticity and heteroscedasticity clear, assume that in the two-variable model $Y_i = \beta_1 + \beta_2 X_i$, Y represents savings and X represents income. Figures 6.2.1 and 6.2.2 show that as income increases, savings on the average also increase. But in Figure 7.1 the variance of savings remains the same at all levels of income, whereas in Figure 6.2.2 it increases with income. It seems that in Figure 6.2.2 the higher income families on the average save more than the lower-income families, but there is also more variability in their savings.

6.2.2 SOURCES OF HETEROSCEDASTICITY

1. As people learn, their errors of behavior become smaller over time. In this case, σ_i^2 is expected to decrease. E.g. Typing errors Vs Hours of typing practice

2. As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence, σ_i^2 is likely to increase with income. Thus in the regression of savings on income one is likely to find σ_i^2 increasing with income (as in Figure 6.2.2) because people have more choices about their savings behavior.
3. As data collecting techniques improve, σ_i^2 is likely to decrease. Thus, banks that have sophisticated data processing equipment are likely to commit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities.
4. Heteroscedasticity can also arise as a result of the presence of outliers. An outlying observation, or outlier, is an observation that is much different (either very small or very large) in relation to the observations in the sample. More precisely, an outlier is an observation from a different population to that generating the remaining sample observations. The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis.
5. Heteroscedasticity may be due to omission of some important variables from the model. For example, in the demand function for a commodity, if we do not include the prices of commodities complementary to or competing with the commodity in question, the residuals obtained from the regression may give the distinct impression that the error variance may not be constant.
6. Another source of heteroscedasticity is **skewness** in the distribution of one or more regressors included in the model. Examples are economic variables such as income, wealth, and education. It is well known that the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top.

Note that the problem of heteroscedasticity is likely to be more common in cross-sectional than in time series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers or their families, firms, industries, or geographical subdivisions such as state, country, city, etc. Moreover, these members may be of different sizes, such as small, medium, or large firms or low, medium, or high income. In time series data, on the other hand, the variables tend to be of similar orders of magnitude because one generally collects the data for the same entity over a period of time. Examples are GNP, consumption expenditure, savings, or employment over some period of time.

6.2.3 OLS ESTIMATION IN THE PRESENCE OF HETEROSCEDASTICITY

In the presence of heteroscedasticity, $\hat{\beta}_2$ is still linear unbiased and consistent estimator. But, $\hat{\beta}_2$ is no longer best (i.e. have no minimum variance). Then what is BLUE in the presence of heteroscedasticity?

The answer is given in the following discussion.

The method of generalized least squares (GLS)

Why is the usual OLS estimator of β_2 not best, although it is still unbiased? Unfortunately, the usual OLS method does not make use of the “information” contained in the unequal variability of the dependent variable Y. It assigns equal weight or importance to each observation. But a method of estimation, known as **generalized least squares (GLS),** takes such information into account explicitly and is therefore capable of producing estimators that are BLUE. To see how this is accomplished, let us continue with the now-familiar two-variable model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \text{ ……………………………... } 6.2.3.1$$

which for ease of algebraic manipulation we write as

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \text{ …………………………… } 6.2.3.2$$

where X_{1i} = 1 for each i. One can see that these two formulations are identical.

Now assume that the heteroscedastic variances σ_i^2 are known. Divide (6.2.3.2) through by σ_i to obtain

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{1i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_{2i}}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \text{ ……………...……… } 6.2.3.3$$

which for ease of exposition we write as

$$Y_i^* = \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + u_i^* \text{ ......……………………… } 6.2.3.4$$

where the starred, or transformed, variables are the original variables divided by (the known) σ_i.We use the notation β_1^* and β_2^*, the parameters of the transformed model, to distinguish them from the usual OLS parameters $\beta_1$and β_2.

What is the purpose of transforming the original model? To see this, notice the following feature of the transformed error term u_i^*:

$$var(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2$$

$$= \frac{1}{\sigma_i^2} E(u_i^2) \qquad \text{since } \sigma_i^2 \text{ is known}$$

$$= \frac{1}{\sigma_i^2} (\sigma_i^2) \qquad \text{since } E\left(u_i^2\right) = \sigma_i^2$$

$$= 1$$

which is a constant. That is, the variance of the transformed disturbance term u_i^* is now homoscedastic. Since we are still retaining the other assumptions of the classical model, the finding that it is u^* that is homoscedastic suggests that if we apply OLS to the transformed model (6.2.3.3) it will produce estimators that are BLUE. In short, the estimated β_1^* and β_2^* are now BLUE and not the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

This procedure of transforming the original variables in such a way that the transformed variables satisfy the standard least-squares assumptions and then applying OLS to them is known as the method of **generalized least squares (GLS)**. The estimators thus obtained are known as **GLS estimators,** and it is these estimators that are BLUE.

The actual mechanics of estimating β_1^* and β_2^* are as follows. First, we write down the SRF of (6.2.3.3)

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1^* \left(\frac{X_{1i}}{\sigma_i}\right) + \hat{\beta}_2^* \left(\frac{X_{2i}}{\sigma_i}\right) + \left(\frac{\hat{u}_i}{\sigma_i}\right)$$

or

$$Y_i^* = \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + u_i^* \text{……………………………..} \; 6.2.3.5$$

Now, to obtain the GLS estimators, we minimize

$$\sum \hat{u}_i^{2^*} = \sum (Y_i^* - \beta_1^* X_{1i}^* - \beta_2^* X_{2i}^*)^2$$

that is,

$$\sum \left(\frac{\hat{u}_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^* \left(\frac{X_{1i}}{\sigma_i}\right) - \hat{\beta}_2^* \left(\frac{X_{2i}}{\sigma_i}\right)\right]^2 \text{………….} \; 6.2.3.6$$

$$\sum w_i \hat{u}_i^2 = \sum w_i \, (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)^2 \,, \quad \text{where } w_i = 1/\sigma_i^2$$

The actual mechanics of minimizing (7.3.6) follow the partial derivative techniques. Using this techniques, the GLS estimator of β_1^* and β_2^* is given as follows

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_{2i} Y_i) - (\sum w_i X_{2i})(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_{2i}^2) - (\sum w_i X_{2i})^2} \text{…………………...….} \; 6.2.3.7$$

$$\hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}_2^* \quad \text{where} \quad \bar{Y}^* = \frac{\sum w_i Y_i}{\sum w_i} \text{ and } \bar{X}_2^* = \frac{\sum w_i X_{2i}}{\sum w_i} \text{...} \; 6.2.3.8$$

Thus, in GLS we minimize a weighted sum of residual squares with $w_i = 1/\sigma_i^2$ acting as the weights, but in OLS we minimize an unweighted or (what amounts to the same thing) equally weighted RSS. As (6.2.3.6) shows, in GLS the weight assigned to each observation is inversely proportional to its σi , that is, observations coming from a population with larger σi will get relatively smaller weight and those from a population with smaller σi will get proportionately larger weight in minimizing the RSS (6.2.3.6).

Since (6.2.3.6) minimizes a weighted RSS, it is appropriately known as **weighted least squares (WLS),** and the estimators thus obtained and given in (6.2.3.7) and (6.2.3.8) are known as **WLS estimators.** But WLS is just a special case of the more general estimating technique, **GLS**. Note that if $w_i = w$, a constant for all i, $\hat{\beta}_2^*$ is identical with $\hat{\beta}_2$.

CONSEQUENCES OF HETEROSCEDASTICITY

I. The least square estimators become inefficient. I.e. no longer with minimum variance property although they are still linear and unbiased.

$$var(\hat{\beta}_2) = \frac{\sum X_{2i}^2 \sigma_i^2}{(\sum X_{2i}^2)^2} \,, \qquad \text{when heteroscedasticity is taken in to account}$$

II. The formulas for obtaining OLS variances of the estimates are biased, thus invalidating tests of significance.

III. The prediction of the Y for a given value of X would be inefficient (since they are based on the $\hat{\beta}$'s which have high variance.

DETECTION OF HETEROSCEDASTICITY

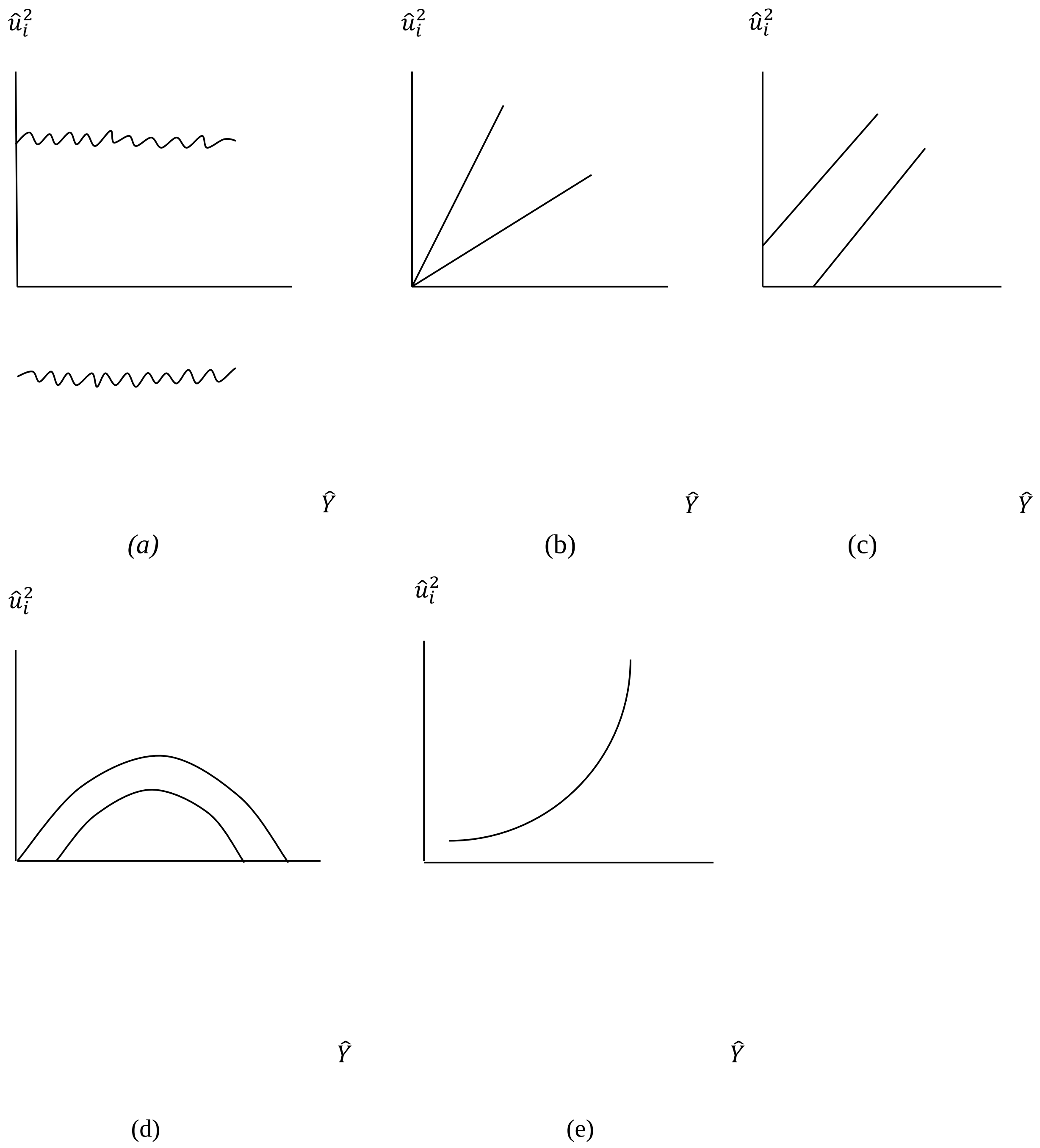
More often than not, in economic studies there is only one sample Y value corresponding to a particular value of X. And there is no way one can know σ_i^2 from just one Y observation. Therefore, in most cases involving econometric investigations, heteroscedasticity may be identified based on the examination of

the OLS residuals $\hat{u}_i$ since they are the ones we observe, and not the disturbances u_i. One hopes that they are good estimates of u_i, a hope that may be fulfilled if the sample size is fairly large.

Informal Methods

1) **Nature of the Problem** Very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered. For example,
   - The residual variance around the regression of consumption on income increased with income.
   - As a matter of fact, in cross-sectional data involving heterogeneous units, heteroscedasticity may be the rule rather than the exception.

2) **Graphical Method** If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then examination of the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern. Although $\hat{u}_i^2$are not the same thing as u_i^2 they can be used as proxies especially if the sample size is sufficiently large. An examination of the $\hat{u}_i^2$ may reveal patterns such as those shown in Figure 7.3.



(a) (b) (c)

(d) (e)

Fig 6.2.3 Hypothetical patterns of estimated squared residuals

In Figure 6.2.3, $\hat{u}_i^2$are plotted against $\hat{Y}_i$, the estimated Y_i from the regression line, the idea being to find out whether the estimated mean value of Y is systematically related to the squared residual. In Figure 6.2.3a it can be seen that there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. Figure 6.2.3 b to e, however, exhibits definite patterns. For instance, Figure 6.2.3c suggests a linear relationship, whereas Figure 6.2.3d and e indicates a quadratic relationship between $\hat{u}_i^2$ and $\hat{Y}_i$. Using such knowledge, one may transform the data in such a manner that the transformed data do not exhibit heteroscedasticity. Instead of plotting $\hat{u}_i^2$against $\hat{Y}_i$, one may plot them against one of the explanatory variables, especially if plotting $\hat{u}_i^2$ against $\hat{Y}_i$ results in the pattern shown in Figure 6.2.3a. This is useful for cross check.

REMEDIAL MEASURES

As we have seen, heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiency makes the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may be called for. There are two approaches to remediation: when σ_i^2 is known and when σ_i^2 is not known.

When σ_i^2 Is Known: The Method of Weighted Least Squares

If σ_i^2 is known, the most straightforward method of correcting heteroscedasticity is by means of weighted least squares, for the estimators thus obtained are BLUE.

When σ_i^2 Is Not Known

Since the true σ_i^2 are rarely known, there is another way of obtaining consistent estimates of the variances of OLS estimators even if there is heteroscedasticity. This is by doing some plausible assumptions about heteroscedasticity pattern. To illustrate this, let us revert to the two-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

We now consider several assumptions about the pattern of heteroscedasticity.

Assumption 1: The error variance is proportional to X_i^2:

$$E(u_i^2) = \sigma^2 X_i^2 \quad \text{.................................. 6.2.3.9}$$

If it is believed that the variance of u_i is proportional to the square of the explanatory variable X, one may transform the original model as follows. Divide the original model through by X_i:

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

$$= \beta_1 \frac{1}{X_i} + \beta_2 + v_i \text{ .... 6.2.3.10}$$

where v_i is the transformed disturbance term, equal to $\frac{u_i}{X_i}$. Now it is easy to verify that

$$E(v_i^2) = E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2} E(u_i^2)$$

$$= \sigma^2 \quad \text{using (7.6.1)}$$

Hence the variance of v_i is now homoscedastic, and one may proceed to apply OLS to the transformed equation (6.2.3.10), regressing $\frac{Y_i}{X_i}$ on $\frac{1}{X_i}$. Notice that in the transformed regression the intercept term β_2 is the

slope coefficient in the original equation and the slope coefficient β_1 is the intercept term in the original model. Therefore, to get back to the original model we shall have to multiply the estimated (6.2.3.10) by X_i.

Assumption 2: The error variance is proportional to Xi. The **square root transformation:**

$$E(u_i^2) = \sigma^2 X_i \text{ ……………………………… 6.2.3.11}$$

If it is believed that the variance of u_i, instead of being proportional to the squared X_i, is proportional to X_i itself, then the original model can be transformed as follows:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$

$$= \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + v_i \text{ ….. 6.2.3.12}$$

Where $v_i = \frac{u_i}{\sqrt{X_i}}$ and where $X_i > 0$

Given assumption 2, one can readily verify that $E(v_i^2) = \sigma^2$, a homoscedastic situation. Therefore, one may proceed to apply OLS to (7.6.4), regressing $\frac{Y_i}{\sqrt{X_i}}$ on $\frac{1}{\sqrt{X_i}}$ and $\sqrt{X_i}$. Note an important feature of the transformed model: It has no intercept term. Therefore, one will have to use the regression-through-the-origin model to estimate β_1 and β_2. Having run (6.2.3.12), one can get back to the original model simply by multiplying (6.2.3.12) by $\sqrt{X_i}$.

Assumption 3: The error variance is proportional to the square of the mean value of Y.

$$E(u_i^2) = \sigma^2[E(Y_i)]^2 \text{ ………………………. 6.2.3.13}$$

Equation (6.2.3.13) postulates that the variance of u_i is proportional to the square of the expected value of Y. Now $E(Y_i) = \beta_1 + \beta_2 X_i$

Therefore, if we transform the original equation as follows,

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)}$$

$$= \beta_1 \frac{1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + v_i \text{ ………. 6.2.3.14}$$

where $v_i = \frac{u_i}{E(Y_i)}$, it can be seen that $E(v_i^2) = \sigma^2$; that is, the disturbances v_i are homoscedastic. Hence, it is regression (6.2.3.14) that will satisfy the homoscedasticity assumption of the classical linear regression model.

The transformation (6.2.3.14) is, however, in operational because $E(Y_i)$ depends on β1 and β2, which are unknown. Of course, we know $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, which is an estimator of $E(Y_i)$. Therefore, we may proceed in two steps: First, we run the usual OLS regression, disregarding the heteroscedasticity problem, and obtain $\hat{Y}_i$. Then, using the estimated $\hat{Y}_i$, we transform our model as follows:

$$\frac{Y_i}{\hat{Y}_i} = \beta_1 \left(\frac{1}{\hat{Y}_i}\right) + \beta_2 \left(\frac{X_i}{\hat{Y}_i}\right) + v_i \text{ .... } 6.2.3.15$$

Where $v_i = \frac{u_i}{\hat{Y}_i}$. In Step 2, we run the regression (6.2.3.15). Although $\hat{Y}_i$ are not exactly $E(Y_i)$, they are consistent estimators; that is, as the sample size increases indefinitely, they converge to true $E(Y_i)$. Hence, the transformation (6.2.3.15) will perform satisfactorily in practice if the sample size is reasonably large.

Assumption 4: A log transformation such as

$$lnY_i = \beta_1 + \beta_2 lnX_i + u_i \text{ .... } 6.2.3.16$$

7. NON LINEAR REGRESSION AND TIME SERIES ECONOMETRICS

7.1 Non Linear Regression Models: Overview

On previous chapters we have seen linear regression models, that is, models that are linear in the parameters and/or models that can be transformed so that they are linear in the parameters. On occasions, however, for theoretical or empirical reasons we have to consider models that are nonlinear in the parameters. In this chapter we take a look at such models and study their special features.

7.1.1 Intrinsically Linear and Intrinsically Nonlinear Regression Models

When we started our discussion of linear regression models on Chapter 3, we stated that our concern is basically with models that are linear in the parameters; they may or may not be linear in the variables. If a model is nonlinear in the parameters it is a nonlinear (in-the-parameter) regression model whether the variables of such a model are linear or not. However, one has to be careful here, for some models may look nonlinear in the parameters but are **inherently** or **intrinsically** linear because with suitable transformation they can be made linear-in-the-parameter regression models. But if such models cannot be linearized in the parameters, they are called **intrinsically nonlinear regression models.** *From now on when we talk about a nonlinear regression model, we mean that it is intrinsically nonlinear.* In short, we will call them **NLRM.**

To drive the distinction between the two, let us consider the following models.

Are the following models linear regression models? Why or why not?

a. $Yi = e^{\beta 1 + \beta 2Xi + ui}$

b. $\ln Yi = \beta_1 + \beta_2(\frac{1}{Xi}) + ui$

c. $Yi = \beta_1 + (0.75 - \beta_1)e^{-\beta 2(Xi-2)} + ui$

d. $Yi = \beta_1 + \beta_2^3 Xi + ui$

Models **c** and **d** are intrinsically nonlinear because there is no simple way to linearize them. Model **b** is obviously a linear regression model. What about Models **a**? Taking the logarithms on both sides of **a,** we obtain $\ln Yi = \beta_1 + \beta_2 Xi + ui$, which is linear in the parameters. Hence Model **a** is *intrinsically* a linear regression model.

Estimation Of Nonlinear Regression Models

To see the difference in estimating linear and nonlinear regression models, consider the following two models:

$Yi = \beta_1 + \beta_2X_i + ui$ ---7.1.1

Yi = $\beta_1 e^{\beta 2Xi}$+ ui---7.1.2

By now you know that (7.1.1) is a linear regression model, whereas (7.1.2) is a nonlinear regression model. Regression (7.1.2) is known as the **exponential regression model** and is often used to measure the growth of a variable, such as population, GDP, or money supply. Suppose we consider estimating the parameters of the two models by OLS. In OLS we minimize the residual sum of squares (RSS), which for model (7.1.1) you should remember from what you have learned in previous chapters. Observe very carefully by remembering that in these equations the unknowns (β's) are on the left-hand side and the knowns (*X* and *Y*) are on the right-hand side. As a result we get explicit solutions of the two unknowns in terms of our data.

Now see what happens if we try to minimize the RSS of (7.1.2).

$\sum Yie^{\beta 2Xi}$= $\beta_1 e^{2\beta 2Xi}$---7.1.3

$\sum YiXie^{\beta 2Xi}$= β1 ∑ Xi $e^{2\beta 2Xi}$---7.1.4

Unlike the normal equations in the case of the linear regression model, the normal equations for nonlinear regression have the unknowns (the βiˆ's) both on the left- and right-hand sides of the equations. As a consequence, we *cannot obtain explicit solutions* of the unknowns in terms of the known quantities. To put it differently, the unknowns are expressed in terms of themselves and the data! Therefore, although we can apply the method of least squares to estimate the parameters of the nonlinear regression models, we cannot obtain explicit solutions of the unknowns. Incidentally, OLS applied to a nonlinear regression model is called **nonlinear least squares (NLLS).** So, what is the solution?

ESTIMATING NONLINEAR REGRESSION MODELS: THE TRIAL-AND-ERROR METHOD

To set the stage, let us consider a concrete example. The data in Table 7.1relates to the management fees that a leading mutual fund in the United States pays to its investment advisors to manage its assets. The fees paid depend on the net asset value of the fund. As you can see, the higher the net asset value of the fund, the lower are the advisory fees, which can be seen clearly from Figure 7.1.

To see how the exponential regression model in (7.1.2) fits the data given in Table 7.1, we can proceed by trial and error.

TABLE 7.1 ADVISORY FEES CHARGED AND ASSET SIZE

	Fee,	**% Asset***
1	0.520	0.5 '
2	0.508	5.0 .
3	0.484	10 '
4	0.46	15 '
5	0.4398	20 '
6	0.4238	25 .
7	0.4115	30 .
8	0.402	35 .
9	0.3944	40 .
10	0.388	45 .
11	0.3825	55 .
12	0.3738	60 .

*Asset represents net asset value, billions of dollars.

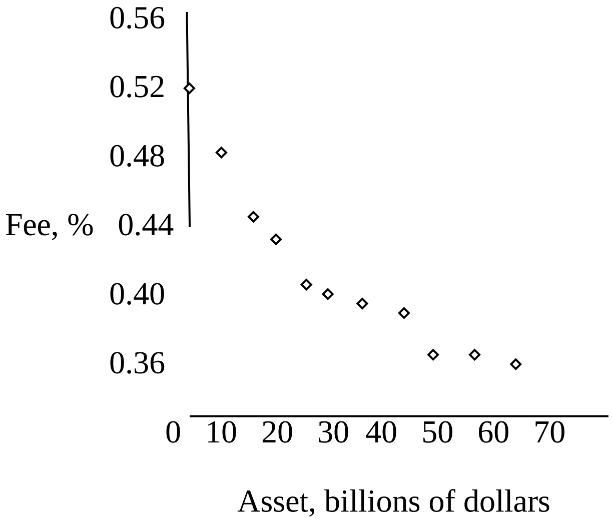


FIGURE 7.1 Relationship of advisory fees to fund assets.

Suppose we assume that initially β1 = 0.45 and β2 = 0.01. These are pure guesses, sometimes based on prior experience or prior empirical work or obtained by just fitting a linear regression model even though it may not be appropriate. At this stage do not worry about how these values are obtained.

Since we know the values of β1 and β2, we can write (7.1.2) as:

$ui = Yi - \beta_1 e^{\beta 2Xi} = Yi - 0.45e^{0.01Xi}$

Therefore,

$\sum\ u_i^2 = \sum\ (Yi - 0.45e^{0.01Xi})^2$--7.1.5

Since *Y*, *X*, β1, and β2 are known, we can easily find the *error sum of squares* in (7.1.5). Remember that in OLS our objective is to find those values of the unknown parameters that will make the error sum of squares as small as possible. This will happen if the estimated *Y* values from the model are as close as possible to the actual *Y* values. With the given values, we obtain

u_i^2= 0.3044. But how do we know that this is the least possible error sum of squares that we can obtain? What happens if you choose another value for β1 and β2, say, 0.50 and −0.01, respectively? Repeating the procedure just laid down, we find that we now obtain & u_i^2 = 0.0073. Obviously, this error sum of squares is much smaller than the one obtained before, namely, 0.3044. But how do we know that we have reached the lowest possible error sum of squares, for by choosing yet another set of values for the β's, we will obtain yet another error sum of squares?

As you can see, such a trial-and-error, or **iterative,** process can be easily implemented. And if one has infinite time and infinite patience, the trial and-error process *may* ultimately produce values of β1 and β2 that may guarantee the lowest possible error sum of squares. But you might ask, how did we go from (β1 = 0.45; β2 = 0.01) to (β1 = 0.50; β2 = −0.1)? Clearly, we need some kind of *algorithm* that will tell us how we go from one set of values of the unknowns to another set before we stop. Fortunately such algorithms are available, and they will be discussed in the next section.

Algorithm Approaches To Estimating Nonlinear Regression Models

There are several approaches, or algorithms, to estimate NLRMs: (1) direct search or trial and error, (2) direct optimization, and (3) iterative linearization.

✓ Direct Search or Trial-and-Error or Derivative-Free Method

In the previous section we showed how this method works. Although intuitively appealing because it does not require the use of calculus methods as the other methods do, this method is generally not used. *First,* if an NLRM involves several parameters, the method becomes very cumbersome and computationally expensive. For example, if an NLRM involves 5 parameters and 25 alternative values for each parameter are considered, you will have to compute the error sum of squares $(25)^5$ = 9,765,625 times! *Second*, there is no guarantee that the final set of parameter values you have selected will necessarily give you the absolute minimum error sum of squares. In the language of calculus, you may obtain a local and not an absolute minimum. In

fact, **no method** guarantees a global minimum.

- ✓ **Direct Optimization**

In direct optimization we differentiate the error sum of squares with respect to each unknown coefficient, or parameter, set the resulting equation to zero, and solve the resulting normal equations simultaneously. We have already seen this in Eqs. (7.1.3) and (7.1.4). But as you can see from these equations, they cannot be solved explicitly or *analytically.* Some iterative routine is therefore called for. One routine is called the **method of steepest descent.** We will not discuss the technical details of this method as they are somewhat involved, but the reader can find the details in the references.

Like the method of trial and error, the method of steepest descent also involves selecting initial trial values of the unknown parameters but then it proceeds more systematically than the hit-or-miss or trial-and-error method. One disadvantage of this method is that it may converge to the final values of the parameters extremely slowly.

- ✓ **Iterative Linearization Method**

In this method we linearize a nonlinear equation around some initial values of the parameters. The linearized equation is then estimated by OLS and the initially chosen values are adjusted. These adjusted values are used to *relinearize* the model, and again we estimate it by OLS and readjust the estimated values. This process is continued until there is no substantial change in the estimated values from the last couple of iterations. The main technique used in linearizing a nonlinear equation is the **Taylor series expansion** from calculus. Rudimentary details of this method are given on (**Gujarati: Basic Econometrics, Fourth Edition for interested reader**) in Appendix 14A, Section 14A.2. Estimating NLRM using Taylor series expansion is systematized in two algorithms, known as the **Gauss–Newton iterative method** and the **Newton–Raphson iterative method.** Since one or both of these methods are now incorporated in several computer packages, and since a discussion of their technical details will take us far beyond the scope of the course at this level, there is no need to dwell on them here.

7.2. Time Series Analysis

7.2.1. Meaning and Components of Time series Analysis

A time series is a set of observations on the value that a variable takes at different times. A time series data is a set of observations taken at specified times, usually, at "equal intervals".

The time series elements are classified into four basic types of variations, which account for the changes in the series over a period of time. These four types of patterns, variations, movements are often called the component or elements of time series. These are:

1) Secular Trend

2) Seasonal Trend

3) Cyclical Variations

4) Irregular Variations

In traditional or classical time series analysis, it is ordinarily assumed that there is a multiplicative relationship between these four components. That is, it is assumed that any particular value in series is the product of factors that can be attributed to the various components. Symbolically, it is given as:

$$Y = T \text{ x } S \text{ x } C \text{ x } I$$

Where T = trend

C = cyclical

S = seasonal

I = irregular

If the above model is employed, the seasonal, cyclical and irregular items are not viewed as absolute amounts, but rather as relative magnitude.

7.2.1.1 Secular Trend

Trend is the variation of the value of a variable that can be observed in a long period of time. It is the general tendency of the data to grow or to decline over a long period of time. Trend is broadly divided under two heads: **linear** and **nonlinear** trends.

The methods of Measuring Trend

The following three methods are used for measuring trend:

- Graphic method
- The semi - average method
- The method of Least Squares

A. Graphic method

This is the simplest method of studying trend. Under this method the given data are plotted on graph paper and a trend line is fitted to the data just by inspecting the graph of the series. There is no formal statistical criterion where the adequacy of such a line can be judged and the judgment depends on the discretion of the individual researcher. This method is not frequently used since its approach is not exact.

B. Methods of Semi- Averages

This method is used in such a way that the given data are divided into two parts preferably, with equal number of years. For example, you are given data from 1982 to 1999, that is over a period of 18 years, the two equal parts will be first nine years, i.e. from 1982 to1990 and from 1991 to 1999. In the case of total number of years not divisible by 2 such as 9, 11,13 etc, two equal parts can be made simply by ignoring the middle year.

Example: Fit a trend line to the following data by the method of semi-averages

Year	Sales
1994	102
1995	105
1996	114
1997	110
1998	108
1999	116
2000	112

Since seven years are given, the middle year should be omitted and an average of the first three years and the last three years shall be obtained. The averages of the first three years is
(102+105+114)/3 = 107 and the average o the last three years is
(108+116+112)/3 = 112

Thus, you get two points 107 and 112, which shall be plotted corresponding to their respective middle years, i.e. 1995 and 1999. By joining these two points; you shall obtain the required trend line.

C. Method of Least Squares

This method is most widely used in practice. When this method is applied, a trend line is fitted to the data in such a way that the following two conditions are satisfied:

1) $\sum(\mathrm{Y}-\mathrm{Yc})=0$. The sum of deviations of the actual values of Y and the computed values of Y is zero
2) $\sum(\mathrm{Y}-\mathrm{Yc})^2$ is the least, that is the sum of the squares of the deviations of the actual and computed values is the least one. The method of least squares can be used either to fit a straight line trend or a parabolic trend. The straight line trend is represented by the equation.

$$\mathrm{Yc} = \mathrm{a}+\mathrm{bX}$$

In order to determine the value of the constants a and b, the following equations are to be solved

$\sum \mathrm{Y} = \mathrm{na} + \mathrm{b}\sum \mathrm{X}^{\cdot}$

$\sum \mathrm{YX} = \mathrm{a}\sum \mathrm{X} + \mathrm{b}\sum \mathrm{X}^2$ where, n represents number of years and X is the time period.

You can measure the variable X from any point of time in origin such as the first year. However this calculations are very much simplified when the midpoint in time is taken as the positive values in the second half so that $\sum \mathrm{X} = 0$, the above two normal equations would take the form:

$\sum \mathrm{Y} = \mathrm{an}^{\cdot}$

$\sum \mathrm{XY} = \mathrm{b}\sum \mathrm{X}^2$, $\quad$ $\mathrm{b} = \sum(\mathrm{X\,Y})/(\sum \mathrm{X}^2)$ $\quad$ $\mathrm{a} = (\sum \mathrm{Y})/\mathrm{n}$

The constant 'a' give the arithmetic mean of Y and the constant 'b' indicates the rate of change.

Example1: The following is production of a sugar factory in thousand quintals

Year	1983	1984	1985	1986	1987	1988	1989
Production	80	90	92	83	94	99	92

Required

1. Fit a straight line trend
2. Estimate the likely sales of the company in 1990

Solution

Year	**Production(Y)**	**Time(X)**	**XY**	**X^2**
1983	80	-3	-240	9
1984	90	-2	-180	4
1985	92	-1	-92	1
1986	83	0	0	0
1987	94	1	94	1
1988	99	2	198	4
1989	92	3	276	9 .
Total	**630**	**0**	**56**	**28**

1) $Yc = a+bX$

$a = (\sum Y)/n = 630/7 = 90$

$b = \sum (XY)/(\sum X^2) = 56/28 = 2$

$Yc = a+bX = 90 + 2X$

2) Forecasting for **1990.** Since 1990 is four years later than the base year, X =4. Therefore it is possible to find the value of the Yc when X = 4, Yc = 90 + 2(4) = 98 units

Exercise Calculate the trend values by the method of least squares from the data given below and fit a straight trend line and estimate the sale for the year 2003.

Year	1996	1997	1998	1999	2000
Sales	12	18	20	23	27

7.2.1.2 Seasonal Variations

Seasonal variations are periodic movements in business activity, which occur regularly every year and have their origin in the nature of the year itself. Seasonal variation exists only when data are given in a period which is less than a year(monthly, weekly, semi-annually, daily etc). However, it does not exist in the data that are given in annual basis or more than a year period interval. Nearly every type of business activity is liable to a seasonal influence to a greater or lesser degree and as such, these variations are regarded as normal phenomenon recurring every year. Although the word 'seasonal' seems to imply a connection with the season of the year, the term is meant to include any kind of variation, which is of periodic nature and whose repeating cycle are of relatively short duration. The factors that cause seasonal variations are climate and weather conditions, Customs, Traditions and Habits

Methods of Measuring Seasonal Variations

When data are expressed annually, there is no seasonal variation. However, monthly or quarterly data frequently exhibit strong seasonal movements and considerable interest attaches to devise a pattern of average seasonal variation. There are several methods of measuring seasonal variations. However, the following methods are popularly used in practice:

- Method of simple averages
- Ratio to trend method

- Ration to moving average method
- Link relatives method

A. Method of simple averages

This is the simplest method of obtaining a seasonal index. The following steps are necessary for computing the index:

- Average the unadjusted data by years and months or quarters if the data are given quarterly.
- Find the totals of the data in each month, quarter or a period in which the data are given.
- Divide each total by the number of years for which data are given.
- Obtain an average of monthly averages by dividing the total of monthly averages by 12.

Taking the average of monthly averages as 100, compute the percentage.

$$\text{Seasonal Index for January} = \frac{\text{Monthly Average for January}}{\text{Average of monthly averages}} \text{ X } 100$$

Example: Consumption of monthly electric power in KW hours of for street lighting in a given company from 1995-1999 is given in the following table.

Year	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
1995	318	281	278	250	231	216	223	245	269	302	325	347
1996	342	309	299	268	249	236	242	262	288	321	342	364
1997	367	328	320	287	269	251	259	284	309	345	367	394
1998	392	349	342	311	290	273	282	305	328	364	389	417
1999	420	378	370	334	314	296	305	330	356	396	422	425

Find out seasonal variation by the method of monthly averages.

Solution;

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Total**	**Aver.**
1995	318	281	278	250	231	216	223	245	269	302	325	347		
1996	342	309	299	268	249	236	242	262	288	321	342	364		
1997	367	328	320	287	269	251	259	284	309	345	367	394		
1998	392	349	342	311	290	273	282	305	328	364	389	417		
1999	420	378	370	334	314	296	305	330	356	396	422	425		
Total	**1839**	**1645**	**1609**	**1450**	**1353**	**1272**	**1311**	**1426**	**1550**	**1728**	**1845**	**1947**	**18975**	**1581.25**
Average	**367.8**	**329**	**321.8**	**290**	**270.6**	**254.4**	**262.2**	**285.2**	**310**	**345.6**	**369**	**394.8**	**3800.4**	**316.7**
%	**116.1**	**103.9**	**101.6**	**91.6**	**85.4**	**80.3**	**82.8**	**90.1**	**97.9**	**109.1**	**116.5**	**124.7**	**1200**	**100**

seasonal index for January =(367.8/316.7)x100= 116.1
seasonal index for February =(329/316.7)x100= 103.9
seasonal index for July =(262.2/316.7)x100= 82.8

B. Ratio-to-Trend Method

This method of calculating a seasonal index is relatively simple and yet an improvement over the method of simple average explained in the preceding section. The method assume that the seasonal variation for a given month is a constant fraction of the trend. First T eliminates the trend component by dividing the original.

$$\frac{T*S*C*I}{T} = S*C*I$$

The random elements are supposed to disappear when the ratio are averaged. A careful selection of a period of years used in the computation is expected to cause the influences of prosperity or depression to offset each other and thus remove the cycle. This method requires the following steps.

Step1: Compute the trend values by applying the method of least squares.

Step2: Divide the original data month by month by the corresponding trend values and multiply the ratio by 100. The values obtained are now free from trend.

Step3: In order to free from irregular and cyclical movements, the irregular given for various years for the months(January, February, etc) should be averaged; and

Step4: The seasonal index for each month is expressed as a percentage of the average month. The sum of 12 values must equal 1,200 or 100%. If it does not, an adjustment is made by multiplying each index by a suitable factor(1200). This gives the final seasonal index.

Example: Find the seasonal variations by ration to trend method from the data given below

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1996	30	40	36	34
1997	34	52	50	44
1998	40	58	54	48
1999	74	76	68	42
2000	80	92	86	82

Solution

To determine seasonal variation by ration to trend method, first you will determine the trend of yearly data and then convert it to quarterly data. First, calculate the trend values.

Year	Yearly total	Yearly average	Time(X)	XY	X^2	Trend values
1996	140	35	-2	-70	4	32
1997	180	45	-1	-45	1	44
1998	200	50	0	0	0	56
1999	260	65	1	65	1	68
2000	340	85	2	170	4	80
Total		**280**	**0**	**120**	**10**	

Yc = a+bX

a = $(\sum Y)/n$ = 280/5 = 56

b = $\sum (XY)/(\sum X^2)$ = 120/10 = 12

Quarterly increment = 12/4 = 3

Calculation of quarterly trend values: Consider 1997. The trend value of 1997 indicates the trend value of the middle quarter of the year. The middle quarter is found half of the second and half of the 3rd quarter. Therefore the trend value of the 2nd quarter is given as 44- 3/2 = 42.5 and the trend value of the 3rd quarter is 44+3/2 = 45.5. After this subtract 3 from the 2nd quarter trend value to get the trend value of the first quarter and add three to get the trend value of the 4th quarter to the trend value of the 3rd quarter. The trend values for each quarter are given in the following table.

Trend values

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1996	27.5	30.5	33.5	36.5
1997	39.5	42.5	45.5	48.5
1998	51.5	54.5	57.5	60.5
1999	63.5	66.5	69.5	72.5
2000	75.5	78.5	81.5	84.5

The ration to trend values can be found by dividing the original data by the trend values expressed in percentage.

Quarterly values as percentage of trend values

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1996	109.1	131.1	107.5	93.1
1997	86.1	122.4	109.9	90.7
1998	77.7	106.4	93.9	79.3
1999	85	114.3	97.8	85.5
2000	106	117.1	105.5	84.5
Total	**463.9**	**591.3**	**514.6**	**445.6**
Average	**92.78**	**118.26**	**102.92**	**89.12**

Since 92.78 + 118.26 + 102.92 + 89.12 =403.08 is greater than 400, you have to find the correction factor and multiply each seasonal index by the correction factor.

CF = $\frac{400}{sum\ of\ the\ values\ of\ the\ 4\ quarters} = \frac{400}{403.08}$, then the adjusted seasonal index will be given as follows:

1st quarter = 92
2nd quarter = 117.4
3rd quarter = 102.2
4th quarter = 88.4

C. Ratio- to- moving average method

The ratio to the moving average is the most widely used method of measuring seasonal variations. The following steps are important in measuring seasonal variations using the ration to moving average method:

Step1: Compute the centered 12 month moving average from the original data. This contains trend and cyclical variations.

Step2: Express the original data for each month as percentage of the centered 12 month moving average.

Step3: Divide each month data by the corresponding centered 12 month moving average and list the quotient.

$$\frac{T*S*C*I}{T*S} = S*I$$

Step4: Compute the average of each month for the quotient that we obtained in step 3. By doing so the irregular component will be removed. $\frac{S*I}{I} = S$

The sum of seasonal index should be 1200. If the sum is different from 1200. If the sum is different from 1200, compute the correction factor and multiply each month's seasonal index by the correction factor. The correction factor is obtained as,

$$CF = \frac{1200}{The\ total\ mean\ for\ 12\ months}$$

D. Link Relatives Method

This method involves the following steps.

Step1: Calculate the link relatives of the seasonal figures

$$LR = \frac{Current\ season's\ figure}{Previous\ season's\ figure} \text{ x } 100$$

Step2: Calculate the average of the link relatives for each season

Step3: Convert the averages into chain relatives on the base of the last season

Step4: Calculate the chain relatives of the first season on the base of the last season

Step5: For correction, chain relatives of the first season calculated by the first method is deducted from the chain relative of the first season calculated by the second method.

Step6: Express corrected chain relatives as percentage of their averages. These Provide the required seasonal indices by the method of link relatives.

7.2.3 Cyclical Variations

The term cycle refers to recurrent variations in time series that usually last longer than a year and regular, neither in amplitude nor in length. Cyclical fluctuations are long term movements that represent consistently recurring rises and declines in activity. They are resulted mainly from business cycles. A business cycle consists of the up and down movements of business activity from some sort of statistical trend. There are four well defined periods or phases in the business cycle. These are prosperity, decline, depression and improvement. The study of cyclical variations is extremely useful in framing suitable policies for stabilizing the level of business activity, i.e. for avoiding the periods of booms and depressions as both are bad for the economy.

7.2.3.1 Measurement of Cyclical Variations

Despite their importance, business cycles are most difficult types of fluctuations to measure. This is because successive cycles vary widely in timing, amplitude and pattern. Because of such reason it is impossible to construct meaningful typical cycle indices of curves similar to those that have been developed for trends and seasonality. The important methods used to measure cyclical variations are:

1. Residual Method
2. Reference Cycle Analysis Method
3. Direct Method
4. Harmonic Analysis Method

Among the above methods the one that is frequently used and convenient is the first method. Therefore only that method will be discussed here.

A. Residual Method

This method is most commonly used method. It consists of eliminating seasonal and then trend variations to obtain the cyclical and irregular movements.

$$\frac{T*S*C*I}{S} = \mathrm{T} * \mathrm{C} * \mathrm{I}$$

$$\frac{T*C*I}{T} = \mathrm{C} * \mathrm{I}$$

The data are usually smoothed in order to obtain cyclical movements, which are sometimes termed as the cyclical relatives since they are always expressed in percentages. This is because cyclical, irregular or the cyclical movements remain residuals. As a result, this procedure is referred to as the **residual method**

7.2.4 Irregular Variations

Irregular variations refers to such variations in business activities which do not repeat in a definite pattern. It includes all types of variations other than those accounting for the trend, seasonal and cyclical movements. Irregular movements are considered to be largely random, being the result of chance factors, which like the fall of a coin, are wholly unpredictable.

Irregular variations are caused by such special occurrences as flood, earthquakes, strikes and wars. Sudden changes in demand or rapid technological progress may also be included in this category. By their nature, these movements are irregular and unpredictable. Quantitatively it is almost impossible to separate-out the irregular movements and the cyclical movements. Therefore while analyzing time series, the trend and seasonal variations are measured separately and cyclical and irregular variations are left altogether.

7.2.4.1 Measurement of Irregular Variations

The irregular component in the time series represents the residue of fluctuations after trend, seasonal and cyclical movements have been accounted for. Thus if the original data is divided by T, S, and C you will get I.   TSCI/TSC = I

In practice the cycle itself is erratic and interwoven with irregular movements that it is impossible to separate them. in the analysis of time series into its components, trend and seasonal movements are usually measured directly, while cyclical and irregular fluctuations are left altogether after the other elements have been removed.

Good luck-----------