

CYNTHIA FRASER

# Business Statistics for Competitive Advantage with Excel 2007

Basics, Model Building, and Cases



Springer

# Business Statistics for Competitive Advantage with Excel 2007

# Business Statistics for Competitive Advantage with Excel 2007

Basics, Model Building,  
and Cases

Cynthia Fraser

University of Virginia, McIntire School of Commerce

Cynthia Fraser  
University of Virginia  
Charlottesville, VA, USA

ISBN: 978-0-387-74402-4      e-ISBN: 978-0-387-74403-2  
DOI: 10.1007/978-0-387-74403-2

Library of Congress Control Number: 2008939440

© Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

springer.com

*To Len Lodish, who introduced me to the competitive advantages of modeling.*

# Contents

Preface		xvii
<b>Chapter 1</b>	<b>Statistics for Decision Making and Competitive Advantage</b>	<b>1</b>
1.1	Statistical Competences Translate Into Competitive Advantages	1
1.2	Attain Statistical Competences And Competitive Advantage With This Text	1
1.3	Follow The Path Toward Statistical Competence and Competitive Advantage	2
1.4	Use Excel for Competitive Advantage	3
1.5	Statistical Competence Is Satisfying	3
<b>Chapter 2</b>	<b>Describing Your Data</b>	<b>5</b>
2.1	Describe Data With Summary Statistics And Histograms	5
	<i>Example 2.1 Yankees' Salaries: Is it a Winning Offer?</i>	5
2.2	Outliers Can Distort The Picture	7
	<i>Example 2.2 Executive Compensation: Is the Board's Offer on Target?</i>	7
2.3	Round Descriptive Statistics	10
2.4	Central Tendency and Dispersion Describe Data	11
2.5	Data Is Measured With Quantitative or Categorical Scales	11
2.6	Continuous Data Tend To Be Normal	12
	<i>Example 2.3 Normal SAT Scores</i>	12
2.7	The Empirical Rule Simplifies Description	13
	<i>Example 2.4 Class of '06 SATs: This Class is Normal &amp; Exceptional</i>	13
2.8	Describe Categorical Variables Graphically: Column and PivotCharts	15
	<i>Example 2.5 Who Is Honest &amp; Ethical?</i>	15
2.9	Descriptive Statistics Depend On The Data	16
Excel 2.1	Produce descriptive statistics and view distributions with histograms	17
Excel 2.2	Sort to produce descriptives without outliers	20
Excel 2.3	Plot a cumulative distribution	23

Excel 2.4	Find and view distribution percentages with a PivotTable and PivotChart	24
Excel 2.5	Produce a column chart from a PivotChart of a nominal variable	27
	<i>Excel Shortcuts at Your Fingertips</i>	29
	<i>Lab 2 Descriptive Statistics</i>	31
	<i>Assignment 2-1 Procter &amp; Gamble's Global Advertising</i>	33
	<i>CASE 2-1 VW Backgrounds</i>	34
<b>Chapter 3</b>	<b>Hypothesis Tests, Confidence Intervals and Simulation to Infer Population Characteristics and Differences</b>	<b>35</b>
3.1	Sample Means Are Random Variables	35
	<i>Example 3.1 Thirsty on Campus: Is there Sufficient Demand?</i>	35
3.2	Use Sample Data to Determine Whether Or Not $\mu$ Is Likely To Exceed A Target	38
3.3	Confidence Intervals Estimate the Population Mean From A Sample	41
3.4	Round $t$ to Calculate Approximate 95% Confidence Intervals With Mental Math	43
3.5	Margin of Error Is Inversely Proportional To Sample Size	43
3.6	Samples Are Efficient	44
3.7	Use Monte Carlo Simulation with Sample Statistics To Incorporate Uncertainty and Quantify Implications Of Assumptions	44
3.8	Determine Whether There Is a Difference Between Two Segments With Student $t$	48
	<i>Example 3.2 Pampers Preemies: Is Income a Useful Base for Segmentation?</i>	48
3.9	Estimate the Extent of Difference between Two Segments With Student $t$	49
3.10	Confidence Intervals Complement Hypothesis Tests	50
3.11	Estimation of a Population Proportion from a Sample Proportion	50
	<i>Example 3.3 Guinea Pigs</i>	50
3.12	Conditions for Assuming Approximate Normality to Make Confidence Intervals for Proportions	53
3.13	Conservative Confidence Intervals for a Proportion	53
3.14	Assess the Difference between Alternate Scenarios or Pairs With Student $t$	54
	<i>Example 3.4 Are "Socially Desirable" Portfolios Undesirable?</i>	55
3.15	Inference from Sample to Population	58
Excel 3.1	Test the level of a population mean with a one sample $t$ test	59
Excel 3.2	Make a confidence interval for a population mean	60

Excel 3.3	Illustrate population confidence intervals with a clustered column chart	61
Excel 3.4	Conduct a Monte Carlo simulation with Crystal Ball	65
Excel 3.5	Test the difference between two segments with a two sample <i>t test</i>	69
Excel 3.6	Construct a confidence interval for the difference between two segments	70
Excel 3.7	Illustrate the difference between two segment means with a column chart	71
Excel 3.8	Construct a pie chart of shares	72
Excel 3.9	Test the difference in levels between alternate scenarios or pairs with a paired <i>t test</i>	74
Excel 3.10	Construct a confidence interval for the difference between alternate scenarios or pairs	76
	<i>Excel Shortcuts at Your Fingertips</i>	78
	<i>Lab Practice 3 Inference</i>	80
	<i>Lab 3 Inference</i>	82
	<i>Assignment 3-1 Bottled Water Possibilities</i>	83
	<i>Assignment 3-2 Immigration in the U.S.</i>	84
	<i>Assignment 3-3 McLattes</i>	84
	<i>Assignment 3-4 A Barbie Duff in Stuff</i>	85
	<i>CASE 3-1 Yankees v Marlins: The Value of a Yankee Uniform</i>	85
	<i>CASE 3-2 Gender Pay</i>	86
	<i>CASE 3-3 Polaski Vodka: Can a Polish Vodka Stand Up to the Russians?</i>	86
	<i>CASE 3-4 American Girl in Starbucks</i>	88
<b>Chapter 4</b>	<b>Quantifying the Influence of Performance Drivers and Forecasting: Regression</b>	<b>91</b>
4.1	The Simple Linear Regression Equation Describes the Line Relating A Decision Variable to Performance	91
	<i>Example 4.1 HitFlix Movie Rentals</i>	92
4.2	<i>F</i> Tests the Significance of the Hypothesized Linear Relationship, <i>RSquare</i> Summarizes Its Strength and Standard Error Reflects Forecasting Precision	93
4.3	The Population Slope Is Tested And Inferred From Our Sample	96
4.4	Analyze Residuals To Learn Whether Assumptions Have Been Met	98
4.5	95% Prediction Intervals Acknowledge That Individual Elements Differ	99
4.6	Use Sensitivity Analysis to Explore Alternative Scenarios	101

4.7	95% Conditional Mean Prediction Intervals Of Average Performance Gauge Average Performance Response To A Driver	101
4.8	Explanation And Prediction Create A Complete Picture	102
4.9	Present Regression Results In Concise Format	103
4.10	We Make Assumptions When We Use Linear Regression	104
4.11	Correlation Is A Standardized Covariance	105
	<i>Example 4.2 HitFlix Movie Rentals</i>	105
4.12	Correlation Coefficients Are Key Components Of Regression Slopes	109
	<i>Example 4.3 Pampers</i>	110
4.13	Correlation Summarizes Linear Association	113
4.14	Linear Regression Is Doubly Useful	113
Excel 4.1	Fit a simple linear regression model	114
Excel 4.2	Construct prediction and conditional mean prediction intervals	118
Excel 4.3	Find correlations between variable pairs	124
	<i>Excel Shortcuts at Your Fingertips</i>	126
	<i>Lab 4 Regression</i>	128
	<i>CASE 4-1 GenderPay (B)</i>	130
	<i>CASE 4-2 GM Revenue Forecast</i>	131
	<i>Assignment 4-1 Impact of Defense Spending on Economic Growth</i>	133
<b>Chapter 5</b>	<b>Marketing Segmentation with Descriptive Statistics, Inference, Hypothesis Tests and Regression</b>	<b>135</b>
	<i>CASE 5-1 Segmentation of the Market for Premie Diapers</i>	135
5.1	Guide to Effective PowerPoint Presentations and Writing Memos that your Audience will Read	145
5.2	Write Memos that Encourage Your Audience to Read and Use Results	147
	MEMO Re: Importance of Fit Drives Trial Intention	148
<b>Chapter 6</b>	<b>Finance Application: Portfolio Analysis with a Market Index as a Leading Indicator in Simple Linear Regression</b>	<b>149</b>
6.1	Rates of Return Reflect Expected Growth of Stock Prices	149
	<i>Example 6.1 Goldman Sachs and Yahoo Returns</i>	149
6.2	Investors Trade Off Risk And Return	152
6.3	Beta Measures Risk	152
	<i>Example 6.2 Four diverse stocks</i>	153

6.4	A Portfolio’s Expected Return, Risk and Beta Are Weighted Averages of Individual Stocks	158
	<i>Example 6.3 Four Alternate Portfolios</i>	158
6.5	Better Portfolios Define The Efficient Frontier	161
	MEMO Re: Recommended Portfolios Include Lockheed Martin and Apple	162
6.6	Portfolio Risk Depends On the Covariances between Individual Stocks’ Rates of Return and The Market Rate Of Return	163
Excel 6.1	Estimate portfolio expected rate of return and risk	164
Excel 6.2	Plot return by risk to identify dominant portfolios and the Efficient Frontier	166
	<i>Assignment 6-1 Individual Stocks’ Beta Estimates</i>	169
	<i>Assignment 6-2 Expected Returns and Beta Estimates of Alternate Portfolios</i>	169
	<i>Assignment 6-3 Portfolio Comparison</i>	170
<b>Chapter 7</b>	<b>Association between Two Categorical Variables: Contingency Analysis with Chi Square</b>	<b>171</b>
7.1	When Conditional Probabilities Differ From Joint Probabilities, There Is Evidence of Association	171
	<i>Example 7.1 Recruiting Stars</i>	172
7.2	Chi Square Tests Association between Two Categorical Variables	174
7.3	Chi Square Is Unreliable If Cell Counts Are Sparse	175
7.4	Simpson’s Paradox Can Mislead	177
	<i>Example 7.2 American Cars</i>	177
	MEMO Re: Country of Manufacture Does Not Affect Older Buyers’ Choices	183
7.5	Contingency Analysis Is Demanding	184
7.6	Contingency Analysis Is Quick, Easy, and Readily Understood	184
Excel 7.1	Construct crosstabulations and assess association between categorical variables with PivotTables and PivotCharts	185
Excel 7.2	Use chi square to test association	187
Excel 7.3	Conduct contingency analysis with summary data	190
	<i>Excel Shortcuts at Your Fingertips</i>	193
	<i>Assignment 7-1 747s and Jets</i>	195
	<i>Assignment 7-2 Fit Matters</i>	195
	<i>Assignment 7-3 Allied Airlines</i>	196
	<i>CASE 7-1 Hybrids for American Car</i>	197
	<i>CASE 7-2 Tony’s GREAT Advertising</i>	198

<b>Chapter 8</b>	<b>Building Multiple Regression Models</b>	<b>201</b>
8.1	Multiple Regression Models Identify Drivers and Forecast	201
8.2	Use Your Logic to Choose Model Components	201
	<i>Example 8.1 Sakura Motors Quest for Cleaner Cars</i>	202
8.3	Multicollinear Variables Are Likely When Few Variable Combinations Are Popular In a Sample	203
8.4	<i>F</i> Tests the Joint Significance of the Set of Independent Variables	204
8.5	Insignificant Parameter Estimates Signal Multicollinearity	205
8.6	Combine or Eliminate Collinear Predictors	205
8.7	<i>Partial F</i> Tests the Significance of Changes in Model Power	207
8.8	Sensitivity Analysis Quantifies the Marginal Impact Of Drivers	211
	MEMO Re: Light, responsive, fuel efficient cars with smaller engines are cleanest	214
8.9	Model Building Begins With Logic and Considers Multicollinearity	215
Excel 8.1	Build and fit a multiple linear regression model	216
Excel 8.2	Use sensitivity analysis to compare the marginal impacts of drivers	221
	<i>Lab Practice 8</i>	228
	<i>Lab 8 Model Building with Multiple Regression</i>	230
	<i>Assignment 8-1</i>	233
<b>Chapter 9</b>	<b>Model Building and Forecasting with Multicollinear Time Series</b>	<b>235</b>
9.1	Time Series Models Include Decision Variables, External Forces, Leading Indicators, And Inertia	237
	<i>Example 9.1 Home Depot Revenues</i>	238
9.2	Indicators of Economic Prosperity Lead Business Performance	238
9.3	Inertia from Loyal Customers Drives Performance	238
9.4	Compare Scatterplots across Time to Choose Length of Lags For Drivers of Delayed Response: Visual Inspection	239
9.5	Hide the Two Most Recent Datapoints to Validate a Time Series Model	241
9.6	Correlations Guide Choice of Lags	241
9.7	The Durbin Watson Statistics Identifies Autocorrelation	242
9.8	Assess Residuals to Identify Unaccounted For Trend or Cycles	243
9.9	Forecast the Recent, Hidden Points to Assess Predictive Validity	246

9.10	Add the Most Recent Datapoints to Recalibrate	246
	MEMO Re: Revenue Decline Forecast Following New Home Sales Downturn	248
9.11	Inertia and Leading Indicator Components Are Powerful Drivers and Often Multicollinear	249
Excel 9.1	Build and fit a multiple regression model with multicollinear time series	250
	<i>Chapter 9 Lab: HP Revenue Forecast</i>	266
	<i>CASE 9-1 Dell: Overcoming Roadblocks to Growth</i>	268
	<i>CASE 9-2 Mattel Revenues Following the Recalls</i>	270
	<i>CASE 9-3 Starbucks in China</i>	272
<b>Chapter 10 Indicator Variables</b>		<b>275</b>
10.1	Indicators Modify the Intercept to Account for Segment Differences	275
	<i>Example 10.1 Hybrid Fuel Economy</i>	275
	<i>Example 10.2 Yankees v Marlins Salaries</i>	276
10.2	Indicators Estimate the Value of Product Attributes	278
	<i>Example 10.3 New PDA Design</i>	278
10.3	Indicators Quantify Seasonality in Time Series	283
	<i>Example 10.4 Tyson’s Farm Worker Forecast</i>	283
	MEMO Re: Declining Supply of Self Employed Agriculture Workers	290
10.4	Indicators Add Structural Shifts in Time Series	291
	<i>Example 10.5 Leadership Changes Influence US Imports by India</i>	291
10.5	Indicators Allow Comparison of Segments and Scenarios And Quantify Structural Shifts	294
Excel 10.1	Use indicators to find part worth utilities and attribute importances from conjoint analysis data	295
Excel 10.2	Add indicator variables to account for segment differences or structural shifts	299
	<i>Lab Practice 10</i>	306
	<i>Assignment 10-1 Conjoint Analysis of PDA Preferences</i>	308
	<i>CASE 10-1 Modeling Growth: Procter &amp; Gamble Quarterly Revenues</i>	309
	<i>CASE 10-2 Store24 (A): Managing Employee Retention and Store24 (B): Service Quality and Employee Skills</i>	312

<b>Chapter 11 Nonlinear Multiple Regression Models</b>	<b>313</b>
11.1 Consider a Nonlinear Model When Response Is Not Constant	313
11.2 Tukey's Ladder of Powers	313
11.3 Rescaling $y$ Builds in Synergies	315
<i>Example 11.1 Executive Compensation</i>	315
11.4 Sensitivity Analysis Reveals the Relative Strength of Drivers	320
MEMO Re: Executive Compensation Driven by Firm Performance and Age	323
11.5 Gains from Nonlinear Rescaling Are Significant	324
11.6 Nonlinear Models Offer the Promise of Better Fit and Better Behavior	325
Excel 11.1 Rescale to build and fit nonlinear regression models with linear regression	326
Excel 11.2 Consider synergies in sensitivity analysis with a nonlinear model	334
<i>Lab Practice 11</i>	338
<i>CASE 11-1 Global Emissions Segmentation: Markets Where Hybrids Might Have Particular Appeal</i>	339
<b>Chapter 12 Indicator Interactions for Structural Differences or Changes in Response</b>	<b>343</b>
12.1 Indicator Interaction with a Continuous Influence Alters Its Partial Slope	343
<i>Example 12.1 Gender Discrimination at Slams Club</i>	344
MEMO Re: Women are Paid More than Men at Slam's Club	350
<i>Example 12.2 Car Sales in China</i>	351
12.2 Indicator Interactions Capture Segment Differences or Structural Differences in Response	358
Excel 12.1 Add indicator interactions to capture segment differences or structural differences in response	359
<i>Lab Practice 12</i>	370
<i>CASE 12-1 Explain and Forecast Defense Spending for Rolls-Royce</i>	372
<i>CASE 12-2 Haier's U.S. Refrigerator Strategy</i>	375
<b>Chapter 13 Logit Regression for Bounded Responses</b>	<b>377</b>
13.1 Rescaling Probabilities or Shares to Odds Improves Model Validity	377
<i>Example 13.1 The Import Challenge</i>	378
MEMO Re: Fuel Efficiency Drives Hybrid Owner Satisfaction	385
<i>Example 13.2 Presidential Approval Proportion</i>	386

13.2	Logit Models Provide the Means to Build Valid Models of Shares And Proportions	390
Excel 13.1	Rescale a limited dependent variable to logits	391
	<i>Assignment 13-1 Big Drug Co Scripts</i>	399
	<i>CASE 13-1 Alltel's Plans to Capture Share in the Cell Phone Service Market</i>	400
	<i>CASE 13-2 Pilgrim Bank (A): Profitability and Pilgrim Bank (B): Customer Retention</i>	403
	<b>Index</b>	<b>405</b>

## **Preface**

Exceptional managers know that they can create competitive advantages by basing decisions on performance response under alternative scenarios. To create these advantages, managers need to understand how to use statistics to provide information on performance response under alternative scenarios. Statistics are created to make better decisions. Statistics are essential and relevant. Statistics must be easily and quickly produced using widely available software, Excel. Then results must be translated into general business language and illustrated with compelling graphics to make them understandable and usable by decision makers.

This book helps students master this process of using statistics to create competitive advantages as decision makers. Statistics are essential, relevant, easy to produce, easy to understand, valuable, and fun, when used to create competitive advantage.

### **The Examples, Assignments, And Cases Used To Illustrate Statistics For Decision Making Come From Business Problems**

McIntire Corporate Sponsors and Partners, such as Rolls-Royce, Procter & Gamble, and Dell, and the industries that they do business in, provide many realistic examples. The book also features a number of examples of global business problems, including those from important emerging markets in China and India. It is exciting to see how statistics are used to improve decision making in real and important business decisions. This makes it easy to see how statistics can be used to create competitive advantages in similar applications in internships and careers.

### **Learning Is Hands On With Excel and Shortcuts**

Each type of analysis is introduced with one or more examples. First, the story of what exactly statistics can provide to decision makers is revealed. Following are examples illustrating the ways that statistics could actually be used to improve decision making. Analyses from Excel is shown and translated so that it is easy to see what the numbers mean to decision makers.

Included in Excel sections which follow are screenshots of an example analysis. Step by step instructions with screen shots allow easy master Excel. Featured are a number of popular Excel shortcuts, which are, themselves, a competitive advantage. Following Excel examples are lab practice problems, designed to closely resemble the chapter examples. Assignments and cases follow, with additional applications to new decision problems.

Powerful PivotTables and PivotCharts are introduced early and used throughout the book. Results are illustrated with graphics from Excel.

Beginning in Chapter 9, Harvard Business School cases are suggested which provide additional opportunities to use statistics to advantage.

## Focus Is On What Statistics Mean to Decision Makers and How to Communicate Results

From the beginning, results are translated into English. In Chapter 5, results are condensed and summarized in memos, the standard of communication in businesses. Later chapters include example memos for students to use as templates, making communication of statistics for decision making an easy skill to master.

Instructors, give your students the powerful skills that they will use to create competitive advantages as decision makers. Students, be prepared to discover that statistics are a powerful competitive advantage. Your mastery of the essential skills of creating and communicating statistics for improved decision making will enhance your career and make numbers fun.

## Acknowledgements

Preliminary editions of *Business Statistics for Competitive Advantage* were used at The McIntire School, University of Virginia, and I thank the many bright, motivated and enthusiastic students who provided comments and suggestions. Special thanks to Senior Associate Dean Rick Netemeyer, The McIntire School, University of Virginia, for his helpful suggestions, support, encouragement and camaraderie, and to Professor Tony Baglioni, also The McIntire School, University of Virginia, for many excellent comments and suggestions.

My appreciation and gratitude goes to John Kimmel, Springer, for sharing my vision and making this text a reality.

*Cynthia Fraser*  
*Charlottesville, VA*

# 1

## Statistics for Decision Making and Competitive Advantage

In the increasingly competitive global arena of business in the Twenty First century, the select few business graduates distinguish themselves by enhanced decision making backed by statistics. Statistics are useful when they are applied to improve decision making. No longer is the production of statistics confined to quantitative analysis and market research divisions in firms. Managers in each of the functional areas of business use statistics daily to improve decision making. Excel and other statistical software live in our laptops, providing immediate access to statistical tools which can be used to improve decision making.

### 1.1 Statistical Competences Translate Into Competitive Advantages

The majority of business graduates can create descriptive statistics and use Excel. Fewer have mastered the ability to frame a decision problem so that information needs can be identified and satisfied with statistical analysis. Fewer can build powerful and valid models to identify performance drivers, compare decision alternative scenarios, and forecast future performance. Fewer can translate statistical results into general business English that is easily understood by everyone in a decision making team. Fewer have the ability to illustrate memos with compelling and informative graphics. Each of these competences provides competitive advantage to those few who have mastery. This text will help you to attain these competences and the competitive advantages which they promise.

### 1.2 Attain Statistical Competences And Competitive Advantage With This Text

Most examples in the text are taken from real businesses and concern real decision problems. A number of examples focus on decision making in global markets. By reading about how executives and managers successfully use statistics to increase information and improve decision making in a variety of mini-case applications, you will be able to frame a variety of decision problems in your firm, whether small or multi-national. The end-of-chapter assignments will give you practice framing diverse problems, practicing statistical analyses, and translating results into easily understood reports or presentations.

Many examples in the text feature bottom line conclusions. From the statistical results, you read what managers would conclude with those results. These conclusions and implications are written in general business English, rather than statistical jargon, so that anyone on a decision team will understand. Assignments ask you to feature bottom line conclusions and general business English.

Translation of statistical results into general business English is necessary to insure their effective use. If decision makers, our audience for statistical results, don't understand the conclusions and implications from statistical analysis, the information created by analysis

will not be used. An appendix is devoted to writing memos that your audience will read and understand, and to effective PowerPoint slide designs for effective presentation of results. Memos and PowerPoints are predominant forms of communication in businesses. Decision making is compressed and information must be distilled, well written and illustrated. Decision makers read memos. Use memos to make the most of your analyses, conclusions and recommendations.

In the majority of examples, analysis includes graphics. Seeing data provides an information dimension beyond numbers in tables. To understand well a market or population, you need to see it, and its shape and dispersion. To become a master modeler, you need to be able to see how change in one variable is driving a change in another. Graphics are essential to solid model-building and analysis. Graphics are also essential to effective translation of results. Effective memos and PowerPoint slides feature key graphics which help your audience digest and remember results. We feature PivotTables and PivotCharts in Chapter Eight. These are routinely used in business to efficiently organize and display data. When you are at home in the language of PivotTables and PivotCharts, you will have a competitive advantage. Practice using PivotTables and PivotCharts to organize financial analyses and market data. Form the habit of looking at data and results whenever you are considering decision alternatives.

### 1.3 Follow The Path Toward Statistical Competence and Competitive Advantage

This text assumes no prior statistical knowledge, but covers basics quickly. Basics form the foundation for essential model building. Chapters Two and Three present a concentrated introduction to data and their descriptive statistics, samples and inference. Learn how to efficiently describe data and how to infer population characteristics from samples.

Model building with simple regression begins in Chapter Four and occupies the focus of the remaining chapters. To be competitive, business graduates must have competence in model building and forecasting. A model-building mentality, focused on performance drivers and their synergies is a competitive advantage. Practice thinking of decision variables as drivers of performance. Practice thinking that performance is driven by decision variables. Performance will improve if this linkage becomes second-nature.

The approach to model building is steeped in logic and begins with logic and experience. Models must make sense in order to be useful. When you understand how decision variables drive performance under alternate scenarios, you can make better decisions, enhancing performance. Model-building is an art that begins with logic.

Model building chapters include nonlinear regression and logit regression. Nearly all aspects of business performance behave in nonlinear ways. We see diminishing or increasing changes in performance in response to changes in drivers. It is useful to begin model building with the simplifying assumption of constant response, but it is essential to

be able to grow beyond simple models to realistic models which reflect nonconstant response. Logit regression, appropriate for the analysis of bounded performance measures such as market share and probability of trial, has many useful applications in business and is an essential tool for managers. Resources and markets are limited, and responses to decision variables are also necessarily limited, as a consequence. Visualize the changing pattern of response when you consider decision alternatives and the ways they drive performance.

#### 1.4 Use Excel for Competitive Advantage

This text features widely available Excel software, including many commonly used shortcuts. Excel is powerful, comprehensive, and user-friendly. Appendices with screenshots follow each chapter to make software interactions simple. Recreate the chapter examples by following the steps in the Excel sections. This will give you confidence using the software. Then forge ahead and generalize your analyses by working through end-of-chapter assignments. The more often you use the statistical tools and software, the easier analysis becomes.

#### 1.5 Statistical Competence Is Satisfying

Statistics and their potential to alter decisions and improve performance are important to you. With more and better information from statistical analysis, we make superior decisions and outperform the competition. You will find your ability to apply statistics to decision making scenarios is satisfying. You will find that the competitive advantages from statistical competence are powerful and yours.

## 2

# Describing Your Data

This chapter introduces *descriptive* statistics, which are almost always included with any statistical analysis to characterize a dataset. The particular descriptive statistics we use depend on the *scale* that has been used to assign numbers to represent the characteristics of entities being studied. When the distribution of continuous data is bell-shaped, we have convenient properties that make description easier. Chapter Two looks at dataset types and their description.

### 2.1 Describe Data With Summary Statistics And Histograms

We use numbers to measure aspects of businesses, customers and competitors. These sets of measured aspects are *data*. Data become meaningful when we use statistics to describe patterns within particular *samples* or collections of businesses, customers, competitors, or other entities.

*Example 2.1 Yankees' Salaries: Is it a Winning Offer?* Suppose that the Yankees want to sign a promising rookie. They expect to offer \$1M, and they want to be sure they are neither paying too much nor too little. What would the General Manager need to know to decide whether or not this is the right offer?

He might first look at how much the other Yankees earn. Their 2005 salaries are in Table 2.1:

Crosby	\$.3	Johnson	\$16.0	Posada	\$11.0	Sierra	\$1.5
Flaherty	.8	Martinez	2.8	Rivera	10.5	Sturtze	.9
Giambi	1.34	Matsui	8.0	Rodriguez	21.7	Williams	12.4
Gordon	3.8	Mussina	19.0	Rodriguez F	3.2	Womack	2.0
Jeter	19.6	Phillips	.3	Sheffield	13.0		

Table 2.1 Yankees' salaries (in \$MM) in alphabetical order

What should he do with this data?

Data are more useful if they are ordered by the aspect of interest. In this case, the Manager would re-sort the data by salary (Table 2.2):

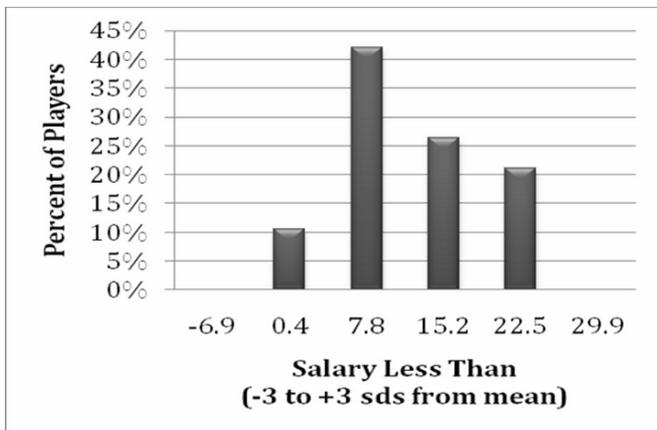
Rodriguez	\$21.7	Williams	\$12.4	Rodriguez F	\$3.2	Sturtze	\$.9
Jeter	19.6	Posada	11.0	Martinez	2.8	Flaherty	.8
Mussina	19.0	Rivera	10.5	Womack	2.0	Crosby	.3
Johnson	16.0	Matsui	8.0	Sierra	1.5	Phillips	.3
Sheffield	13.0	Gordon	3.8	Giambi	1.3		

Table 2.2 Yankees sorted by salary (in \$MM)

Now he can see that the lowest Yankee salary, the *minimum*, is \$300,000, and the highest salary, the *maximum*, is \$21,700,000. The difference between the maximum and the minimum is the *range* in salaries, which is \$21,400,000, in this example. From these statistics, we know that the salary offer of \$1MM falls in the lower portion of this range. Additionally, however, he needs to know just how unusual the extreme salaries are to better assess the offer.

He'd like to know whether or not the rookie would be in the better-paid half of the Team. This could affect morale of other players with lower salaries. The *median*, or middle, salary is \$3,800,000. We know this because the lower-paid half of the team earns between \$300,000 and \$3,800,000, and the higher-paid half of the team earns between \$3,800,000 and \$21,700,000. Thus, he would be in the bottom half. The Manager needs to know more to fully assess the offer.

Often, a *histogram* and a *cumulative distribution plot* are used to visually assess data, as shown in Figures 2.1 and 2.2.



<i>salary (\$MM)</i>	
25%	1.42
<i>median</i>	3.8
75%	12.7

The histogram of team salaries shows us that more than 40% of the players earn more than \$400,000, but less than the average, or *mean*, salary of \$7,800,000.

Figure 2.1 Histogram of Yankee salaries

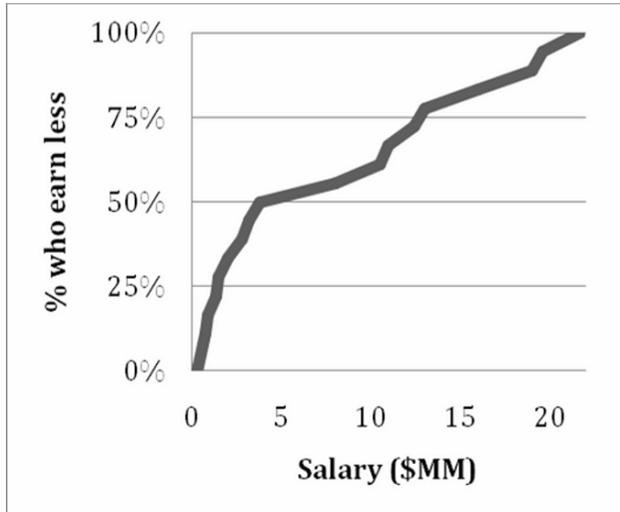


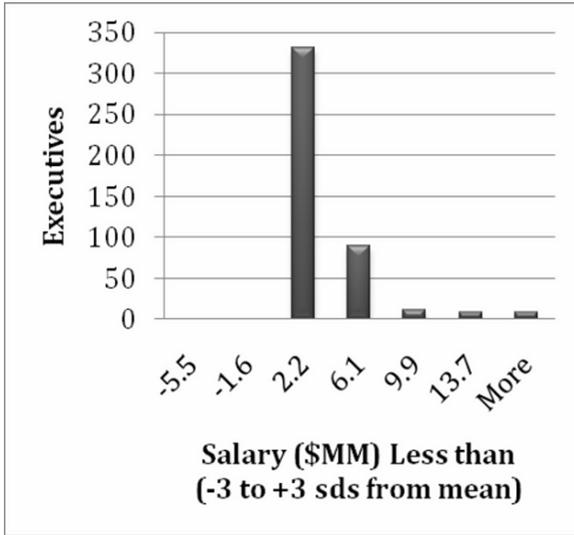
Figure 2.2 Cumulative distribution of salaries

The cumulative distribution reveals that the *Interquartile Range* between the 25<sup>th</sup> percentile and the 75<sup>th</sup> percentile is more than \$10 million. A quarter earn less than \$1.42 million, the 25<sup>th</sup> percentile, half earn between \$1.42 and \$12.7 million, and quarter earn more than \$12.7 million, the 75<sup>th</sup> percentile. Half of the players have salaries below the *median* of \$3.8 million and half have salaries above \$3.8 million.

## 2.2 Outliers Can Distort The Picture

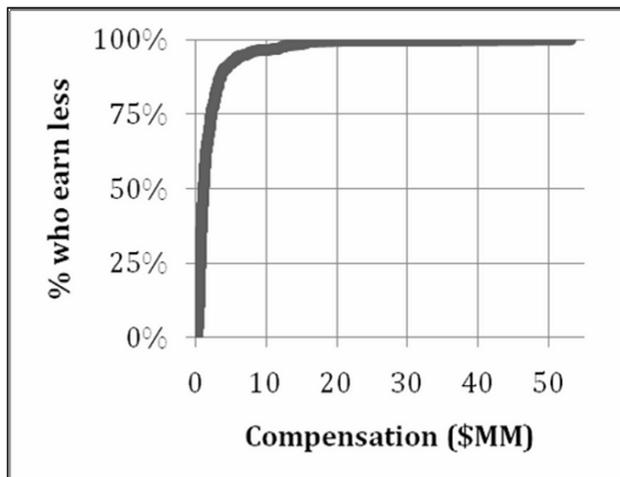
Outliers are extreme elements, considered unusual when compared with other sample elements. Because they are extraordinary, they can distort descriptive statistics.

*Example 2.2 Executive Compensation: Is the Board's Offer on Target?* The Board of a large corporation is pondering the total compensation package of the CEO, which includes salary, stock ownership, and fringe benefits. Last year, the CEO earned \$2,000,000. For comparison, The Board consulted Forbes' summary of the total compensation of the 500 largest corporations. The histogram, cumulative frequency distribution and descriptive statistics are shown in Figures 2.3 and 2.4.



<i>Total Compensation (sds from mean -3 to +3)</i>	<i>Frequency</i>
-5.46	0
-1.62	0
2.22	331
6.06	90
9.9	10
13.74	8
More	8

Figure 2.3 Histogram of executive compensation



<i>Total Compensation (\$MM)</i>	
<i>mean</i>	2.22
<i>sd</i>	3.84
<i>75<sup>th</sup> percentile</i>	2.26
<i>median</i>	1.13
<i>25<sup>th</sup> percentile</i>	0.72

Figure 2.4 Cumulative distribution of total compensation

The average executive compensation in this sample of large corporations is \$2.22 million. The least well-compensated executive earns \$29,000 and the best-compensated executive earns more than \$53,000,000. Half the sample of 447 executives earns \$1.13 million (the median) or less. One quarter earns less than \$.72 million, the middle half, or *interquartile range*, earns between \$.72 million and \$2.26 million, and one quarter earns more than \$2.26 million.

Why is the *mean*, \$2.22 million, so much larger than the *median*, \$1.13 million? There is a group of eight *outliers*, shown as *MORE* than three standard deviations above the mean in Figure 2.3, who are compensated extraordinarily well. Each collects a compensation package of more than \$13.7 million, a compensation level that is more than three standard deviations greater than the mean.

When we exclude these eight outliers, eleven additional outliers emerge. This cycle repeats, since the distribution is highly skewed. When we removed outliers, the new mean is adjusted, making other executives appear to be more extreme. As a rule of thumb, remove no more than ten percent of the sample. In this case, removing about ten percent, or the 44 best-compensated executives, gives us a better picture of what “typical” compensation is, shown in Figure 2.5:

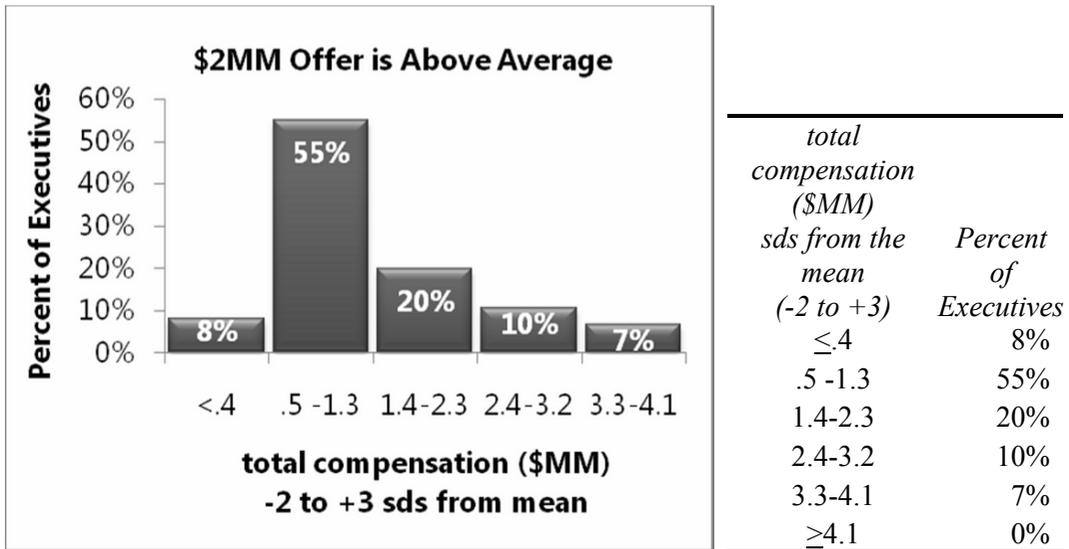


Figure 2.5 Histogram and descriptive statistics with 44 outliers excluded

Ignoring the 44 outliers, the average compensation is about \$1,400,000, and the *median* compensation is about \$1,000,000, shown in Figure 2.6:

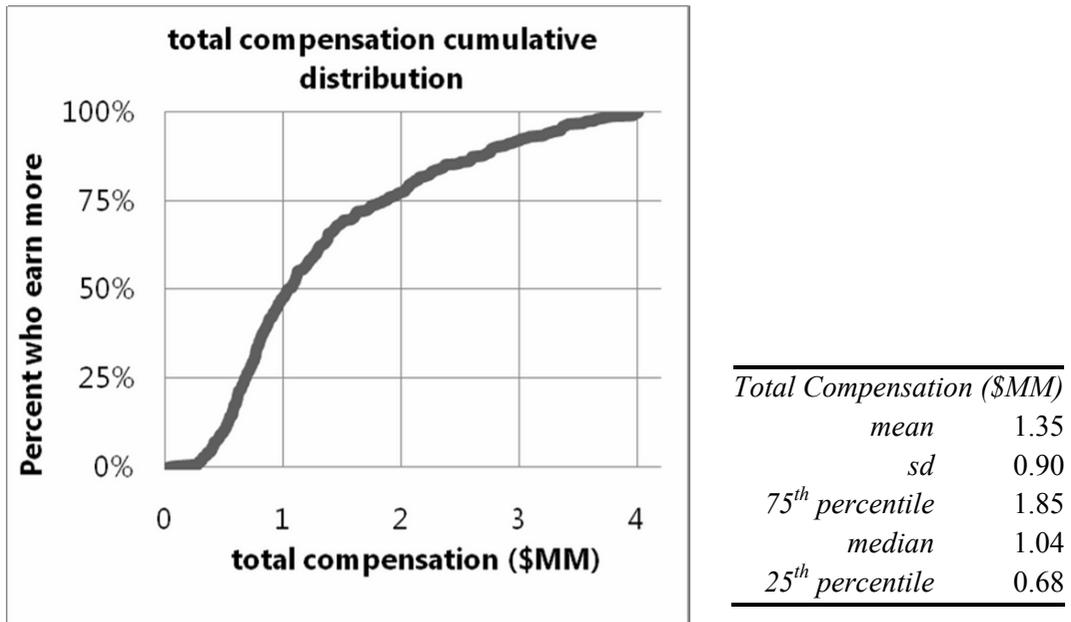


Figure 2.6 Cumulative distribution of total compensation

The *mean* and *median* are closer. With this more representative description of executive compensation in large corporations, The Board has an indication that the \$2,000,000 package is well above average. More than three quarters of executives earn less. Because extraordinary executives exist, the original distribution of compensation is *skewed*, with relatively few exceptional executives being exceptionally well compensated.

### 2.3 Round Descriptive Statistics

In the examples above, statistics in the output from statistical packages are presented with many decimal points of accuracy. The Yankee manager in Example 2.1 and The Board considering executive compensation in Example 2.2 will most likely be negotiating in hundred thousands. It would be distracting and unnecessary to report descriptive statistics with significant digits more than two or three. In the **Yankees** example, the average salary is \$7,800,000 (*not* \$7,797,000). In the **Executive Compensation** example, average total compensation is \$1,400,000 (*not* \$1,387,494). It is deceptive to present results with many significant digits, creating an illusion of accuracy. In addition to being honest, statistics in two or three significant digits are much easier for decision makers to process and remember.

## 2.4 Central Tendency and Dispersion Describe Data

The baseball salaries and executive compensation examples focused on two measures of *central tendency*: the *mean*, or average, and the *median*, or middle. Both examples also refer to a measure of *dispersion* or variability: the *range* separating the minimum and maximum. To describe data, we need statistics to assess both central tendency and dispersion. The statistics we choose depends on the *scale* which has been used to code the data we are analyzing.

## 2.5 Data Is Measured With Quantitative or Categorical Scales

If the numbers in a dataset represent amount, or magnitude of an aspect, **and** if differences between adjacent numbers are equivalent, the data are *quantitative* or *continuous*. Data measured in dollars (i.e., revenues, costs, prices and profits) or percents (i.e., market share, rate of return, and exam scores) are continuous. We can add, subtract, divide or multiply quantitative variables to find meaningful results.

When we have quantitative data, we report central tendency with the *mean*,

$$\mu = \frac{\sum x_i}{N} \text{ for describing a } \textit{population} \text{ and}$$

$$\bar{X} = \frac{\sum x_i}{N} \text{ for describing a } \textit{sample} \text{ from a population,}$$

where  $x_i$  are data point values, and

$N$  is the number of data points that we are describing.

We also use the *median* to assess central tendency and the *range*, *variance*, and *standard deviation* to assess dispersion. The *variance* is the average squared difference between each of the data points and the mean:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \text{ for a population and}$$

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{(N - 1)} \text{ for a sample from a population.}$$

The *standard deviation*  $\sigma$  (for a population) and  $s$  (for a sample) is the square root of the variance, which gives us a measure of dispersion in the more easily interpreted, original units, rather than squared units.

If numbers in a dataset are arbitrary and used to distinguish categories, the data are *nominal*, or *categorical*. Football jersey numbers and your student ID are nominal. A larger number doesn't mean that a player is better or a student is older or smarter. We can tabulate nominal data to find the most popular number occurring most frequently, the *mode*, which we use to report central tendency. We cannot add, subtract, divide or multiply nominal numbers.

Quantitative measures convey the most information, including direction and magnitude, while categorical measures convey the least and merely identify category membership. In between quantitative and categorical scales are *ordinal* scales that we use to rank order data, or to convey direction, but not magnitude. With ordinal data, an element (which could be a business, a person, a country) with the most or best is coded as '1', second place as '2', etc. With ordinal numbers, we can sort the data, but we cannot add, subtract, divide or multiply the rankings. Just as with other categorical data, we rely on the mode to report central tendency of ordinal data.

When focus is on membership in a particular category, the *proportion* of sample elements in the category is a continuous measure of central tendency. Proportions are quantitative and can be added, subtracted, divided or multiplied, though they are bounded by zero, below, and by one, above.

## 2.6 Continuous Data Tend To Be Normal

Continuous variables are often *Normally distributed*, and their histograms resemble bell-shaped curves, with the majority of data points clustered around the mean. Most elements are "average" with values near the mean; fewer elements are unusual and far from the mean. If continuous data are Normally distributed, we need only the mean and standard deviation to describe this data and our description is simplified.

*Example 2.3 Normal SAT Scores.* Standardized tests, such as SAT, capitalize on Normality. Math and verbal SATs are both specifically constructed to produce Normally distributed scores with *mean* = 500 and *standard deviation* = 100 over the population of students (Figure 2.7):

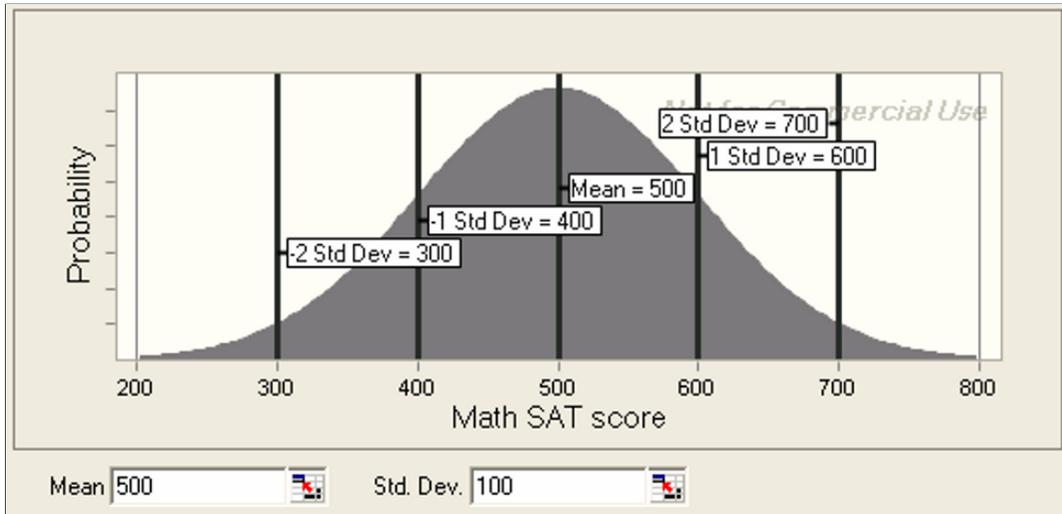


Figure 2.7 Normally distributed SAT scores

## 2.7 The Empirical Rule Simplifies Description

Normally distributed data have a very useful property known as the *Empirical Rule*:

- 2/3 of the data lie within one standard deviation of the mean
- 95% of the data lie within two standard deviations of the mean

This is a powerful rule! *If data are Normally distributed, we can describe the data with just two statistics: the mean and the standard deviation.*

Returning to SAT scores, if we know that the average score is 500 and the standard deviation is 100, we also know that

- 2/3 of SAT scores will fall within 100 points of the mean of 500, or between 400 and 600,
- 95% of SAT scores will fall within 200 points of the mean of 500, or between 300 and 700.

*Example 2.4 Class of '06 SATs: This Class is Normal & Exceptional.* Descriptive statistics and a histograms of Math SATs of a third year class of business students reveal an interquartile range from 640 to 730, with mean of 685 and standard deviation of 70, as shown in Figure 2.8:

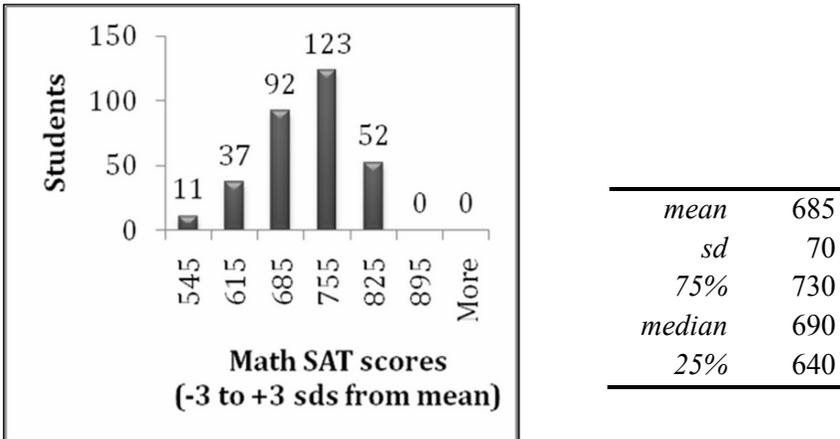


Figure 2.8 Histograms and descriptive statistics of class '06 math SATs

Are Class '06 Math SATs Normally distributed? Approximately. Class '06 scores are bell shaped, though negatively skewed. There are “too many” perfect scores of 800.

The Empirical Rule would predict that  $2/3$  of the class would have scores within one standard deviation of 70 points of the mean of 685, or within the interval 616 to 755. There actually 68% (=29%+39%), though there are more scores one standard deviation above the mean than below.

The Empirical Rule would also predict that only 2-1/2% of the class would have scores more than two standard deviations below or above the mean of 685: scores below 545 and above 825. We find that 3% actually do have scores below 545, though none score above 825 (since a perfect SAT score is 800). This class of business students has Math SATs that are nearly Normal, but not exactly Normal.

To summarize Class '06 students' SAT scores, we would report:

- Class '06 students' Math SAT scores are approximately Normally distributed with *mean* of 685 and *standard deviation* of 70.
- Relative to the larger population of all SAT-takers, the smaller *standard deviation* in Class '06 students' Math SAT scores, 70 versus 100, indicates that Class '06 students' are a more homogeneous group than the more varied population.

## 2.8 Describe Categorical Variables Graphically: Column and PivotCharts

Numbers representing category membership in nominal, or categorical, data are described by tabulating their frequencies. The most popular category is the *mode*. Visually, we show our tabulations with a *Pareto* chart, which orders categories by their popularity.

*Example 2.5 Who Is Honest & Ethical?* Figure 2.9 shows a column chart of results of a survey of 1,014 adults by Gallup in 2004:

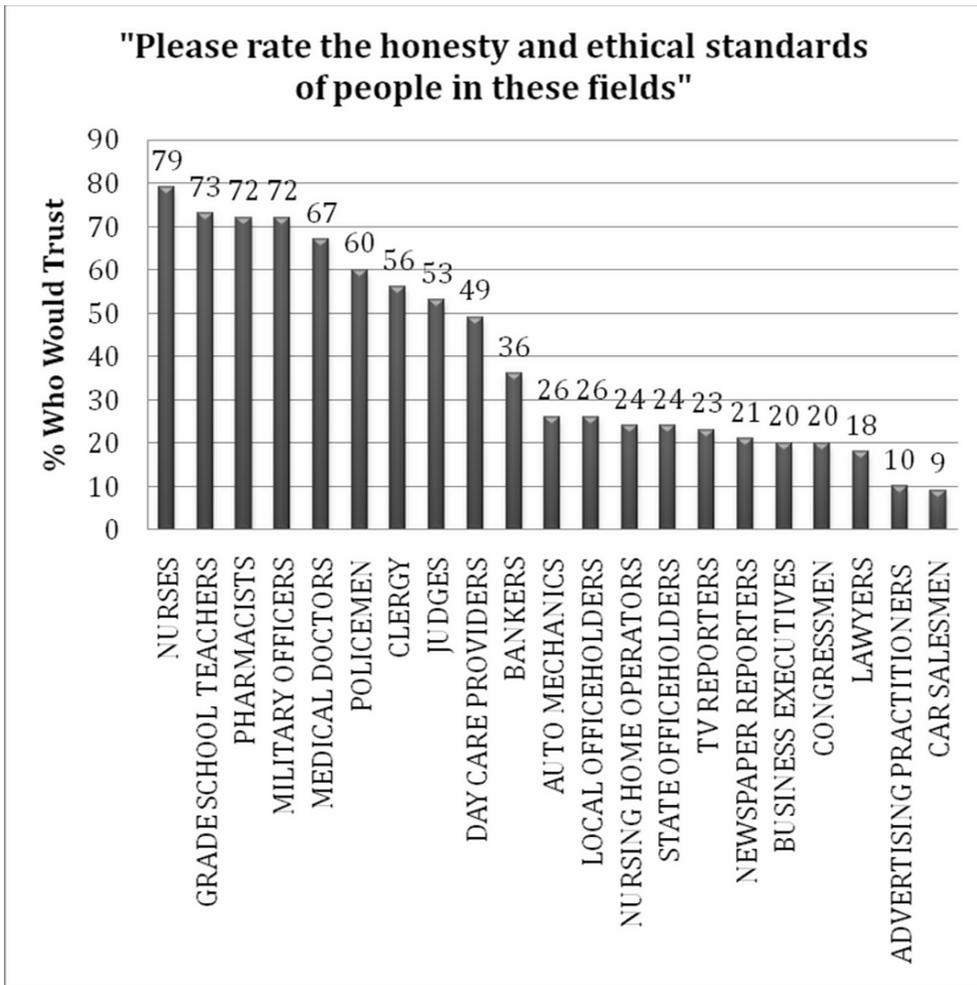


Figure 2.9 Pareto charts of the percents who judge professions honest

More Americans trust and respect nurses (79%, the *modal* response) than people in other professions, including doctors, clergy and teachers. Though a small minority judge business executives (20%) and advertising professionals (10%) as honest and ethical, most do not judge people in those fields to be honest (which highlights the importance of ethical business behavior in the future).

## 2.9 Descriptive Statistics Depend On The Data

Descriptive statistics, graphics, central tendency and dispersion, depend upon the type of scale used to measure data characteristics (i.e., quantitative or categorical). Table 2.3 summarizes the descriptive statistics (graph, central tendency, dispersion) that we use for both types of data:

	Quantitative	Categorical
<b>Central Tendency</b>	<i>mean</i> <i>median</i>	<i>mode</i> <i>proportion</i>
<b>Dispersion</b>	<i>range</i> <i>standard deviation</i>	
<b>Graphics</b>	<i>histogram</i> <i>cumulative distribution</i>	<i>Pareto chart</i> <i>pie chart</i> <i>column chart</i>

Table 2.3 Descriptive statistics (central tendency, dispersion, graphics) for two types of data

If continuous data are Normally distributed, we can completely describe a dataset with just the mean and standard deviation. We know from the *Empirical Rule* that 2/3 of the data will lie within one standard deviation of the mean and that 95% of the data will lie within two standard deviations of the mean.



**Histograms.** To make a histogram of salaries, Excel needs to know what ranges of values to combine. We will set these *bins*, or categories to differences from the sample mean that are in widths of standard deviations.

The **histogram bins.xls** uses formulas to find cutoff values for histogram bins of three standard deviations below the mean to three standard deviations above the mean using a default mean of zero and standard deviation of 1. We will change these to the sample mean and standard deviation.

Open **histogram bins.xls**, select **A1:E9**, then use the shortcut **Cntl+C** to copy. In the **Executive Compensation** file, select **C1**, [**Enter**], to paste the **histogram bins** formulas into columns C through E.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	mean	standard deviation	histogram bins in sds from the mean (-3 to +3)	Normal	sds from mean												
2		0	1	-3	0.1%	≤-3 outliers											
3				-2	2.1%	≤ -2											
4				-1	13.6%	≤-1											
5				0	34.1%	≤ mean											
6				1	34.1%	≤+1											
7				2	13.6%	≤+2											
8				3	2.1%	≤+3											
9					0.1%	>3 outliers											

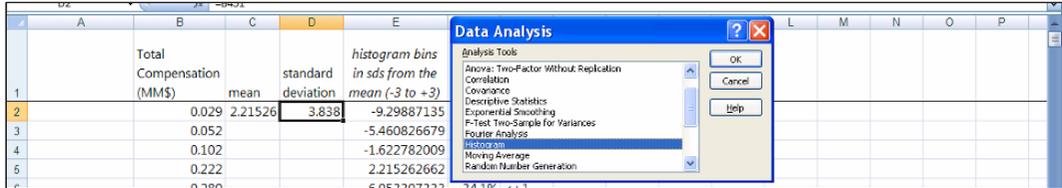
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		Total Compensation (M\$)	mean	standard deviation	histogram bins in sds from the mean (-3 to +3)	Normal	sds from mean									
1																
2		0.029	0	1	-3	0.1%	≤-3 outliers									
3		0.052			-2	2.1%	≤ -2									
4		0.102			-1	13.6%	≤-1									
5		0.222			0	34.1%	≤ mean									
6		0.280			1	34.1%	≤+1									
7		0.281			2	13.6%	≤+2									
8		0.291			3	2.1%	≤+3									
9		0.293				0.1%	>3 outliers									

In **C2**, replace the mean of zero with the sample mean by entering **=B450 [Enter]**.

In **D2**, replace the standard deviation of one with the sample standard deviation by entering **=B451 [Enter]**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		Total Compensation (M\$)	mean	standard deviation	histogram bins in sds from the mean (-3 to +3)	Normal	sds from mean									
1																
2		0.029	2.21526	3.838	-9.29887135	0.1%	≤-3 outliers									
3		0.052			-5.460826679	2.1%	≤ -2									
4		0.102			-1.622782009	13.6%	≤-1									
5		0.222			2.215262662	34.1%	≤ mean									
6		0.280			6.053307333	34.1%	≤+1									
7		0.281			9.891352004	13.6%	≤+2									
8		0.291			13.72939667	2.1%	≤+3									
9		0.293				0.1%	>3 outliers									

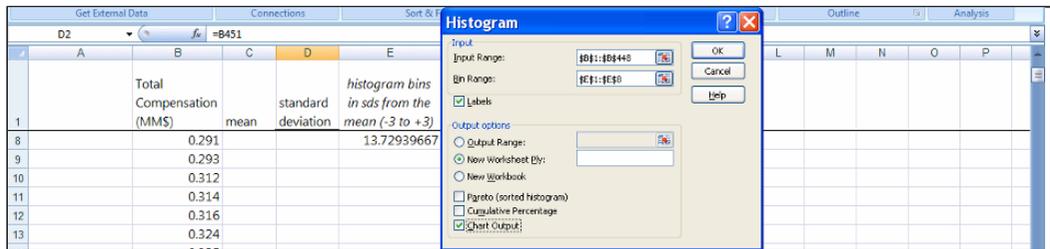
To see the distribution of *Total Compensation*, activate shortcuts with **Alt AY2 Histogram, OK.** (**Alt AY2** selects the Data menu and the Data Analysis menu.)



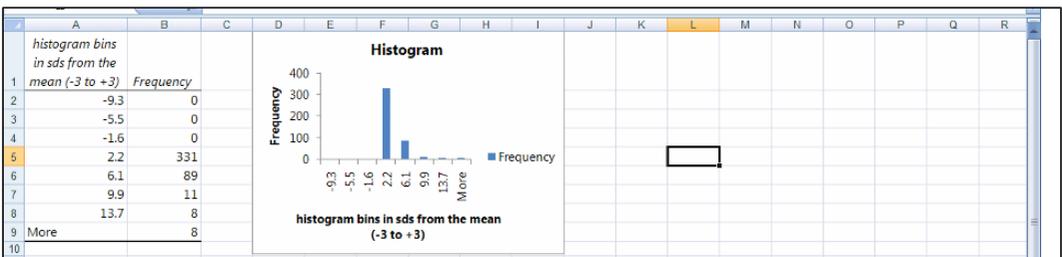
For **Input Range**, select **B1**, then use shortcuts to select the *Total Compensation* data in column **B** with **Cntl+Shift+down arrow**.

For **Bin Range**, select **E1**, then use shortcuts to select the histogram bins in column **E** with **Cntl+Shift+down arrow**.

Select **Labels and Chart Output**, then **OK**:



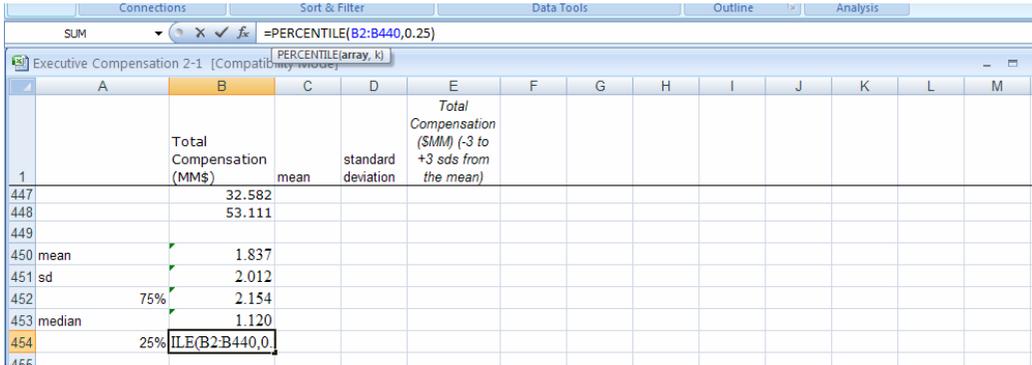
To reduce the unnecessary decimals, select **A2:A7**, then activate shortcuts **Alt H9** to reduce decimals. (**H** selects the Home menu and **9** selects the reduce decimals function of the Number menu.)





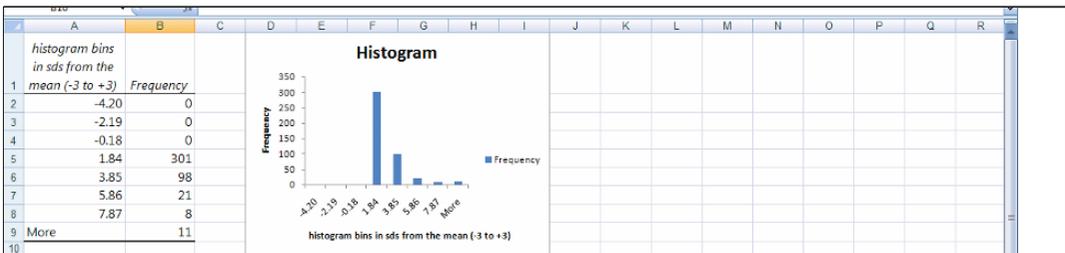
Recalculate the mean, standard deviation, 25%, median, and 75% percentile, including only rows with *total compensation* less than 13.7 million.

Change the end of the array in each Excel function from **454** to **440**.  
 (The histogram bins formulas will automatically update bin cutoffs with your new mean and standard deviation.)



	A	B	C	D	E	F	G	H	I	J	K	L	M
		Total Compensation (MM\$)	mean	standard deviation	Total Compensation (\$MM) (-3 to +3 sds from the mean)								
447		32.582											
448		53.111											
449													
450	mean	1.837											
451	sd	2.012											
452	75%	2.154											
453	median	1.120											
454	25%	=PERCENTILE(B2:B440,0.25)											

Re-run the histogram tabulation, excluding the outliers, changing the array end in **Input Data** from **448** to **440**:



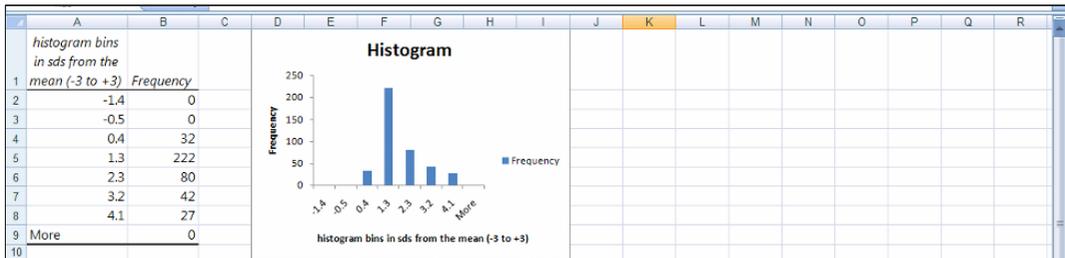
Update the descriptive statistics in **B450:B454** and re-run the histogram with only rows with compensation less than \$7.9 million, **B1:B429**.

The mean, \$1.60 million, and the median, \$1.11 million, are now much closer, though a new set of twelve outliers appears.

Continue excluding outliers, stopping before you have excluded 10% of the sample, or 45 executives. Since the distribution of total compensation is highly skewed, *outliers* will continue to appear. We will use the rule of thumb to exclude no more than 10% of a sample.

With rows **B1:B404**, including executives whose total compensation is less than \$4.1 million, the descriptive statistics are more representative:

	Total Compensation (MMS)	mean	standard deviation	histogram bins in sds from the mean (-3 to +3)	sds from Normal mean
450	mean	1.34997			
451	sd	0.903			
452	75%	1.847			
453	median	1.036			
454	$\Phi\%$	0.678			



The Board can be confident that the \$2 million package is an attractive one, better than 75% of other executives packages. There are also a number of better-paid executives, some earning as much as \$4.1 million, making \$2 million a reasonable offer for a talented executive.

## Excel 2.3 Plot a cumulative distribution

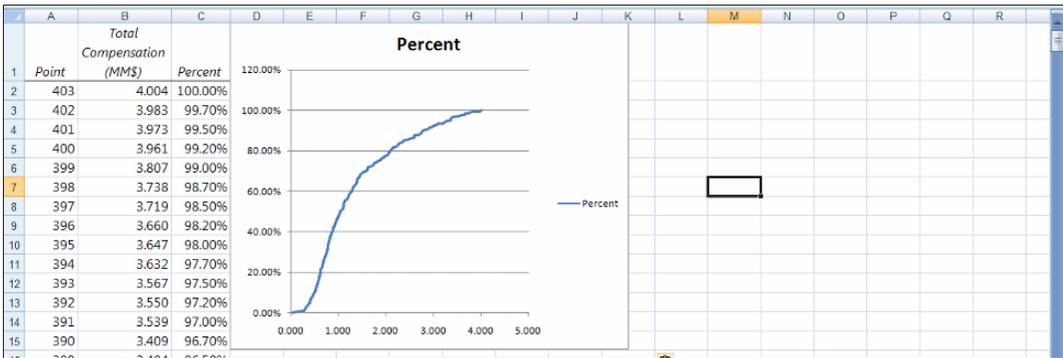
To see the cumulative distribution of total compensation, choose **Rank and Percentile** from the Data Analysis menu (**Alt AY2, Rank and Percentile, OK**), with **Input Range B1:B404, OK**:



To make the cumulative distribution plot from the **Rank and Percentiles**, first, for convenience, delete column C.

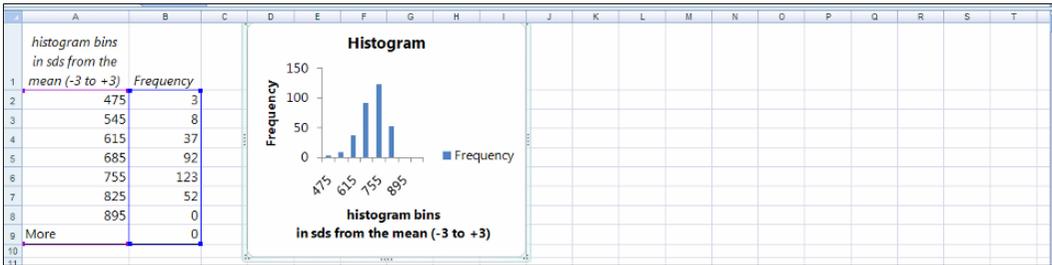
Select C, then use shortcuts to delete: **Alt HDC**. (**H** selects the Home menu, **D** selects the Delete menu, and **C** deletes the column.)

To plot *Total Compensation* in **B** by *Percent* in **C**, select **B** and **C**, then use shortcuts to insert a scatterplot (**Alt ND**):

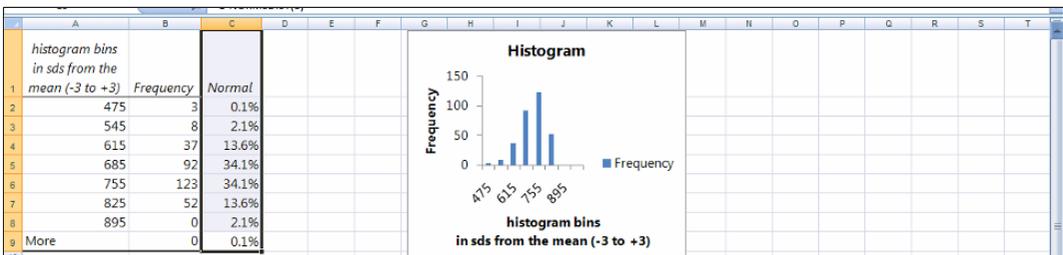




Order the histogram tabulation of *MathSATs*.

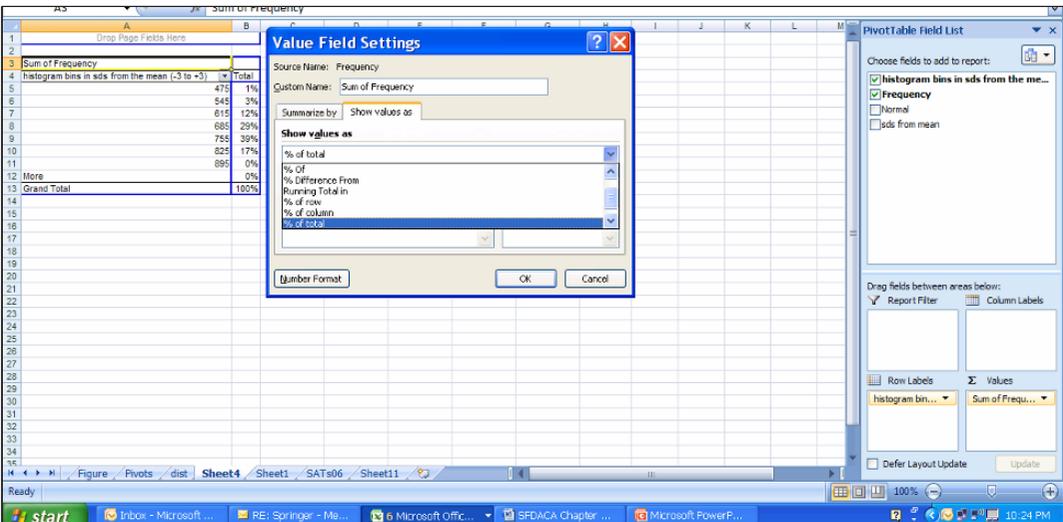


**PivotTable and PivotChart of a distribution in percents.** Reduce decimals in A2:A7, copy from the SATs '06 sheet the percents we would find in each bin were the distribution Normal, H1:H9, and paste into C1:C9 of the histogram sheet:

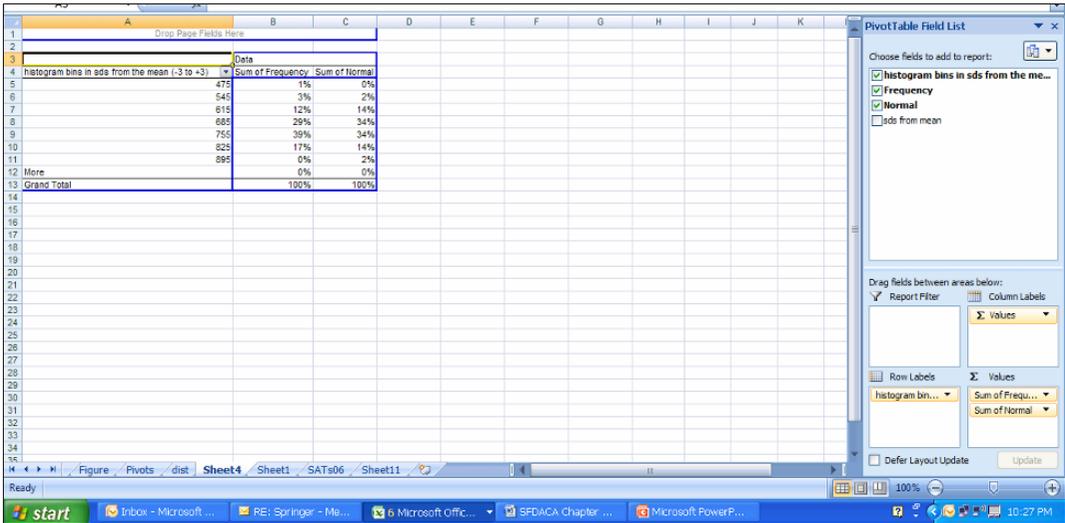


Select A1:C8, and make a PivotTable with shortcuts **Alt NVT**. (N selects the Insert menu, V selects the Pivot menu, and T inserts a PivotTable.)

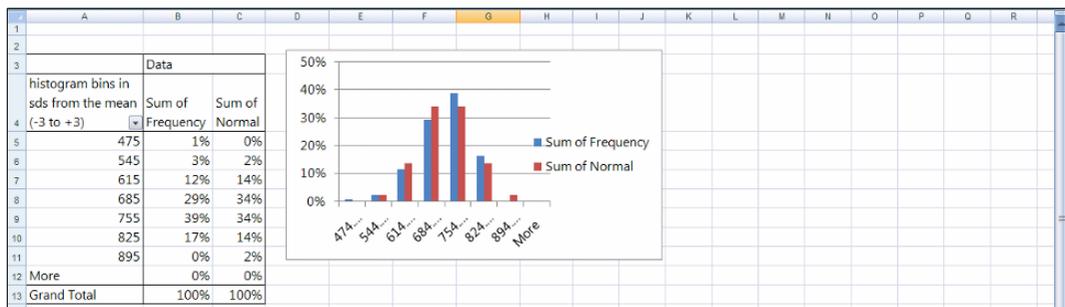
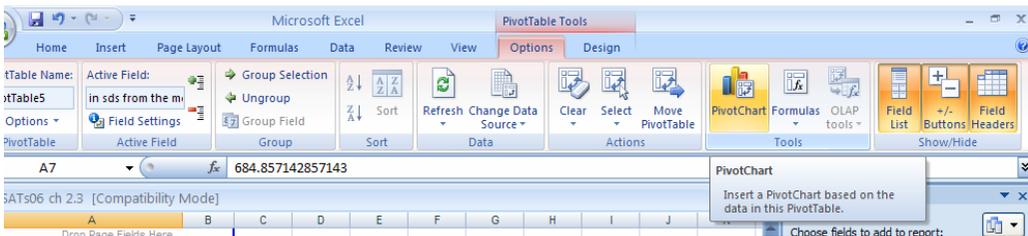
Set up your PivotTable, putting *histogram bins* in **ROW** and *Frequency* in **DATA**. Change the table to percents by double clicking *Sum of Frequency*, **Show values as, % of total**, **Ok**.



Add the *Normal* percents by dragging *Normal* to the  $\Sigma$  values box.



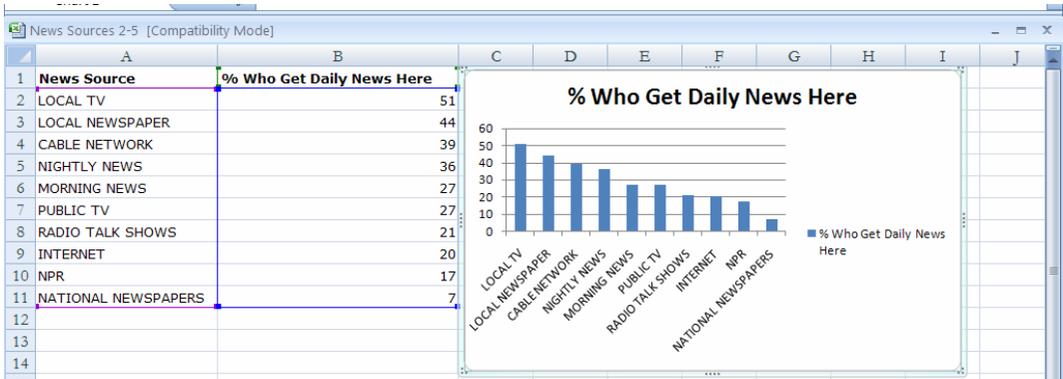
Click inside the table then choose the **Options** tab and click the **PivotChart** icon, **column**, **ok**:



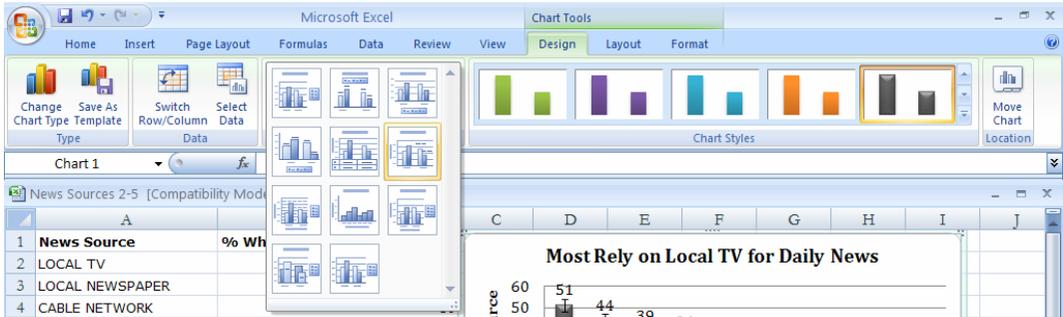
## Excel 2.5 Produce a column chart from a PivotChart of a nominal variable

A firm is targeting customers who consult a news source daily. Management wants to compare the popularity of news sources. To facilitate comparisons, we will make a PivotChart from a recent (December 2004) Gallup Poll of 992 Americans. Data are in **Excel 2.5 News Sources.xls**.

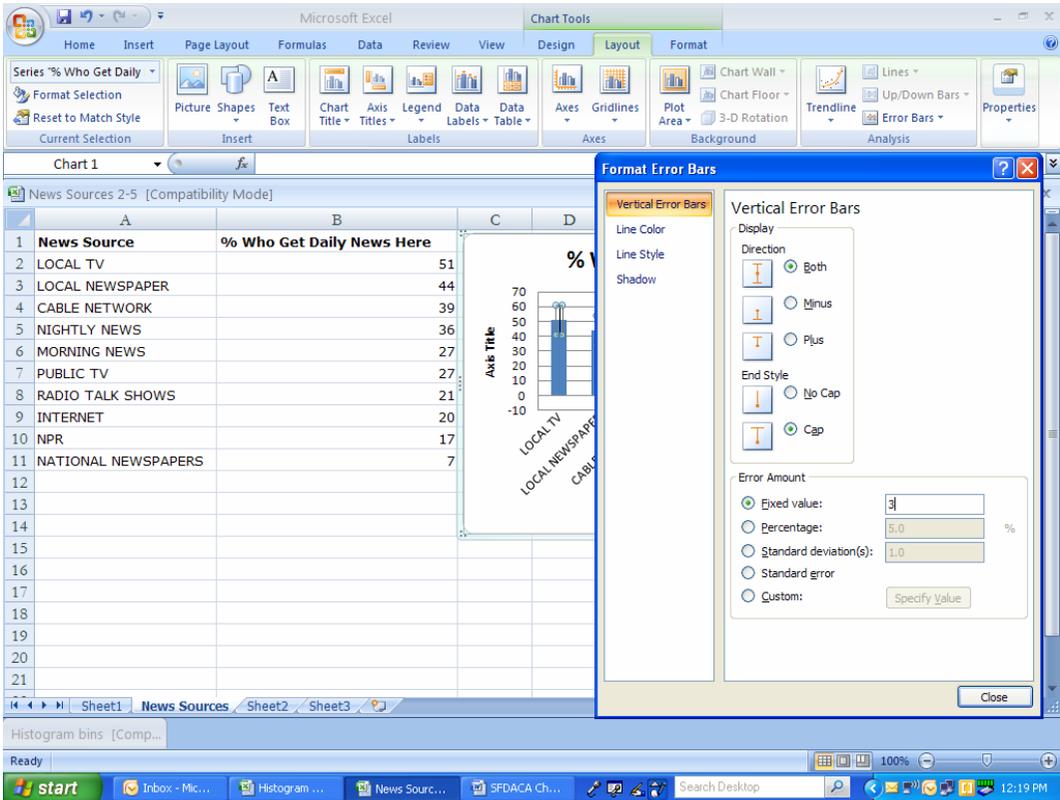
Open **Excel 2.5 News Sources.xls**, and select **A1:B11**, the use shortcuts to insert a column chart (**Alt NC**):



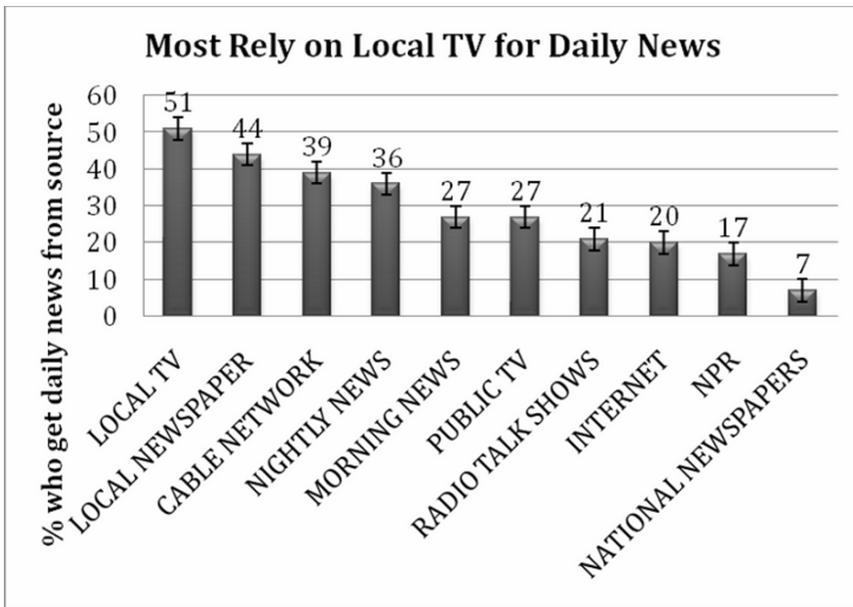
Choose **Design Chart Layout 6** (which features a vertical axis label) and select a **Design Chart Style**.



To add vertical margin of error bars, click inside a column, then use shortcuts to add error bars of 3, the approximate margin of error: **Alt JARM, Fixed value: 3, Close.** (**JA** selects the Layout menu, **R** selects the Error Bar menu, and **M** selects the custom Error Bar menu.)



Add data labels, a bottom line title and add the vertical axis title:



More Americans, 48 to 54%, get daily news from local TV than from any other news source.

## Excel Shortcuts at Your Fingertips

### By Shortcut Key

**Alt** activates the shortcuts menus, linking keyboard letters to Excel menus. Press **Alt**, then release and press letters linked to the menus you want.

The following are examples of shortcuts. Press **Alt**, then

**H 9** to select the Home menu and the reduce decimals function

**H DC** to select the Home menu and the Delete function to delete column(s)

**H IC** to select the Home menu and Insert function and to insert a column to the left of the selected cell or column

**AY2** to select the Data and Data Analysis menus

**AS** to select the Data and the Sort menus

**NC** to select the Insert function and to insert a column chart

**ND** to select the Insert function and to insert a scatterplot

**NE** to select the Insert function and to insert a pie chart

**NVT** to select the Insert function, the Pivot menu, and to insert a PivotTable

**NX** to select the Insert function and to insert a text box

**WFR** to select the View and Freeze panes menus, and to Freeze rows

**JAB** to select the Layout and Data Labels menus

**JARM** to select the Layout, the Error Bar, and the custom Error Bar menus

**Shift+arrow** selects cells scrolled over

**Cntl+C** to copy

**Cntl+down arrow** scrolls through all cells in the same column that contain data and stops at the last filled cell.

**Cntl+R** fills in values of empty cells using a formula from the first cell in a selected array

**Cntl+Shift+down arrow** selects all filled cells in the column.

## By Goal

If you want to

Activate shortcuts menus, press **Alt**, then release.

Add data labels in a column chart: select a column, then **Alt JAB**

Add error bars in a column chart: select a column, then **Alt JARM**

Analyze data: **Alt AY2**

Copy cells: select the cells, then **Cntl+C**

Delete a column: **Alt HDC**

Freeze the top row: **Alt WFR**

Insert a column: **Alt HIC**

Insert a column chart: **Alt NC**

Insert a pie chart: **Alt NE**

Insert a PivotTable: **Alt NVT**

Insert a row: **Alt HIR**

Insert a scatterplot: **Alt ND**

Insert a text box: **Alt NX**

Move to the end of a column: **Cntl+down arrow**

Reduce decimals: **Alt H9**

Select all of the filled cells in a column: select the first cell in the column, then **Cntl+Shift+down arrow**

Sort data: **Alt AS**

## Lab 2 Descriptive Statistics

### A Typical Executive's Compensation

Help the Board of firm in the financial industry evaluate the \$2MM compensation package that they expect to offer the CEO. Summarize the Forbes data on executives' compensation in **Lab 2 Executive Compensation.xls**.

1. Find the sample mean and standard deviation, and then make a histogram of *compensation* in financial firms.  
(See your text for a similar example in Excel.)

Average *compensation*: \_\_\_\_\_

How many executives earn an unusually high or low package (**more than 3 sds** above or below the average)? \_\_\_\_\_

2. Find the sample mean and standard deviation, excluding outliers, and then make a histogram of *compensation* in financial firms.

Average *compensation*, excluding outliers: \_\_\_\_\_

Excluding outliers first identified, how many executives earn an unusually high or low package? \_\_\_\_\_

3. Find the
  - i. *sample mean*
  - ii. *standard deviation,*
  - iii. *25% compensation value*
  - iv. *median, and*
  - v. *75% compensation value,*

excluding outliers, then make a histogram of *compensation* in financial firms to confirm that your sample is typical.

Average *compensation* among typical large financial firms: \_\_\_\_\_

25% of executives in typical, large financial firms earn less than: \_\_\_\_\_

25% of executives in typical, large financial firms earn more than: \_\_\_\_\_

Half of executives in typical, large financial firms earn between \_\_\_\_\_ and \_\_\_\_\_

4. Make a PivotTable and PivotChart from the histogram table excluding outliers. Compare the distribution of *compensation* in the financial sector with a *Normal* distribution with the same mean and standard deviation. (See your text for a similar example in Excel.)

<i>Compensation (sds from mean)</i>	<i>% if Normal</i>	<i>Actual %</i>
<-3 (outliers)		
< -2		
<-1		
< mean		
<+1		
<+2		
<+3		

How does the actual distribution differ from *Normal*?

---

What can the Board say to the CEO to describe the \$2MM package proposal?

---

One Board member has heard rumors that American Express, a competitor, may try to hire the CEO. Will the \$2MM package be competitive? **Y or N**

### *Hollywood Politics*

Managers of a political campaign are considering launch of an effort to attract Hollywood celebrity endorsements.

- Summarize public opinion of celebrity endorsements reported in a recent CBS News/ New York Times poll. Data are in **Lab 2 Hollywood Politics.xls**.

What percent of Republicans prefer celebrities to stay out of politics? \_\_\_\_\_ to \_\_\_\_\_ %

What percent of Democrats prefer celebrities to stay out of politics? \_\_\_\_\_ to \_\_\_\_\_ %

- Make a PivotTable and PivotChart (column chart) comparing the percents of Republicans, Democrats and Independents who prefer celebrities to stay out of politics. Add a label that summarizes poll results. (See your text for a similar example in Excel.)

## *Assignment 2-1 Procter & Gamble's Global Advertising*

Procter & Gamble spent \$5,960,000 on advertising in 51 global markets in 2003. This data, from *Advertising Age*, Global Marketing, 2004 edition, is in **Assignment 2-1 P&G Global Advertising.xls**.

P&G Corporate is reviewing the firm's global advertising strategy, which is the result of decisions made by many brand management teams. Corporate wants to be sure that these many brand level decisions produce an effective allocation when viewed together.

Describe *Procter & Gamble's advertising* spending across the 51 *countries* that make up the global markets.

- Identify *countries* which are **outliers**. Does P&G spend a lot more or a lot less in these markets?
  - Find the sample mean and standard deviation, then use those to make a histogram.
  - Sort the *countries* by *advertising*, then recalculate the sample mean and standard deviation and make a second histogram, **excluding outliers**.
  - Repeat the process of removing outliers and updating the sample mean, standard deviation and histogram until (i) there are **no more outliers**, or (ii) you have excluded **10% of the sample**.
- Is advertising distributed *Normally* across *countries*?
  - After excluding **as many as 10% as outliers**, create a chart of histogram of percentages.  
Compare the percents that are one and two standard deviations above and below the mean with the percents you would expect from a *Normal* distribution and describe what you find.
- Summarize your analysis by describing *P&G's advertising* in *countries* around the world, excluding outliers.  
Include
  - one or more measures of central tendency, such as the mean and median,
  - one or more measures of dispersion, such as the standard deviation and range,
  - the similarity of the distribution to a *Normal* distribution
- Which advertising strategy describes the P&G strategy better: (i) advertise at a moderate level in many global markets, (ii) advertise heavily to a small number of key markets and spend a little in many other markets.

## *CASE 2-1 VW Backgrounds*

Volkswagon management comissioned background music for the New Beetle commercials. The advertising message is that the New Beetle is unique. . . “round in a world of squares.” To be effective, the background music must support this message.

Thirty customers were asked to write down the first word that came to mind when they listened to background music featured in Volkswagon’s Beetle commercials. The music clip is in **Case 2-1 VW background.wav** and words that they wrote are contained in **Case 2-1 VW backgrounds.xls**. Listen to the clip, then describe market response.

- Create a PivotTable of the percent who associate each image with the music
- Sort the PivotTable rows so that the modal image is first
- Create a PivotChart to illustrate the images associated with the background music.
- What is the modal image created by the VW commercial’s background music?

Is this music is a good choice for the VW commercial? Explain.

### 3

## Hypothesis Tests, Confidence Intervals and Simulation to Infer Population Characteristics and Differences

We study a sample to estimate population characteristics. Chapter Three explores the practice of *inference*: how we reliably test *hypotheses* about what may be true in the population and estimate population statistics with *confidence intervals*. Included in this chapter are tests of hypotheses and confidence intervals for

- (i) a population mean from a single sample,
- (ii) the difference between means of two populations, or segments from two independent samples, and
- (iii) the mean difference within one population between two time periods or two scenarios from two matched or paired samples.

In some cases, it is useful to simulate random samples using decision makers' assumptions about a population, to estimate demand and its sensitivity to those assumptions. Monte carlo simulation is introduced in this chapter.

### 3.1 Sample Means Are Random Variables

*Example 3.1 Thirsty on Campus: Is there Sufficient Demand?* An enterprising New Product Development class has an idea to sell on campus custom-flavored, enriched bottles of water from dispensers which would add customers' desired vitamins and natural flavors to each bottle. To assess profit potential, they need an estimate of demand for bottled water on campus. If demand exceeds the breakeven level of seven bottles per week per customer, the business would generate profit. Each of the fifteen student teams in the class independently surveyed a sample of thirty consumers from the campus and then calculated the sample mean and standard deviation from their sample. Team 1, for example, found that average demand in their sample is 11.2 bottles per week, with standard deviation of 4.5 bottles. Each of team's descriptive statistics from the fifteen samples is shown in Figure 3.1.

Teams' Sample Statistics		
Student Research Team	Average Demand per consumer per week $\bar{X}_i$	Standard deviation $s_i$
1	11.2	4.5
2	10.9	4.0
3	10.6	4.3
4	9.5	3.4
5	9.0	3.9
6	10.8	4.6
7	9.6	3.8
8	9.9	4.1
9	9.7	3.7
10	10.7	4.2
11	9.0	3.8
12	9.8	3.6
13	10.5	3.1
14	12.2	4.9
15	11.6	4.2

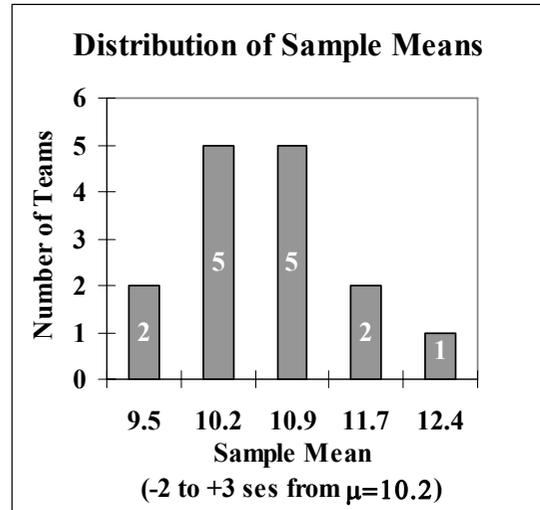


Figure 3.1 Fifteen teams' samples

Each team's sample mean  $\bar{X}$  is close to the true, unknown, population mean,  $\mu=10.2$ . Each of the sample standard deviations is close to the true, unknown population standard deviation  $\sigma=4$ . But each team's sample provides slightly different statistics. Sample means are approximately Normal around the unknown population mean. Sample means will be approximately Normal if "large" ( $N \geq 30$ ) random samples are drawn.

On average, across all random samples of the same size  $N$ , the average difference between sample means and the population mean is the standard error of sample means:

$$s_{\bar{X}} = \sigma / \sqrt{N}$$

where  $\sigma$  is the standard deviation in the population, and  $N$  is the sample size.

Since the population standard deviation is almost never known, but estimated from a sample, the standard error is also estimated from a sample, using the estimate of the population standard deviation  $s$ :

$$s_{\bar{X}} = s / \sqrt{N}$$

When the standard deviation is estimated from a sample (which is nearly always), the distribution of standardized sample means  $\bar{X} / s_{\bar{X}}$  is distributed as *Student t*, which is

approximately Normal. *Student t* has slightly fatter tails since we are estimating the standard deviation. This makes more of a difference if sample size is small. For sample sizes of about thirty or more, there is little difference between Student t and Normal. Our estimate of the standard deviation from the sample becomes closer to the true population value once the sample size meets or exceeds thirty, as shown in Figure 3.2.

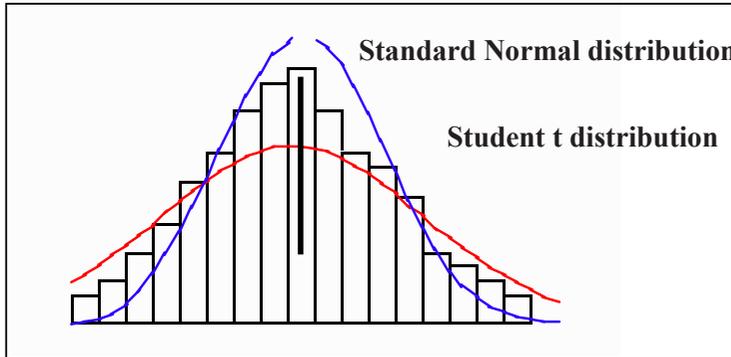


Figure 3.2 Distribution of sample means

With random samples of thirty, population mean  $\mu=10.2$  and standard deviation  $\sigma=4.0$ , the sampling standard error would be  $s_{\bar{x}} = \sigma / \sqrt{30} = 4 / 5.5 = .73$ . From the Empirical Rule introduced in Chapter 2, we would expect 2/3 of the teams' sample means to fall within one standard error of the population mean:

$$\begin{aligned}\mu - s_{\bar{x}} &\leq \bar{X} \leq \mu + s_{\bar{x}} \\ 10.2 - .73 &\leq \bar{X} \leq 10.2 + .73 \\ 9.5 &\leq \bar{X} \leq 10.9,\end{aligned}$$

and we expect 95% of the teams' *sample means* to fall within two standard errors of the population mean:

$$\begin{aligned}\mu - 2s_{\bar{x}} &\leq \bar{X} \leq \mu + 2s_{\bar{x}} \\ 10.2 - 2(.73) &\leq \bar{X} \leq 10.2 + 2(.73) \\ 8.7 &\leq \bar{X} \leq 11.7\end{aligned}$$

We expect nearly all of sample means to fall within three standard errors of the mean, 8.0 to 12.4. Sample means across the fifteen teams ranged from 9.0 to 12.2 bottles per week per consumer.

Rearranging the Empirical Rule formula, we see that *Student t* counts the standard errors between a sample mean and the population mean:

$$|\bar{X} - \mu| / s_{\bar{X}} = t_{N-1}$$

### 3.2 Use Sample Data to Determine Whether Or Not $\mu$ Is Likely To Exceed A Target

Sample statistics can be used to test hypotheses about the population mean or proportion. In the bottled water example, the entrepreneurial class needs to know whether or not demand exceeds seven bottles per consumer per week, because below this level of demand, revenues wouldn't cover expenses.

Hypotheses are formulated as *null* and *alternative*. In this case, the null hypothesis states a limiting conclusion about the population mean. This default conclusion is accepted unless the data indicate that it is highly unlikely.

In the **Thirsty** example, the null hypothesis is a conclusion of insufficient demand, which would lead the class to stop development:

$H_0$ : Campus consumers drink no more than seven bottles of water per week on average:

$$\mu \leq 7$$

Unless sample data indicates sufficient demand, the class will stop development.

In this case, the alternative hypothesis states a conclusion that the population mean exceeds the qualifying condition. The alternative hypothesis is accepted only with sufficient evidence from a sample that the null hypothesis is unlikely to be true.

In **Thirsty**, the alternate hypothesis concludes that population demand is sufficient and would lead to a decision to proceed with the new product's development:

$H_1$ : Campus consumers drink more than seven bottles of water per week on average:

$$\mu > 7$$

Given sufficient demand in a sample, the class would accept the alternate hypothesis and proceed with the project.

Sample statistics are used to determine whether or not the population mean is likely to be less than seven, using the sample mean as the estimate. We ask, "How likely is it that we would observe this sample mean, were the population mean seven or less?" From the

Empirical Rule, we know that sample means are within approximately two standard errors of the population mean 95% of the time. A difference between a sample mean and the break-even level of seven that is more than approximately two standard errors ( $t > 2$ ) is a signal that population demand is unlikely to be seven or less.

In the **Thirsty** example, each team would first calculate the number of standard errors by which their sample mean exceeds seven. Next, each would refer to a table of Student  $t$  values or their statistical software to find the area under the right distribution tail, called the  $p$  value or significance level. Were true demand less than seven, it would be unusual to observe a sample mean more than  $t_{.05; 29} = 1.7$  standard errors greater than seven. The larger a  $t$  value, the smaller the corresponding  $p$  value will be, and the less likely the sample statistics would be observed were the null hypothesis true:

$p \text{ value} > .05$  . . . if the null hypothesis were true, it would not be unusual to observe the data.

The conclusion of insufficient demand  $H_0$  cannot be rejected.

The Team recommends halting development.

$p \text{ value} \leq .05$  . . . if the null hypothesis were true, it would be unusual to observe the data.

Reject the null hypothesis and accept the alternate conclusion  $H_1$  of sufficient demand.

The Team recommends proceeding with development.

Each team used software to test the hypothesis that demand exceeds seven. Team 8's analyses are in Figure 3.3, as an example:

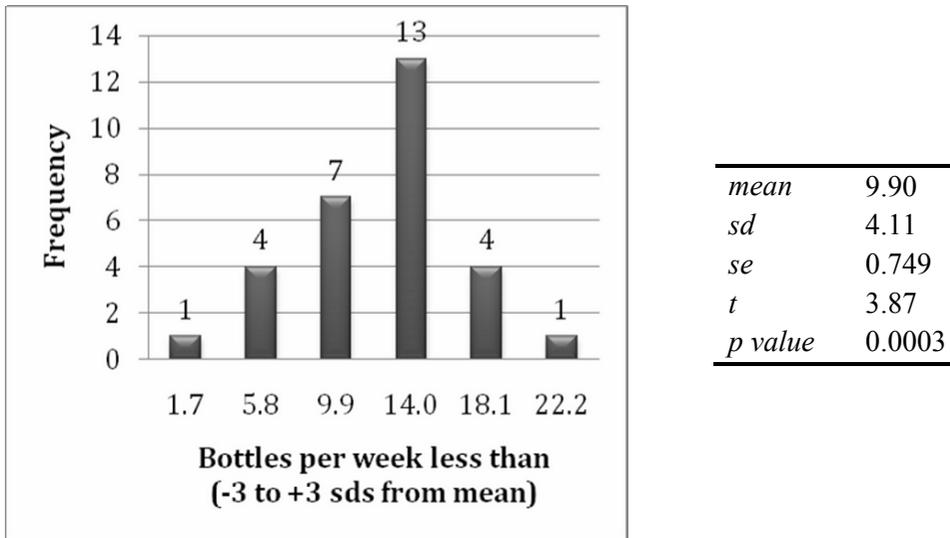


Figure 3.3 *t* test of the hypothesis that population demand is seven or less

Reviewing these results Team Eight would conclude:

*Demand in our sample of thirty ranged from zero to nineteen bottles per person per week, averaging 9.9 bottles per person per week. With this sample of thirty, the standard error is .75 bottles per week. Our sample mean is 3.9 standard errors greater than breakeven of seven. (The *t* statistic is 3.9.) Were population demand seven or less, it would be unusual to observe demand of 9.9 in a sample of thirty. The *p*-value is .0003. We conclude that demand is not seven or less. Sample evidence suggests that demand exceeds seven bottles per person per week.*

In a test of the level of demand for bottles of water, each team used a “one-tail” test. Regardless of how much demand exceeds seven bottles per consumer per week, a team would vote to proceed with development as long as they can be reasonably sure demand exceeds breakeven. They require only that the chance of observing the data be less than 5%, were true demand less than seven. We can then be at least 95% ( $= 1 - p$  value) certain that the true demand is not insufficient. Thus, it is only the area under the right tail that concerns us.

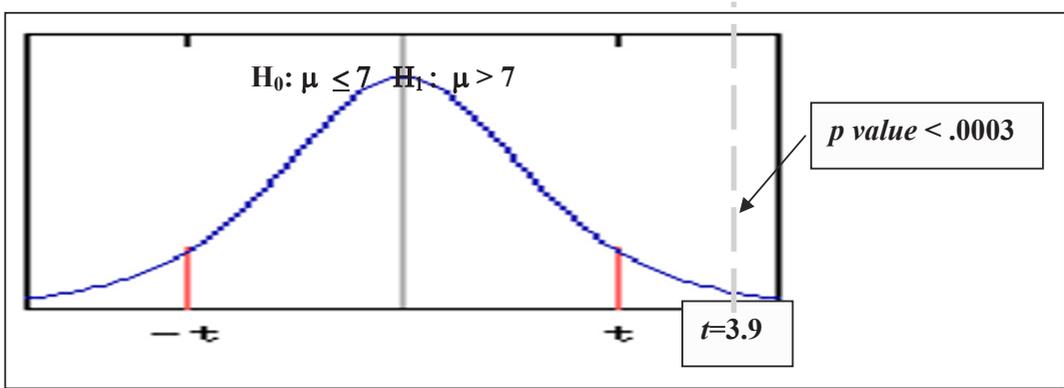


Figure 3.4 *t* test of population mean difference from seven

### 3.3 Confidence Intervals Estimate the Population Mean From A Sample

Since the class of entrepreneurs in the **Thirsty** example doesn't know that the population mean is 10.2 bottles per customer per week, each team will estimate this mean using their sample data. Rearranging the formula for a *t* test, we see that each team can use their sample standard error, the Student *t* value for their sample size and the desired level of confidence to estimate the range that is likely to contain the true population mean:

$$\bar{X} - t_{\alpha, N-1} s_{\bar{x}} < \mu < \bar{X} + t_{\alpha, N-1} s_{\bar{x}}$$

Where  $\alpha$  is the chance that a sample is drawn from one of the sample distribution tails,  
 and  $t_{\alpha, (N-1)}$  is the particular *Student t* value for a chosen level of certainty  $(1-\alpha)$  and sample size  $N$ .

The *confidence level*  $(1-\alpha)$  allows us to specify the level of certainty that an interval will contain the population mean. Generally, decision makers desire a 95% level of confidence  $(\alpha=.05)$ , insuring that in 95 out of 100 samples, the interval would contain the population mean. The Student *t* value for 95% confidence with a sample of thirty  $(N=30)$  is  $t_{\alpha, (N-1)=29} = 2.05$ . In 95% of random samples of thirty drawn, we expect the sample means to be no further than 2.05 standard errors from the population mean:

$$\bar{X} - 2.05s_{\bar{x}} \leq \mu \leq \bar{X} + 2.05s_{\bar{x}}$$

Each team's sample standard error and 95% confidence interval from the **Thirsty** example are shown in Table 3.1:

student team $i$	average demand/ consumer/week, $\bar{X}_i$	standard deviation $s_i$	standard error $s_{\bar{x}}$	margin of error $2.05 \times s_{\bar{x}}$	95% confidence interval $\bar{X} \pm 2.05s_{\bar{x}}$
1	11.2	4.5	0.84	1.61	9.59 12.81
2	10.9	4.0	0.74	1.43	9.47 12.33
3	10.6	4.3	0.80	1.54	9.06 12.14
4	9.5	3.4	0.63	1.22	8.28 10.72
5	9.0	3.9	0.72	1.40	7.60 10.40
6	10.8	4.6	0.85	1.65	9.15 12.45
7	9.6	3.8	0.71	1.36	8.24 10.96
8	9.9	4.1	0.75	1.47	8.43 11.37
9	9.7	3.7	0.69	1.32	8.38 11.02
10	10.7	4.2	0.78	1.50	9.20 12.20
11	9.0	3.8	0.71	1.36	7.64 10.36
12	9.8	3.6	0.67	1.29	8.51 11.09
13	10.5	3.1	0.58	1.11	9.39 11.61
14	12.2	4.9	0.91	1.75	10.45 13.95
15	11.6	4.2	0.78	1.50	10.10 13.10

Table 3.1 Confidence intervals from each team's sample

In practice, we would not collect fifteen samples. We would collect a single sample, just as each individual team did in their market research. Team 8's analysis is shown in Table 3.2 as an example:

<i>mean</i>	9.90	Team 8 would conclude: "Average demand in our sample of thirty is 9.9 bottles per person per week. It is likely that average campus demand is between 8.4 and 11.4 bottles per person per week."
<i>standard error</i>	0.749	
<i>critical t</i>	2.05	
<i>margin of error</i>	1.47	
<i>95% lower</i>	8.43	
<i>95% upper</i>	11.37	

Table 3.2 Confidence interval for bottled water demand  $\mu$

### 3.4 Round $t$ to Calculate Approximate 95% Confidence Intervals With Mental Math

When the sample size is “large,”  $N \geq 30$ , we can use an approximate  $t_{.05; (N-1)} \cong 2.0$  to produce approximate confidence intervals with mental math. Using  $t_{.05; 29} \cong 2$  for an approximate 95% level of confidence, the fifteen student teams each calculated the likely ranges for bottled water demand in the population, shown in Table 3.3.

student team $i$	average demand/ consumer/ week $\bar{X}_i$	standard error $s_{\bar{X}}$	margin of error $2.05 s_{\bar{X}}$	95% confidence interval $\bar{X} \pm 2.05 s_{\bar{X}}$		approximate margin of error $2s_{\bar{X}}$	approximate 95% confidence interval $\bar{X} \pm 2s_{\bar{X}}$	
1	11.2	0.84	1.71	9.5	12.9	1.67	9.5	12.9
2	10.9	0.74	1.52	9.4	12.4	1.49	9.4	12.4
3	10.6	0.80	1.64	9.0	12.2	1.60	9.0	12.2
4	9.5	0.63	1.29	8.2	10.8	1.26	8.2	10.8
5	9.0	0.72	1.48	7.5	10.5	1.45	7.6	10.5
6	10.8	0.85	1.75	9.0	12.6	1.71	9.1	12.5
7	9.6	0.71	1.45	8.2	11.0	1.41	8.2	11.0
8	9.9	0.75	1.50	8.4	11.4	1.52	8.4	11.4
9	9.7	0.69	1.41	8.3	11.1	1.37	8.3	11.1
10	10.7	0.78	1.60	9.1	12.3	1.56	9.1	12.3
11	9.0	0.71	1.45	7.6	10.4	1.41	7.6	10.4
12	9.8	0.67	1.37	8.4	11.2	1.34	8.5	11.1
13	10.5	0.58	1.18	9.3	11.7	1.15	9.4	11.7
14	12.2	0.91	1.87	10.3	14.1	1.82	10.4	14.0
15	11.6	0.78	1.60	10.0	13.2	1.56	10.0	13.2

Table 3.3 Each Team’s Approximate Confidence Interval

With the approximation, Team 8’s conclusion remains: expected demand will range from 8.4 to 11.4 bottles per week per customer.

### 3.5 Margin of Error Is Inversely Proportional To Sample Size

The larger our sample  $N$  is, the smaller our 95% confidence interval is,

$$\bar{X} - 2s_{\bar{X}} \leq \mu \leq \bar{X} + 2s_{\bar{X}}$$

since the standard error  $s_{\bar{X}}$  and margin of error, roughly  $2s_{\bar{X}}$  are inversely proportional to the square root of our sample size  $N$ , shown in Table 3.4.

<i>Sample Size</i> $N$	<i>Approximate Margin of Error</i> $2s / \sqrt{N}$
25	.4s
100	.2s
400	.1s

To double precision, we must quadruple the sample size. Gains in precision become increasingly more expensive.

Table 3.4 Margin of error, given sample size

### 3.6 Samples Are Efficient

We rely on samples to estimate population statistics because it is often neither possible nor feasible to measure all population elements. The time and expense involved in identifying and measuring all elements is prohibitive. To survey the bottled water consumption of each faculty member, student, and staff member on campus would take many hours. We accept an estimate of demand inferred from a random, representative sample which includes faculty, students, and staff. Though we know that our estimates will not be exactly the same as population statistics because of sampling error, samples are amazingly efficient if properly drawn and representative of the population.

### 3.7 Use Monte Carlo Simulation with Sample Statistics To Incorporate Uncertainty and Quantify Implications Of Assumptions

The Team 8 partners were concerned that they might either pass up a profitable opportunity or invest in an unprofitable business. Their estimate of average bottles of water demanded per customer per week seemed promising, though there was a fairly large difference between breakeven and the profit they felt necessary to warrant the investment.

Demand depended on bottles per customer, as well as share of bottles sold on campus. They were unsure whether they would be successful in capturing five percent share of bottled water sold on campus, but this was the best estimate.

With their sample estimate of demand and their assumptions about demand and market share, they want to know the chances that demand would exceed 500,000 bottles in the first year.

The Team decided to use a Monte Carlo simulation to incorporate both demand and share uncertainty and their assumptions into their forecast and decision. Results will show the outcomes under their assumptions.

In a spreadsheet, they will specify the links between demand, market share and bottles sold, and they will use their sample statistics to specify the hypothetical demand distribution. They will use their judgment to specify the hypothetical range of shares thought possible. The simulation will take these inputs and draw a sample of one thousand random hypothetical levels of demand and market share from distributions specified by the Team. Each pair of demand and share values in the simulated sample will feed into the bottles sold worksheet. The Team will then have a better idea of the possible profit levels attainable from the proposed business, given their assumptions.

**Demand, Market Share and Net Profit.** The Team constructed a demand worksheet, highlighting uncertainties, demand and market share, as well as the key performance measure, bottles sold, shown in Table 3.5.

<i>Bottles/customer/week</i>	9.90
<i>Share</i>	5%
<i>Bottles sold (K)</i>	579

Table 3.5 Worksheet for Bottled Water Demand

The Team input their assumptions regarding the distribution of demand using their sample statistics.

**Demand assumptions.** They assumed that

- demand for bottled water was Normally distributed,
- there was a 90% chance that demand would be greater than 8.7 and less than 11.1, and
- 9.9, their sample mean, was the most likely level.

Crystal Ball allows input of 5% and 95% values to specify assumptions about a distribution. This is a 90% (=95%-5%) confidence interval. Since The Team is using their sample statistics to specify assumptions, they will use the 90% *lower* and *upper* confidence limits, shown in Figure 3.5.

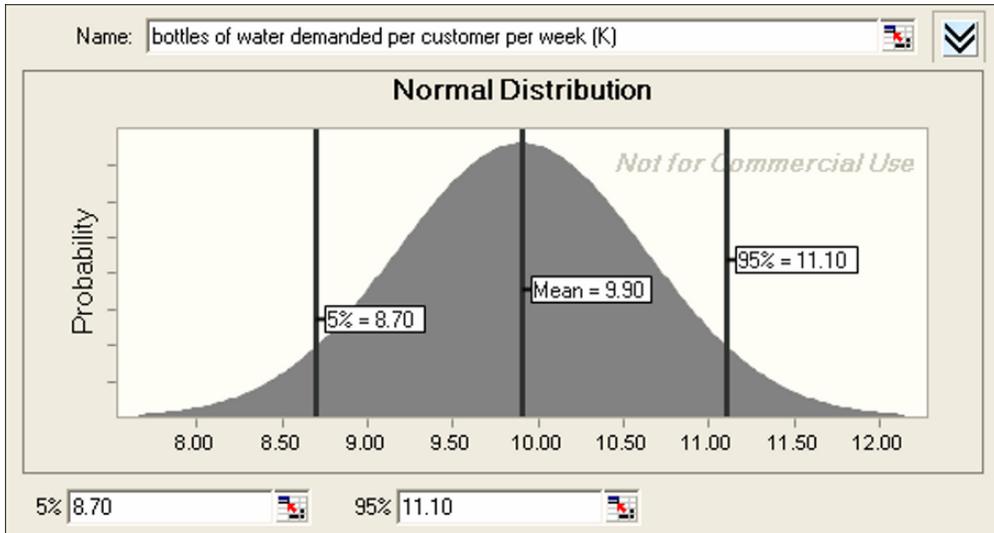


Figure 3.5 Demand assumptions

**Share assumptions.** The Team thought five percent was the most likely market share that could be achieved, though they felt that market share could be as low as two percent or as high as fifteen percent. They chose a triangular distribution for share, shown in Figure 3.6.

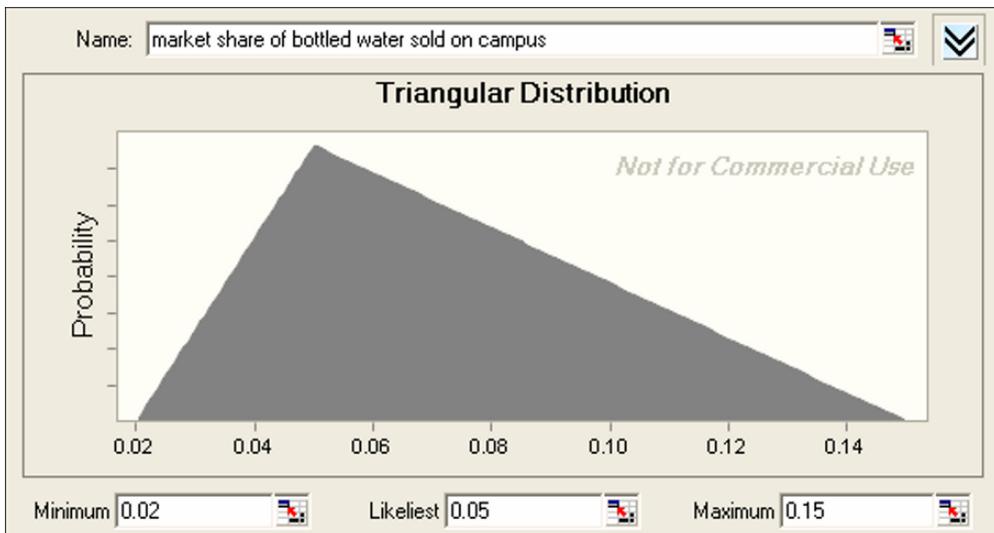


Figure 3.6 Share assumptions

Crystal Ball offers a selection of distributions to match assumptions. When sample data are used to specify assumptions, a Normal distribution can be assumed. When assumptions are based on judgment or expert opinion, a triangular distribution is often used. With less information (no sample), we can use the triangular distribution with *minimum*, *likeliest*, and *maximum* assumptions.

The simulation made one thousand random draws from the assumed distributions of bottles per customer per week and share, which were combined in the demand worksheet.

If the assumed distributions are valid, there is a 86% chance that the Team will sell at least 500,000 bottles in the first year, shown in Figure 3.7

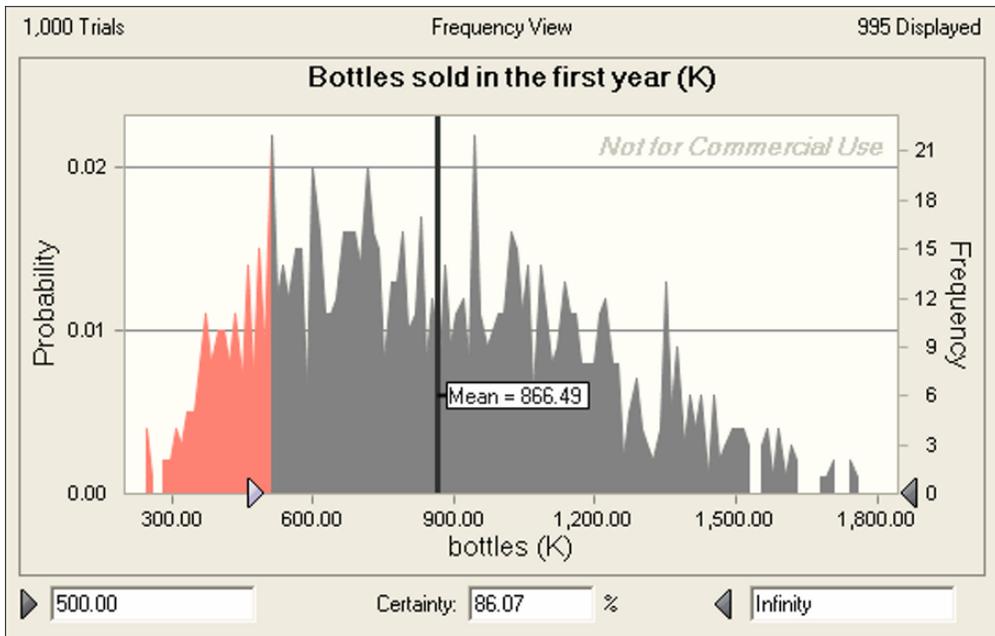


Figure 3.7 Simulated distribution of demand

Since the minimally acceptable level of 500,000 bottles seemed likely, given The Team’s assumptions, The Team is more confident that the potential demand warrants their investment.

### 3.8 Determine Whether There Is a Difference Between Two Segments With Student t

#### *Example 3.2 Pampers Preemies: Is Income a Useful Base for Segmentation?*

Procter & Gamble would like to identify the demographic segment with the highest demand for its new preemie diaper concept. Ninety-seven mothers of premature infants were surveyed and asked to indicate the likelihood that they would try the new diapers if they were available at a (premium) price of \$.36. Fifty-six of the mothers intend to try the new diapers and forty-one do not. Since the new diaper concept is priced relatively high, the Likely Triers may have higher incomes.

Procter & Gamble needs to determine whether or not income is a useful demographic indicator of interest. The null hypothesis states the conclusion that the average annual household income of Likely Triers is not greater than that of Unlikely Triers:

$H_0$ : Average annual household income of Likely Triers is equal to or less than that of Unlikely Triers of the new product concept.

$$\mu_{LT} \leq \mu_{UT}$$

OR

$$|\mu_{LT} - \mu_{UT}| \leq 0$$

Alternatively:

$H_1$ : Average annual household incomes of Likely Triers exceeds that of Unlikely Triers of the new product concept:

$$\mu_{LT} > \mu_{UT}$$

OR

$$|\mu_{LT} - \mu_{UT}| > 0.$$

If there is no difference in incomes between the two segments, or if Likely Triers earn lower incomes, the null hypothesis would be supported by the data.

A test of the significance of the difference between the two segments' average annual household incomes is based on the difference between the two sample means,

$$\bar{X}_{LT} - \bar{X}_{UT}, \text{ and the standard error of the difference } s_{\bar{X}_{LT} - \bar{X}_{UT}}.$$

The standard error of average difference in annual household income (in thousands) is:

$$s_{\bar{X}_{LT} - \bar{X}_{UT}} = \sqrt{s^2_{\bar{X}_{LT}} / N_{LT} + s^2_{\bar{X}_{UT}} / N_{UT}} = \sqrt{[2,300/41 + 2,670/56]} = \$10.2$$

The number of standard errors of difference between sample means is measured with Student t:

$$t_{\alpha,(N-1)} = (\bar{X}_{LT} - \bar{X}_{UT}) / s_{\bar{X}_{LT} - \bar{X}_{UT}}$$

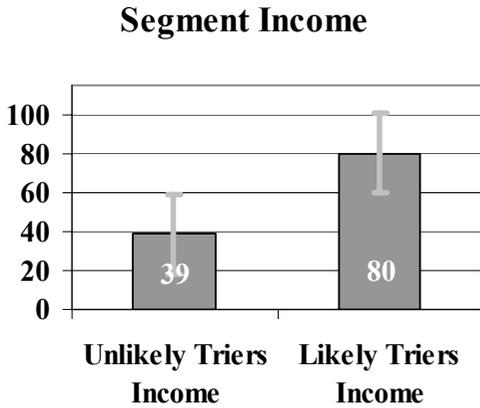


Figure 3.8 Difference between segments

From the *t test* of difference between segment incomes, shown in Figure 3.8, Procter & Gamble brand management could conclude:

*“In our sample of 97, the average incomes of Likely and Unlikely Trier segments are \$80K and \$38K, a difference of \$42K. Were there no difference in segment mean incomes in the population, it would be unusual to observe this difference in segment average incomes in a sample. Based on sample evidence, we conclude that average incomes of Likely Triers exceed the average incomes of Unlikely Triers. Income is a useful basis for segmentation.”*

### 3.9 Estimate the Extent of Difference between Two Segments With Student t

From the sample data, market researchers estimate the average annual household income difference (in thousands) between Likely and Unlikely Triers:

$$\bar{X}_{LT} - \bar{X}_{UT} = \$80.1 - \$38.5 = \$41.6$$

The approximate 95% confidence interval of the difference in annual household incomes between Likely and Unlikely Triers is:

$$(\bar{X}_{LT} - \bar{X}_{UT}) - 2s_{\bar{X}_{LT} - \bar{X}_{UT}} \leq (\mu_{LT} - \mu_{UT}) \leq (\bar{X}_{LT} - \bar{X}_{UT}) + 2s_{\bar{X}_{LT} - \bar{X}_{UT}}$$

$$\begin{aligned} \$41.6 - 2 (\$10.2) &\leq (\mu_{LT} - \mu_{UT}) \leq \$41.6 + 2 (\$10.2) \\ \$21.2 &\leq (\mu_{LT} - \mu_{UT}) \leq \$62.0 \end{aligned}$$

Thus, the firm estimates that the average difference in annual household income between Likely and Unlikely Triers is \$21,000 to \$62,000.

Marketing management will conclude that annual household income can be used to differentiate the two market segments, and that Likely Triers are wealthier than Unlikely Triers.

*In our sample of 97, the average difference in income between Likely and Unlikely Trier segments is \$42K, and the standard error of the difference is \$10K. Relative to Unlikely Triers, we estimate that Likely Triers earn \$21,000 to \$62,000 more on average each year.*

To construct confidence intervals for the difference in means of two samples, we assume that either (i) both segments' characteristics are bell-shaped (distributed approximately Normal) and we've randomly sampled both segments, or (ii) "large" random samples from both segments have been collected.

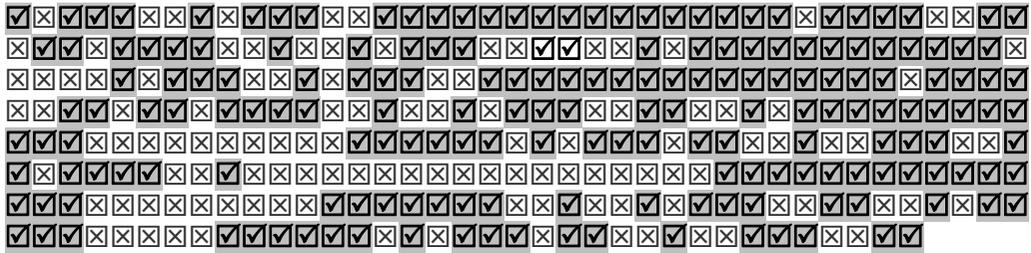
### 3.10 Confidence Intervals Complement Hypothesis Tests

Confidence intervals and hypothesis tests are consistent and complementary, but are used to make different decisions. If a decision maker needs to make a qualitative Yes/No decision, a hypothesis test is used. If a decision maker instead requires a quantitative estimate, such as level of demand, confidence intervals are used. Hypothesis tests tell us whether demand exceeds a critical level or whether segments differ. Confidence intervals quantify demand or magnitude of differences between segments.

### 3.11 Estimation of a Population Proportion from a Sample Proportion

*Example 3.3 Guinea Pigs.* A pharmaceutical company gauges reactions to their products by applying them to animals. An animal rights activist has threatened to start a campaign to boycott the company's products if the animal testing doesn't stop. Concerned managers have hired four public opinion polling organizations to learn whether medical testing on animals is accepted or not.

Below are the opinions in one sample slice of American adults, where 60% agree  that medical testing on animals is morally acceptable:



Four independent pollsters each surveyed thirty Americans and found the proportions shown in Table 3.6 agree that medical testing on animals is morally acceptable:

<i>Poll</i>	<i>Sample Approval Proportion</i>
1	$P_1 = 16 / 30 = .53$
2	$P_2 = 19 / 30 = .63$
3	$P_3 = 17 / 30 = .57$
4	$P_4 = 21 / 30 = .70$

Table 3.6 Sample approval proportions by poll

If numerous random samples are taken, sample proportions  $P$  will be approximately Normally distributed around the unknown population proportion  $\pi=.6$ , as long as this true proportion is not close to either zero or one.

The standard deviation of the sample proportions  $P$ , the *standard error of the sample proportion*, measures dispersion of samples of size  $N$  from the population proportion  $\pi$ :

$$\sigma_{\pi} = \sqrt{\pi(1 - \pi) / N}$$

which we estimate with the sample proportion  $P$ :

$$s_p = \sqrt{P(1 - P) / N}$$

The four poll organizations would each estimate the proportion of Americans who agree that medical testing on animals is morally acceptable, shown in Table 3.7.

<i>Poll</i> <i>i</i>	<i>Sample</i> <i>Proportion,</i> $P_i$	<i>Standard</i> <i>Error,</i> $s_{P_i}$ ( $N=30$ )	<i>Approximate</i> <i>Margin of Error</i> <i>for 95%</i> <i>Confidence,</i> $2s_{P_i}$	<i>Interval containing the</i> <i>Population Proportion with 95%</i> <i>confidence</i> $P_i \pm 2s_{P_i}$
1	0.53	0.091	0.18	0.35 to 0.72
2	0.63	0.088	0.18	0.46 to 0.81
3	0.57	0.090	0.18	0.39 to 0.75
4	0.70	0.084	0.17	0.53 to 0.87

Table 3.7 Confidence interval of approval proportion by poll,  $N=30$ 

We see that with samples of just thirty, margins of error are relatively large and we are uncertain whether a minority or a sizeable majority approves. In practice, polling organizations use much larger samples, which shrink margins of error and corresponding confidence intervals. Had samples of 1,000 been collected instead, the poll results would be as shown in Table 3.8.

<i>Poll</i> <i>i</i>	<i>Sample</i> <i>Proportion,</i> $P_i$	<i>Standard</i> <i>Error,</i> $s_{P_i}$ ( $N=1000$ )	<i>Approximate Margin of</i> <i>Error for 95%</i> <i>Confidence,</i> $2s_{P_i}$	<i>Approximate 95% Confidence</i> <i>Interval</i> $P_i \pm 2s_{P_i}$	
1	0.57	0.016	0.031	0.54	0.60
2	0.61	0.015	0.031	0.58	0.64
3	0.58	0.016	0.031	0.55	0.61
4	0.63	0.015	0.031	0.60	0.66

Table 3.8 Confidence interval of approval proportion by poll,  $N=1000$ 

With much larger samples and correspondingly smaller margins of error, it becomes clear that the majority approves of medical testing on animals.

The second polling organization would report:

*The majority of a random sample of 1,000 Americans approves of medical testing on animals. 61% believe medical testing on animals is morally acceptable, with a margin of error of 3%.*

### 3.12 Conditions for Assuming Approximate Normality to Make Confidence Intervals for Proportions

It is appropriate to use the Normal distribution to approximate the distribution of possible sample proportions if sample size is “large” ( $N \geq 30$ ), and both  $N \times P \geq 5$  and  $N \times (1-P) \geq 5$ . When the true population proportion is very close to either zero or one, we cannot reasonably assume that the distribution of sample proportions is Normal. A rule of thumb suggests that  $P \times N$  and  $(1-P) \times N$  ought to be at least five in order to use Normal inferences about proportions. For a sample of thirty, the sample proportion  $P$  would need to be between .17 and .83 to use Normal inferences. For a sample of 1,000, the sample proportion  $P$  would need to be between .01 and .99. Drawing larger samples allows us to confidently infer population proportions from samples.

### 3.13 Conservative Confidence Intervals for a Proportion

Polling organizations report the sample proportion and margin of error, rather than a confidence interval. For example, “61% approve of medical testing on animals. (The margin of error from this poll is 3 percentage points.)” A 95% level of confidence is the industry standard. Because the true proportion and its standard deviation are unknown, and because pollsters stake their reputations on valid results, a *conservative* approach, which assumes a true proportion of .5, is used. This conservative approach

$$s_p = \sqrt{.5(1-.5)/N}$$

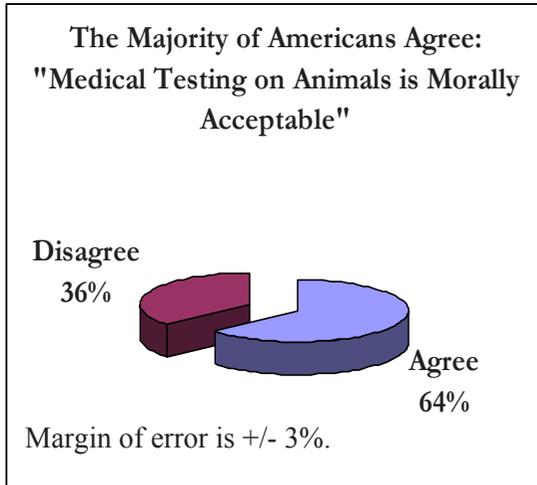
yields the largest possible standard error for a given sample size and makes the margin of error ( $2s_p$ ) a simple function of the square root of the sample size  $N$ .

With this conservative approach and samples of  $N=1,000$ , the pollsters’ results are shown in Table 3.9.

<i>Poll</i> <i>i</i>	<i>Sample</i> <i>Proportion,</i> <i>P</i>	<i>Approximate Conservative</i>	
		<i>Margin of Error for 95% Confidence,</i> $2s_p$	<i>Approximate Conservative 95% Confidence Interval</i> $P - 2s_p \leq \pi \leq P + 2s_p$
1	.57	.032	.54 .60
2	.61	.032	.58 .64
3	.58	.032	.55 .61
4	.63	.032	.60 .66

Table 3.9 Conservative confidence intervals for approval proportions,  $N=1000$

An effective display of proportions or shares is a *pie chart*. The second poll organization used Excel to create this illustration of their survey results, shown in Figure 3.9:



The second polling organization would report:

*"Sixty-one percent of American adults agree that medical testing on animals is morally acceptable. Poll results have a margin of error of 3 percentage points. The majority of Americans supports medical testing on animals."*

Figure 3.9 Pie chart of approval percentage

Other appropriate applications for confidence intervals to estimate population proportions or shares include:

- Proportion who prefer a new formulation to an old one in a taste test
- Share of retailers who offer a brand
- Market share of a product in a specified market
- Proportion of employees who call in sick when they're well
- Proportion of new hires who will perform exceptionally well on the job

### 3.14 Assess the Difference between Alternate Scenarios or Pairs With Student t

Sometimes management is concerned with the comparison of means from a single sample taken under varying conditions—at different times or in different scenarios—or comparison of sample pairs, like the difference between an employee's opinion and the opinion of the employee's supervisor.

- Financial management might be interested in comparing the reactions of a sample of investors to "socially desirable" stock portfolios, excluding stocks of firms that manufacture or market weapons, tobacco, or alcohol, versus alternate portfolios which promise similar returns at similar risk levels, but which are not "socially desirable."

- Marketing management might be interested in comparing taste ratings of sodas which contain varying levels of red coloring—do redder sodas taste better to customers?
- Management might be interested in comparing satisfaction ratings following a change which allows employees to work at home.

These examples compare *repeated samples*, where participants have provided multiple responses that can be compared.

- Financial management might also be interested in comparing the risk preferences of husbands and wives.
- Marketing management might want to compare children and parents' preferences for red sodas.
- Management might also be interested in comparing the satisfaction ratings of those employees with their supervisors' satisfaction ratings.

In these examples, we are interested in comparing means from *matched pairs*.

In either case of repeated or matched samples, we can find the difference and use a *t test* to determine whether or not the difference is non-zero.

*Example 3.4 Are “Socially Desirable” Portfolios Undesirable?* An investment consulting firm's management believes that they have difficulty selling “socially desirable” portfolios because potential investors assume those funds are inferior investments. Socially Desirable funds exclude stocks of firms which manufacture or market weapons, tobacco or alcohol. There may be a perceived sacrifice associated with socially desirable investment which causes investors to avoid portfolios labeled “socially desirable.” The null hypothesis is:

$H_0$ : Investors rate “socially desirable” portfolios at least as attractive as equally risky, conventional portfolios promising equivalent returns:

$$\mu_{\text{SOCIALLY DESIRABLE}} - \mu_{\text{CONVENTIONAL}} \geq 0.$$

If investors do not penalize “socially desirable” funds, the null hypothesis would be supported.

The alternative hypothesis is:

$H_1$ : Investors rate “socially desirable” portfolios as less attractive than other equally risky portfolios promising equivalent returns:

$$\mu_{\text{SOCIALLY DESIRABLE}} - \mu_{\text{CONVENTIONAL}} < 0.$$

Thirty-three investors were asked to evaluate two stock portfolios on a scale of attractiveness (-3 = “Not At All Appealing” to 3 = “Very Appealing”). The two portfolios promised equivalent returns and were equally risky. One contained only “socially desirable” stocks, while the other included stocks from companies which sell tobacco, alcohol and arms. These are shown in Table 3.10.

<i>appeal of conventional portfolio</i>	<i>appeal of socially desirable portfolio</i>	<i>difference in appeal = appeal of conventional – appeal of socially desirable</i>	<i>appeal of conventional portfolio</i>	<i>appeal of socially desirable portfolio</i>	<i>difference in appeal = appeal of conventional – appeal of socially desirable</i>
-3	1	-4	2	-1	3
-3	2	-5	2	-1	3
-3	3	-6	2	-2	4
-3	3	-6	2	2	0
0	-1	1	2	1	1
0	1	-1	2	2	0
1	-3	4	2	2	0
1	-3	4	2	3	-1
1	-1	2	3	-3	6
1	-1	2	3	-3	6
1	-1	2	3	-3	6
1	1	0	3	-1	4
1	1	0	3	-1	4
1	2	-1	3	-3	6
2	-3	5	3	3	0
2	-3	5	3	3	0
2	-2	4			

Table 3.10 Paired ratings of other & socially desirable portfolios

From a random sample of 33 investors’ ratings of conventional and Socially Desirable portfolios of equivalent risk and return, the average difference is 1.5 points on a 7-point scale of attractiveness.

$$\bar{X}_{dif} = \bar{X}_{SD} - \bar{X}_C = -.2 - 1.3 = -1.5$$

With this sample of 33, the standard error of the difference is .6.

$$s_{\bar{X}_{dif}} = s_{dif} / \sqrt{N} = 3.4 / \sqrt{33} = .6$$

The average difference in attractiveness between the Conventional and the Socially Desirable portfolio is 2.5 standard errors:

$$t = \bar{X}_{dif} / s_{\bar{X}_{dif}} = -1.5 / .6 = -2.5$$

The *p* value for *t* = -2.5, for a sample size of 33 is .0097. Were the Socially Desirable portfolio at least as attractive as the Conventional portfolio with equivalent risk and return, it would be unusual to observe such a large sample mean difference in ratings. Based on sample evidence, shown in Figure 3.10, we conclude that a “socially desirable” label reduces portfolio attractiveness.

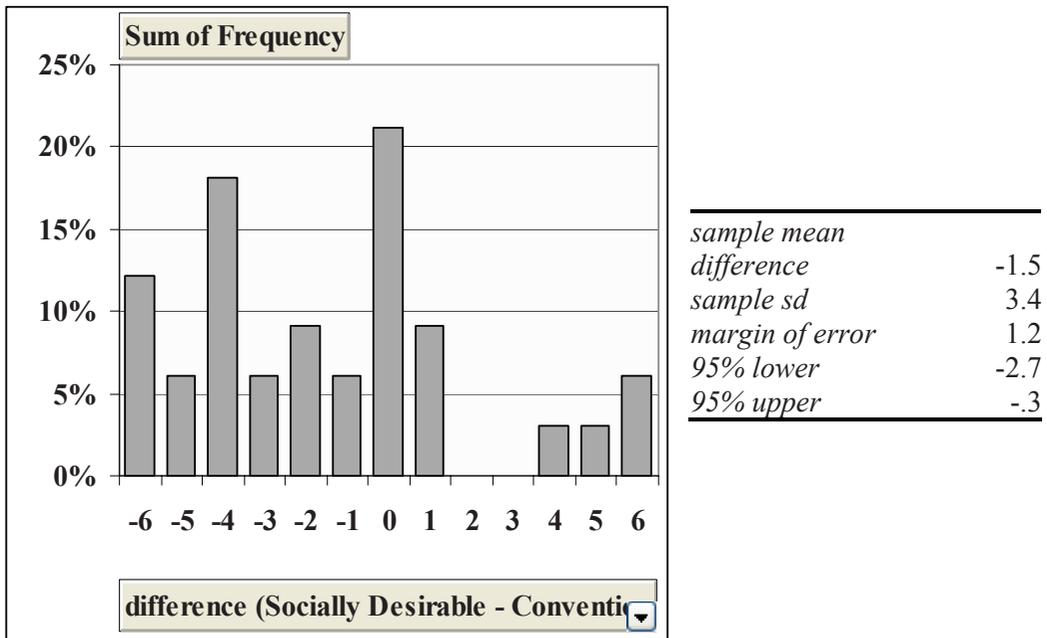


Figure 3.10 *t* test of differences between paired ratings of socially desirable & conventional portfolios

The approximate 95% confidence interval for the difference is

$$\bar{X}_{dif} \pm 2s_{\bar{X}_{dif}}$$

$$-1.5 \pm 2(.6)$$

OR -2.7 to -0.3 on the 7-point scale.

The investment consultants would conclude:

A “socially desirable” label reduces investors’ judged attractiveness ratings. Investors downgrade the attractiveness of “socially desirable” portfolios by about 1 to 3 points on a 7-point scale, relative to other equivalent portfolios.

### 3.15 Inference from Sample to Population

Managers use sample statistics to infer population statistics, knowing that inference from a sample is efficient and reliable. Because sample standard errors are approximately *Normally* distributed, we can use the Empirical Rule to build confidence intervals to estimate population means and to test hypotheses about population means with *t tests*. We can determine whether a population mean is likely to equal, be less than, or exceed a target value, and we can estimate the range which is likely to include a population mean.

Our certainty that a population mean will fall within a sample-based confidence interval depends on the amount of population variation and on the sample size. To double precision, sample size must be quadrupled, because the margin of error is inversely proportional to the square root of sample size.

Differences are important to managers, since differences drive decision making. If customers differ, segments are targeted in varying degrees. If employee satisfaction differs between alternate work environments, the workplace may be altered.

Inference about differences between two populations is similar, and relies on differences between two independent samples. With a *t test*, we can determine whether there is a likely difference between two population means, and with a confidence interval, we can estimate the likely size of difference.

Excel 3.1 Test the level of a population mean with a one sample *t* test

*Thirsty on Campus.* Team 8 wants to know whether the demand for bottled water exceeds a break-even level of 7 bottles per day. We will analyze their data, using a *t* test of *Bottles* purchased per day.

Open **Excel 3.1 Bottled Water Demand.xls**.

Add labels *mean*, *standard deviation*, *standard error*, *t*, *p value* in **A33:A37**, then enter their formulas in **B33:B37**:

In **B33**, enter **=AVERAGE(B2:B31)** [Enter].

In **B34**, enter **=STDEV(B2:B31)** [Enter].

Find the standard error by dividing the sample standard deviation in **B34** by the square root of sample size, **30**:

In **B35**, enter **=B34/SQRT(30)** [Enter].

The screenshot shows an Excel spreadsheet with the following data and formulas:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		<b>bottles</b>											
15		10											
16		10											
17		10											
18		10											
19		10											
20		11											
21		11											
22		11											
23		12											
24		12											
25		13											
26		13											
27		15											
28		15											
29		15											
30		16											
31		19											
32													
33	mean	9.90											
34	sd	4.11											
35	se	0.749											

Find *t* by finding the difference between the sample mean in **B33** and the critical value **7**, then dividing that difference by the standard error in **B35**.

In **B36**, enter **=(B33-7)/B35** [Enter].

Find the  $p$  value for this  $t$  using the Excel function **TDIST(t,df,tails)**, entering the  $t$  in **B36**. For degrees of freedom, **df**, enter the sample size, minus one, **29** (=30-1). For **tails**, enter **1** for a one-tail test:

In **B37**, enter **=TDIST(B36,29,1)** [Enter]:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		bottles											
15		10											
16		10											
17		10											
18		10											
19		10											
20		11											
21		11											
22		11											
23		12											
24		12											
25		13											
26		13											
27		15											
28		15											
29		15											
30		16											
31		19											
32													
33	mean	9.90											
34	sd	4.11											
35	se	0.749											
36	t	3.87											
37	p-value	0.0003											

### Excel 3.2 Make a confidence interval for a population mean

We will determine for Team 8 the range which is likely to contain average demand in the population. We will construct the 95% confidence interval for the population mean *Bottles* demanded.

In **A38:A40**, enter the labels *margin of error*, *95% lower* and *95% upper*.

Use the Excel function **CONFIDENCE(alpha, standard deviation, sample size)** to find the 95% margin of error. For **alpha**, enter **.05** for a 95% level of confidence. For **standard deviation**, enter the sample standard deviation in **B34**, and for **sample size**, enter 30:

In **B38**, enter **=CONFIDENCE(.05, B34, 30)**[Enter].

The **CONFIDENCE** function returns the margin of error in **B38**. Add and subtract this to find the 95% upper and lower confidence interval limits:

In **B39**, enter **=B33-B38[Enter]**.

In **B40**, enter **=B33+B38[Enter]**.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		<b>bottles</b>											
15		10											
16		10											
17		10											
18		10											
19		10											
20		11											
21		11											
22		11											
23		12											
24		12											
25		13											
26		13											
27		15											
28		15											
29		15											
30		16											
31		19											
32													
33	mean	9.90											
34	sd	4.11											
35	se	0.749											
36	t	3.87											
37	p-value	0.0003											
38	margin of error	1.47											
39	95% lower	8.43											
40	95% upper	11.37											
41													

### Excel 3.3 Illustrate population confidence intervals with a clustered column chart

*t-mobile's Service.* t-mobile managers have conducted a survey of customers in 32 major metropolitan areas to assess the quality of service along three key areas: coverage, absence of dropped calls, and static. Customers rated t-mobile service along each of these three dimensions using a five-point scale (1=poor to 5=excellent). Management's goal is to be able to offer service that is not perceived as inferior. This goal translates into mean ratings of at least 3 on the 5-point scale in the national market across all three service dimensions. We will make 95% confidence intervals to estimate the average perceived quality of service.

Open **Excel 3.3 t-mobile.xls**.

**95% Confidence Intervals.** In **B34:B38** type in labels *sample mean, sample standard deviation, margin of error, 95% lower* and *95% upper*.

Find the sample mean and standard deviation:



To see confidence intervals for *dropped call rating* and *static rating*, select **C34:C38**, then use shortcuts to fill in statistics for *dropped call rating* and *static rating*.

**Shift+>** through **E**, **Cntl+R**.

(**Shift+arrow** selects cells scrolled over, and **Cntl+R** fills in values of empty cells using formulas from the first column.)

	A	B	C	D	E	F	G	H	I	J
			coverage rating (1=Poor to 5=Excellent)	dropped calls rating (1=Poor to 5=Excellent)	static rating (1=Poor to 5=Excellent)					
1	city	service								
17	washingtondc	tmobile	1	3	2					
18	birmingham	tmobile	2	3	3					
19	albany	tmobile	1	3	3					
20	cincinnati	tmobile	3	3	3					
21	austin	tmobile	3	3	2					
22	albuquerque	tmobile	3	3	3					
23	fort worth	tmobile	3	4	3					
24	sacramento	tmobile	1	3	3					
25	madison	tmobile	4	5	3					
26	baltimore	tmobile	1	3	2					
27	trenton	tmobile	2	3	3					
28	tucson	tmobile	3	4	3					
29	portland	tmobile	1	3	3					
30	las vegas	tmobile	3	4	4					
31	kansas city	tmobile	3	4	3					
32	miami	tmobile	2	3	4					
33	raleigh	tmobile	1	3	2					
34		sample mean	2.25	3.375	2.9375					
35		sample standard deviation	0.983738754	0.609071213	0.56440091					
36		margin of error	0.340841825	0.211028531	0.195551345					
37		95% lower	1.909158175	3.163971469	2.741948655					
38		95% upper	2.590841825	3.586028531	3.133051345					

**Clustered column chart of confidence intervals.** To see the confidence intervals for all three service dimension ratings, first insert a row above row 37 for chart labels, using shortcuts: select **37**, **Alt HIR**.

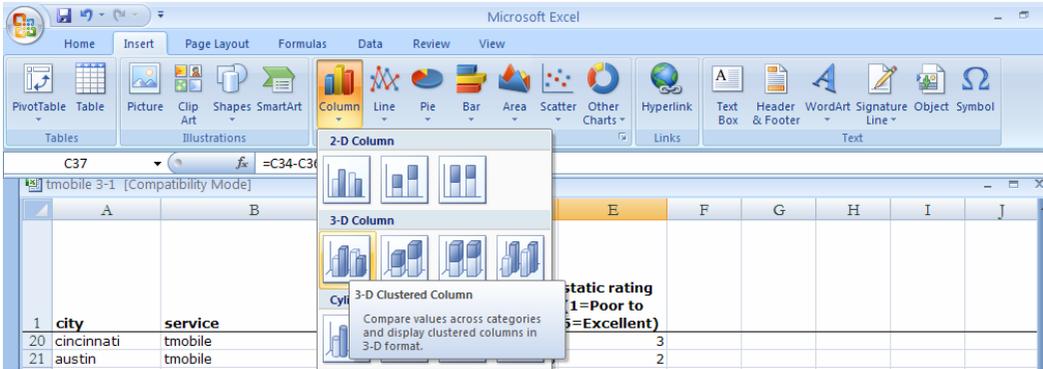
(**Alt** activates shortcuts, **H** selects the Home menus, **I** selects **Insert** menus, and **R** inserts a row.)

Type in labels, *coverage*, *dropped calls*, *static*:

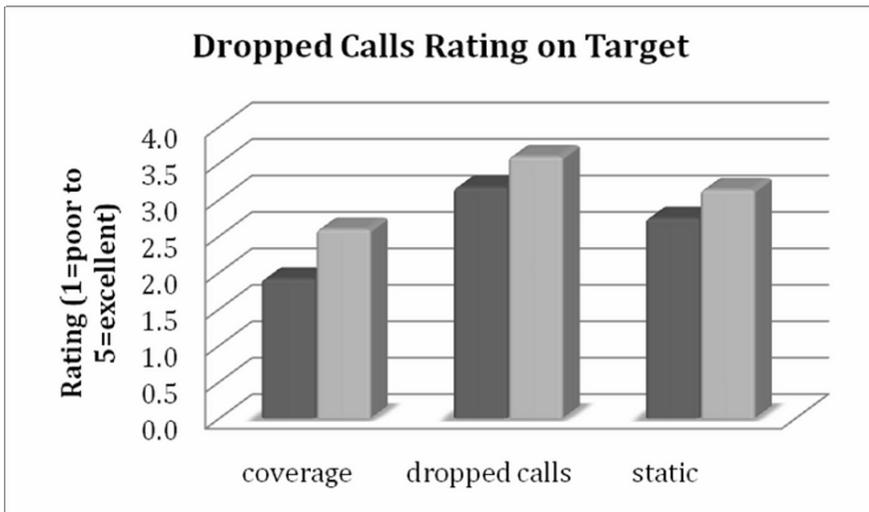
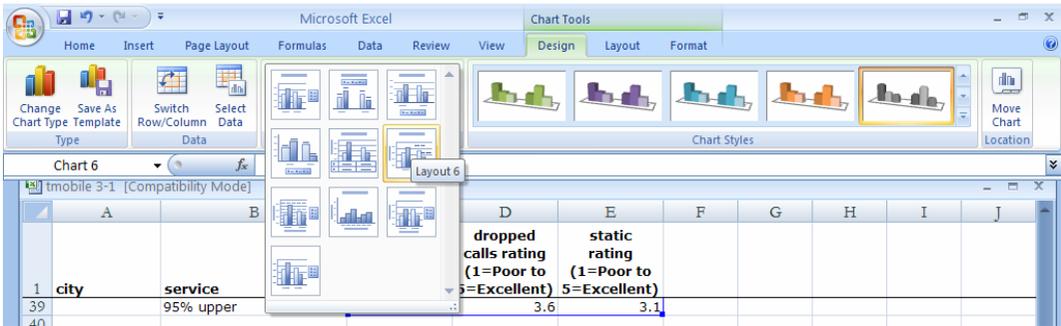
	A	B	C	D	E	F	G	H	I	J
			coverage rating (1=Poor to 5=Excellent)	dropped calls rating (1=Poor to 5=Excellent)	static rating (1=Poor to 5=Excellent)					
1	city	service								
36		margin of error	0.340841825	0.211028531	0.195551345					
37			coverage	dropped calls	static					
38		95% lower	1.9	3.2	2.7					
39		95% upper	2.6	3.6	3.1					

Use shortcuts to make a clustered column chart.

Select **C37:E39**, **Alt NC**, **3-D Clustered Column**.

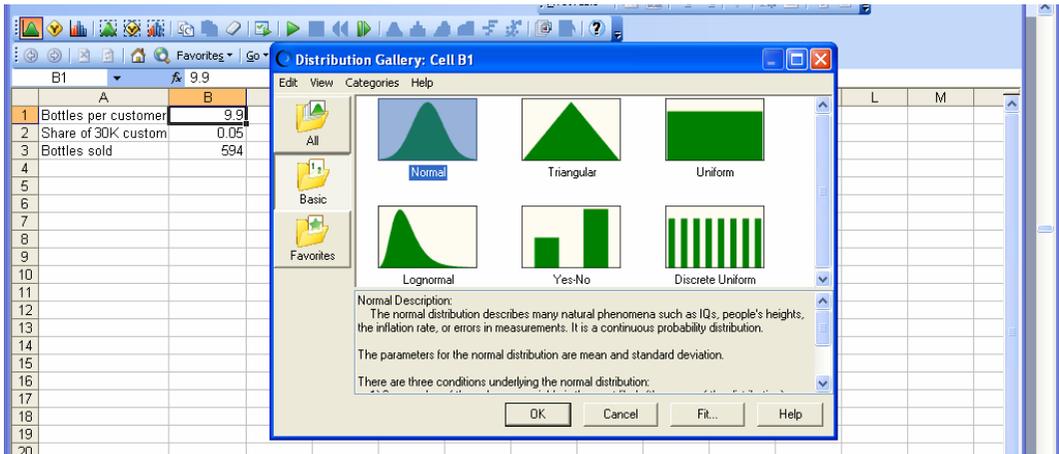


Choose **Design**, **Chart Layout 6** to add a vertical axis label:



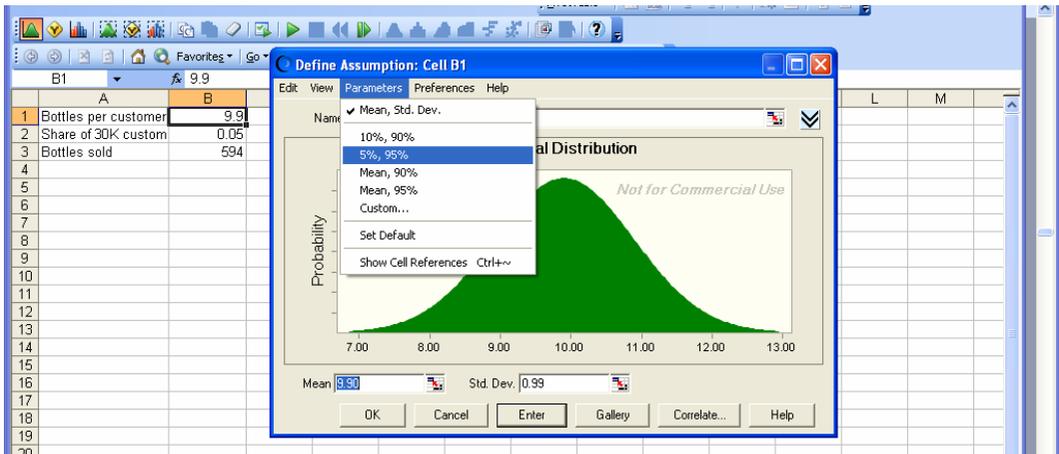


**Define assumptions.** Select **B1**, then choose the assumptions icon.

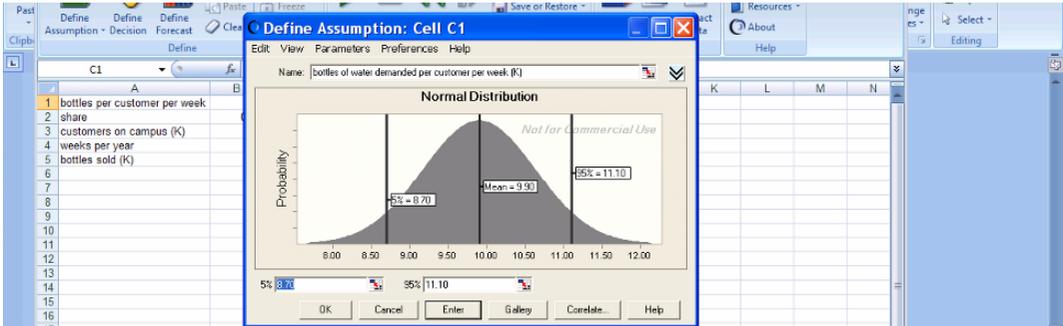


Since we are using sample data to specify assumptions, select **Normal**, **OK**:

We want to specify the distribution center, range and shape with the 90% confidence interval from the sample. Select **Parameters**, **5%, 95%**:

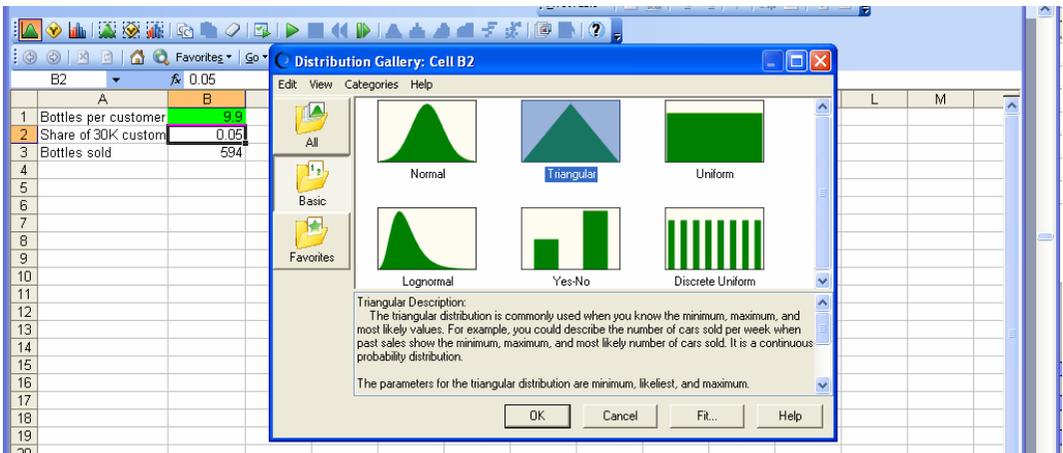


For **5%**, enter the *90% lower* confidence limit from the sample, **8.7** and for **95%**, enter the *90% upper* confidence limit from the sample, **11.1**, **Enter**, **OK**:



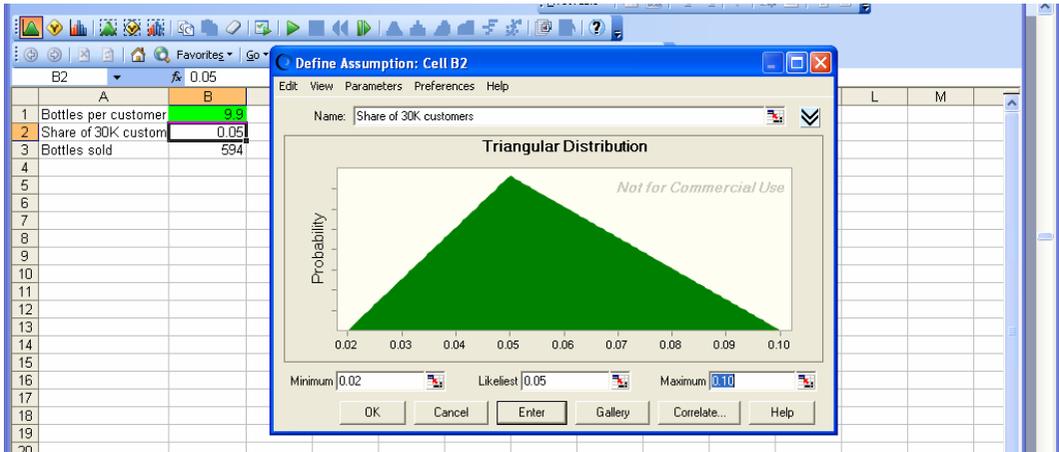
Our market share estimate is based on subjective judgment, and we aren't sure of the shape. We are comfortable specifying the minimum, likeliest, and maximum market share values, so we will assume a triangular distribution.

Select the market share cell, **B2**, then choose the assumptions icon.

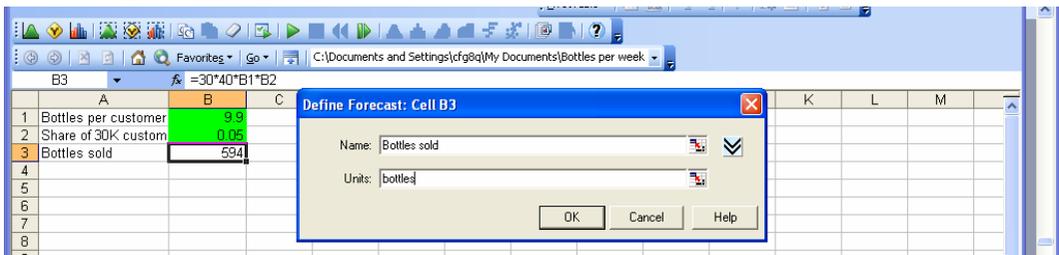


From the **Gallery** choose **Triangular**, **OK**:

Enter the market share assumptions: **Minimum .02, Likeliest, .05, Maximum, .15, Enter, OK.**



**Define forecast.** To record the forecast bottles sold with each demand and market share combination drawn, select the performance outcome cell, *bottles sold*, **B3**, then choose the forecast icon  and enter a name for the forecast and the units.



Start the simulation by selecting the run icon: 



### Excel 3.6 Construct a confidence interval for the difference between two segments

Procter & Gamble would like to estimate with 95% certainty the difference in incomes between the Unlikely and Likely Trier segments.

Open **Excel 3.6 Pampers Segment Income.xls**. At the end of the dataset, add the means and standard deviations of the two segments.

In **A59:A60**, enter labels *mean* and *standard deviation*, and

in **B59**, enter `=AVERAGE(B2:B42)`[Enter].

In **B60**, enter `=STDEV(B2:B42)`[Enter].

In **C59**, enter `=AVERAGE(C2:C57)`[Enter].

In **C60**, enter `=STDEV(C2:C57)`[Enter].

Find the difference between segments and the standard error of the difference.

In **A61:A62**, type in the labels *segment mean difference* and *standard error*.

Find the difference between segment sample means by entering in **F61** `=C59-B59` [Enter].

Find the standard error of the difference by taking the square root of the sum of segment variances (equal to the standard deviations in **B60** and **C60**, squared), each divided by the segment sample size.

In **F62**, enter `=SQRT(B60^2/41+C60^2/56)` [Enter]:

F62		=SQRT(C60^2/41+B60^2/56)													
	A	B	C	D	E	F	G	H	I	J	K	L	1		
1		Likely Triers Income	Unlikely Triers Income	Income (K)	Trier	p value									
56		141		199	likely										
57		156		24	likely										
58															
59	mean	80.14286	38.53659	24	likely										
60	standard deviation	51.67184	47.95628	48	likely										
61	difference			48	likely	41.6062718									
62	standard error			120	likely	10.1868053									

Find the approximate margin of error, which will be twice the standard error, then make the 95% confidence interval for the difference by adding and subtracting the margin of error from the mean difference:

In **A63:A65**, enter the labels *approximate margin of error*, *95% lower*, *95% upper*.

In **F63**, find the approximate margin of error by entering **=2\*F62 [Enter]**.  
 In **F64**, find the 95% lower confidence interval bound by entering **=F61-D63 [Enter]**.  
 In **F65**, find the 95% upper confidence interval bound by entering **=F61+D63 [Enter]**:

	A	B	C	D	E	F	G	H	I	J	K	L
1		Likely Triers Income	Unlikely Triers Income	Income (K)	Trier	p value						
56		141		199	likely							
57		156		24	likely							
58												
59	mean	80.14286	38.53659	24	likely							
60	standard deviation	51.67184	47.95628	48	likely							
61	difference			48	likely	41.6062718						
62	standard error			120	likely	10.1868053						
63	margin of error			24	likely	20.3736106						
64	95% lower			24	likely	21.2326612						
65	95% upper			142	likely	61.9798824						

### Excel 3.7 Illustrate the difference between two segment means with a column chart

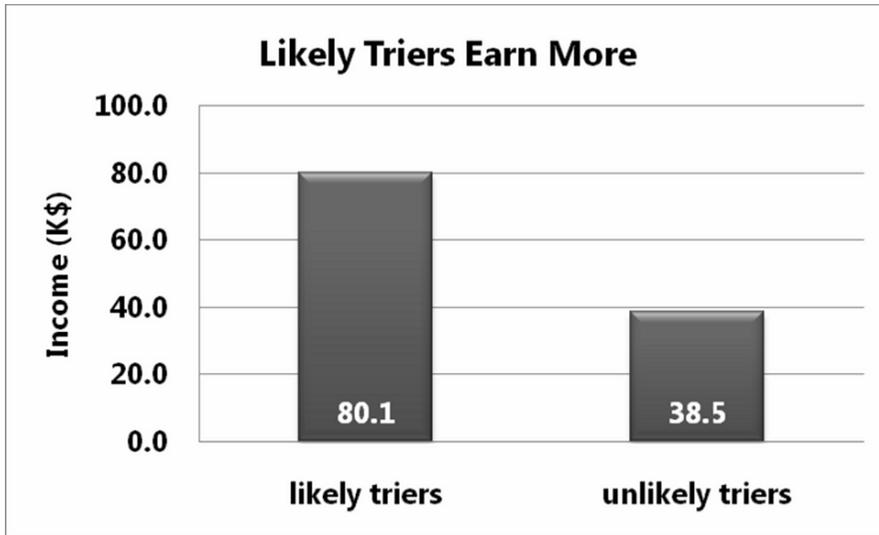
We want to show the average incomes of Likely and Unlikely Triers.  
 Add a row above **59**: Select **59**, **Alt HIR**, then in **B59** and **C59** enter labels *likely triers* and *unlikely triers*.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Likely Triers Income	Unlikely Triers Income	Income (K)	Trier	p value							
58													
59		likely triers	unlikely triers										
60	mean	80.1	38.5	24	likely	41.6							

Select the two labels and sample means in **B59:C60**, then use short cuts to insert a column chart: **Alt NC**.

Choose **Design, Chart Layout 6**, and add a title and vertical axis title.

Use shortcuts to add data labels: **Alt JAB**.



### Excel 3.8 Construct a pie chart of shares

*Moral Acceptance of Medical Testing on Animals.* We will construct a pie chart to illustrate how sample ratings of the acceptability of medical testing on animals are split.

Open a new workbook and type in two new columns, *rating* and *proportion*. In the *rating* column, type in *acceptable* and *unacceptable*.

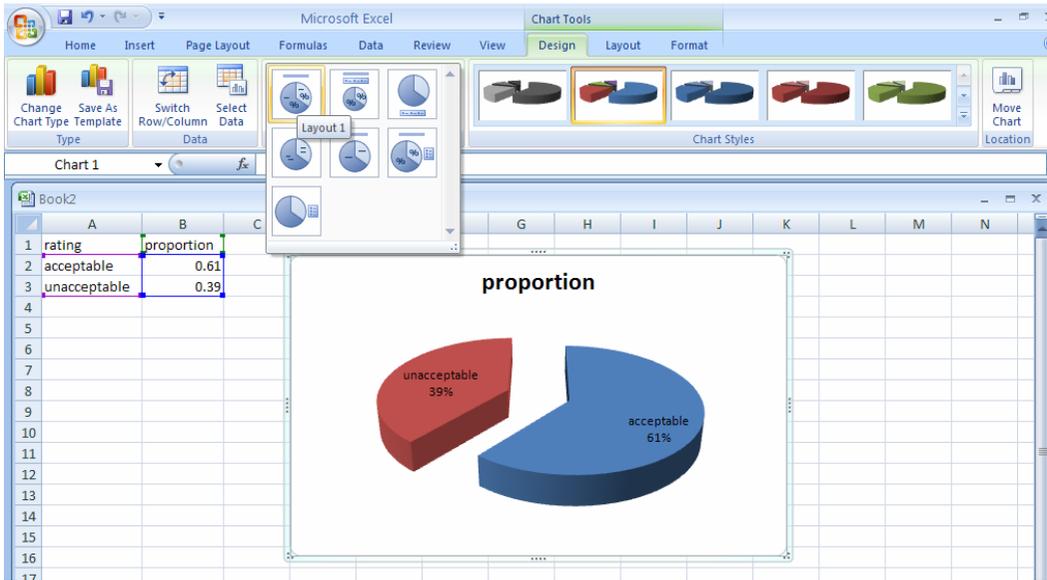
In the *proportion* column, type in the sample proportions that found medical testing on animals acceptable, *.61* and unacceptable *.39*.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	rating	proportion											
2	acceptable	0.61											
3	unacceptable	0.39											
4													

To make a pie chart, select **A1:B3**, then use shortcuts to insert a pie chart: **Alt NE**. (**E** selects a pie chart from the Insert menu.)

Click the three dimensional chart type.

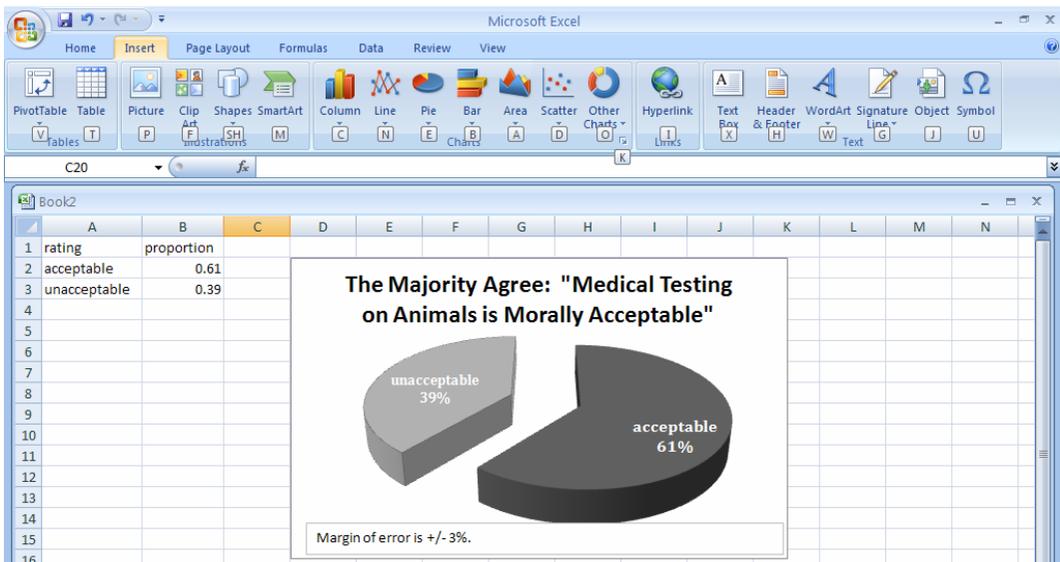
### Choose Design, Chart Layout 1:



Add a chart title.

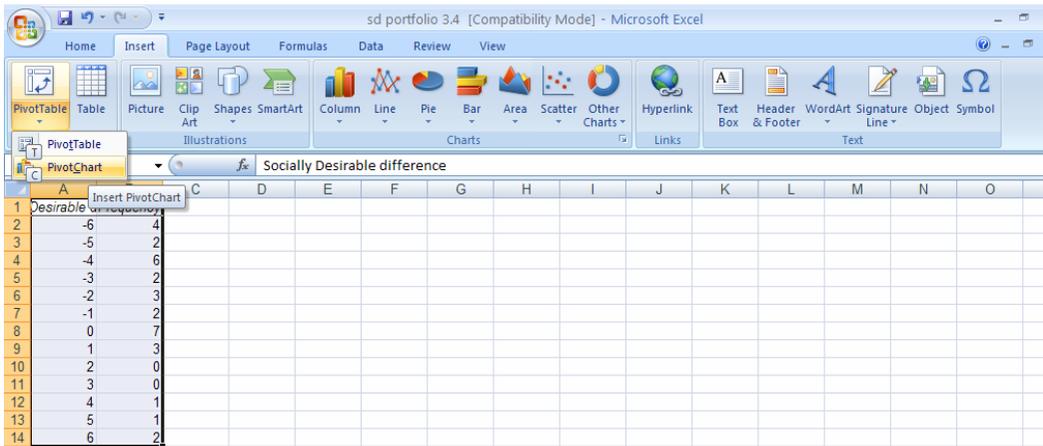
To add the margin of error, use shortcuts to insert a text box below the pie: **Alt NX**. (X selects Text Box from the Insert menu.)

Type in *Margin of error is +/- 3%*:



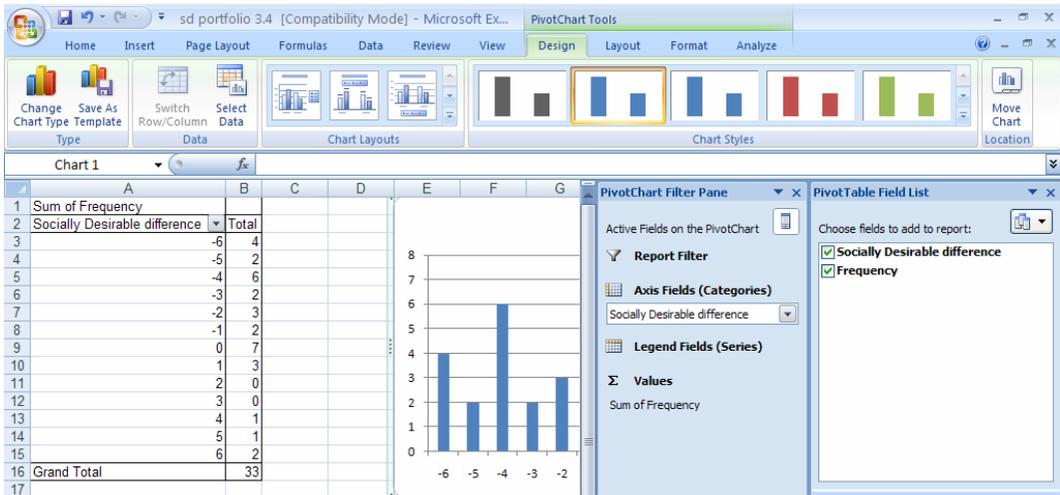


Use shortcuts to order a histogram tabulation of the *socially desirable differences* in **D**.

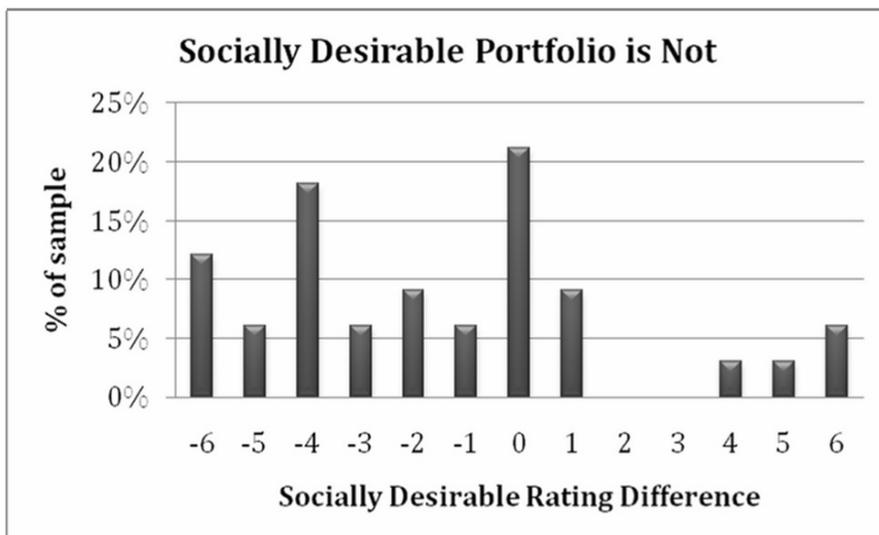


Use shortcuts to insert a PivotTable of the tabulation.

Drag *Socially Desirable difference* to **ROW** and drag *Frequency* to **DATA**:



Double click the **Sum of Frequency** box and **Show values as, % of total**. Reduce decimals in **B**, then click the PivotChart icon.



Though many potential investors rate Socially Desirable and conventional portfolios equally attractive, a sizeable percent rate Socially Desirable portfolios as less desirable than conventional portfolios.

### Excel 3.10 Construct a confidence interval for the difference between alternate scenarios or pairs

To estimate the population difference in investors' ratings of Socially Desirable and Conventional portfolios from our sample data, we will construct a confidence interval of the average rating difference.

In **C36:C40**, enter labels *sample mean difference*, *sample standard deviation*, *margin of error*, *95% lower*, *95% upper*.

To find the mean difference, in **C36**, enter **=AVERAGE(C2:C34)[Enter]**.

To find the standard deviation of the difference, in **C37**, enter **=STDEV(C2:C34)[Enter]**.

Use the standard deviation in **C37** to find the margin of error of the difference.

In **C38**, enter **=CONFIDENCE(.05, C37, 33)[Enter]**.



## Excel Shortcuts at Your Fingertips

### By Shortcut Key

**Alt** activates the shortcuts menus, linking keyboard letters to Excel menus. Press **Alt**, then release and press letters linked to the menu you want.

The following are examples of shortcuts. Press **Alt**, then

**H 9** to select the Home menu and the reduce decimals function

**H DC** to select the Home menu and the Delete function to delete column(s)

**H IC** to select the Home menu and Insert function and to insert a column to the left of the selected cell or column

**AY2** to select the Data and Data Analysis menus

**AS** to select the Data and the Sort menus

**NC** to select the Insert function and to insert a column chart

**ND** to select the Insert function and to insert a scatterplot

**NE** to select the Insert function and to insert a pie chart

**NVT** to select the Insert function, the Pivot menu, and to insert a PivotTable

**NX** to select the Insert function and to insert a text box

**WFR** to select the View and Freeze panes menus, and to Freeze rows

**JAB** to select the Layout and Data Labels menus

**JARM** to select the Layout, the Error Bar, and the custom Error Bar menus

**Shift+arrow** selects cells scrolled over

**Cntl+C** to copy

**Cntl+down arrow** scrolls through all cells in the same column that contain data and stops at the last filled cell.

**Cntl+R** fills in values of empty cells using a formula from the first cell in a selected array

**Cntl+Shift+down arrow** selects all filled cells in the column.

## By Goal

If you want to

Activate shortcuts menus, press **Alt**, then release.

Add data labels in a column chart: select a column, then **Alt JAB**

Add error bars in a column chart: select a column, then **Alt JARM**

Analyze data: **Alt AY2**

Copy cells: select the cells, then **Cntl+C**

Delete a column: **Alt HDC**

Freeze the top row: **Alt WFR**

Insert a column: **Alt HIC**

Insert a column chart: **Alt NC**

Insert a pie chart: **Alt NE**

Insert a PivotTable: **Alt NVT**

Insert a row: **Alt HIR**

Insert a scatterplot: **Alt ND**

Insert a text box: **Alt NX**

Move to the end of a column: **Cntl+down arrow**

Reduce decimals: **Alt H9**

Select all of the filled cells in a column: select the first cell in the column, then **Cntl+Shift+down arrow**

Sort data: **Alt AS**

## Lab Practice 3 Inference

### *cingular's Position in the Cell Phone Service Market*

cingular's managers have conducted a survey of customers in 21 major metropolitan areas to assess the quality of service along three key areas: *coverage*, *absence of dropped calls*, and *static*. Customers rated cingular service along each of these three dimensions using a five-point scale (1=poor to 5=excellent). Data are in **Lab Practice 3 cingular.xls**

Management's goal is to be able to offer service that is not perceived as inferior. This goal translates into mean ratings of at least 3 on the 5-point scale in the national market across all three service dimensions.

Based on this sample, average ratings in all major metropolitan areas are

\_\_\_\_\_ to \_\_\_\_\_ for *coverage*,

\_\_\_\_\_ to \_\_\_\_\_ for absence of *dropped calls*,

\_\_\_\_\_ to \_\_\_\_\_ for *static*, with 95% confidence.

Management can conclude that they have achieved their goal along:

\_\_\_\_ *coverage* \_\_\_\_ *dropped calls* \_\_\_\_ *static*

### *Value of a Nationals Uniform*

The Nationals General Manager is concerned that his club may not be paying competitive salaries. He has asked you to compare Nationals' salaries with salaries of players for the closest team in the National League East, the Phillies. He suspects that the Phillies may win more games because they are attracting better players with higher salary offers. Data are in **Lab Practice 3 Nationals.xls**.

This is a \_\_\_\_\_ tail *t test*.

*p value* from one tail *t test* of difference in team *salary* means: \_\_\_\_\_

The General Manager can conclude that, relative to the Phillies, the Nationals are paid \_\_\_\_  
Less \_\_\_\_ the same.

**Extra Value of a Phillies Uniform.** If you conclude that the Phillies do earn higher salaries, estimate the average difference at a 95% level of confidence.

On average, players for the Phillys earn \_\_\_\_\_ to \_\_\_\_\_ more than players for the Nationals.

The pooled standard error of the difference in mean salaries is: \_\_\_\_\_

*Illustrate the two teams' salaries with a column chart.*

### *Confidence in Chinese Imports*

Following the recall of a number of products imported from China, the Associated Press-Ipsos Poll asked 1,005 randomly selected adults about the perceived safety of products imported from China. Poll results are below:

**“When it comes to the products that you buy that are made in China, how confident are you that those products are safe . . . ?”**

	Not	
Confident	Confident	Unsure
%	%	%
42	57	1

Use this data to construct an *approximate, conservative 95% confidence interval* for the *proportion Not Confident* that Chinese imports are safe.

\_\_\_\_\_ to \_\_\_\_\_ percent are not confident that products made in China are safe.

Illustrate your result with a pie chart which includes the margin of error in a text box. Add a “bottom line” title.

## Lab 3 Inference

### I. Dell PDA Plans

Managers at Dell are considering a joint venture with a Chinese firm to launch a new PDA equipped with Qwerty keyboard and loaded with Microsoft Office.

In a concept test using a random sample of 1,000 PDA owners, **20%** indicated that they would probably or definitely replace their PDA with the new product within the next quarter.

Norms from past market research indicate that **80%** who indicate intent to replace actually will.

\_\_\_\_\_ to \_\_\_\_\_% of PDA owners are expected to replace with the new Dell PDA in the next quarter.

Construct a pie chart showing the percents of all PDA owners (i) who are expected to replace their PDAs with the new Dell PDA and (ii) who aren't expected to replace their PDAs with the new Dell PDA. Include a descriptive title and add a text box showing the conservative, approximate margin of error.

The percent of PDA owners who are expected to replace is Dell's best estimate of market share.

*Dell market share* in the third quarter of 2008 is most likely to be **16%** if

$$\begin{aligned} \text{Dell market share}_q &= \text{intent \% to replace with Dell} \times \text{expected replacement \% per intent \%} \\ &= 20\% \times 80\% \\ &= 16\% \end{aligned}$$

*The proportion who will replace their PDAs with the Dell PDA is approximately **Normal**.*

The world PDA market declined in the first two quarters of 2008, down 40% from shipments in 2007.

- *World shipments* in the third quarter of 2008 are most likely to be **600,000**.
- *World shipments* will fall between **500,000** and **800,000 with 90% certainty**.
- Lower potential *world shipments* are more likely, similar to a **triangular** distribution.

Managers want to know the likelihood that shipments of the new PDA will exceed **80,000 in the third quarter of 2008**.

Build a spreadsheet linking *Dell shipments* to *world shipments* and *Dell market share*:

$$\text{Dell shipments}_t = \text{Dell market share}_t \times \text{world shipments}_t$$

Then, use Crystal Ball to create 1,000 samples, specifying managers' assumptions.

Given these assumptions, what is the chance that *shipments* will exceed **80,000 in the third quarter of 2008**? \_\_\_%

### *Assignment 3-1 Bottled Water Possibilities*

The students in Team 8, Stephanie, Shawn, Erica, and Tyler, want to know how their assumptions regarding

- *demand* for bottled water and
- *market share*

affect the chances that *bottles sold* will **exceed 500,000**.

Stephanie has convinced her teammates to consider a broader range of possibilities for the 5% and 95% demand assumptions in their monte carlo simulation. Two other teams in the class reported that average *demand* for bottled water could be **as low as 7.6**, while a third team reported that average *demand* could be **as high as 14 bottles** per customer per week.

Use Crystal Ball to conduct a monte carlo simulation of bottles sold with these two assumptions:

- *average demand for bottled water* will be **less than 7.6** in 5% of samples and **less than 14.0** in 95% of samples
- *market share* that Team 8 could achieve with their custom bottled water dispensers could be **as low as 2%** and **as high as 10%**, and the *market share* possibilities within this range are equally likely, or **uniformly distributed**.

What are the chances that Team 8 could sell **at least 500,000 bottles** in the first year, given these assumptions?

Include the distribution of bottles sold to illustrate your answer.

### Assignment 3-2 Immigration in the U.S.

The FOX News/Opinion Dynamics Poll, July 11-12, 2006, of (N=) 900 registered voters nationwide, reports public opinion concerning immigrants and proposed immigration legislation:

“In general, do you think immigrants who come to the United States today join society and give to the country or stay separate from society and take from the country?”				
	Join Society/ Give	Stay Separate/ Take	Depends (vol.)	Unsure
	%	%	%	%
7/11-12/06	41	36	17	6
“Do you think the United States should increase or decrease the number of legal immigrants allowed to move to this country?”				
	Increase	Decrease	No Change (vol.)	Unsure
	%	%	%	%
7/11-12/06	24	51	17	8

Use this data to construct *approximate, conservative 95% confidence intervals* for the *proportions* who (i) agree that immigrants contribute positively to society and (ii) agree that the U.S. should increase the number of legal immigrants.

Briefly summarize the opinions of **all registered voters** using language that American adults would understand.

Illustrate your summary with pie charts embedded in your report.

Be sure to include the margin of error in your pie chart.

### Assignment 3-3 McLattes

McDonalds recently sponsored a blind taste test of lattes from Starbucks and their own McCafes. A sample of thirty Starbucks customers tasted both lattes from unmarked cups and provided ratings on a -3 (=worst latte I’ve ever tasted) to +3 (=best latte I’ve ever tasted) scale. These data are in **Assignment 3-3 Latte.xls**.

Can McDonalds claim that their lattes taste every bit as good as Starbucks’ lattes? (Please use 95% confidence.)

What evidence allows you to reach this conclusion?

### *Assignment 3-4 A Barbie Duff in Stuff*

Mattel recently sponsored a test of their new Barbie designed by Hillary Duff. The Duff Barbie is dressed in Stuff, Hillary Duff clothing designs, and resembles Hillary Duff. Mattel wanted to know whether or not the Duff Barbie could compete with rival MGA Entertainment's Bratz dolls.

A sample of thirty 7-year-old girls attended Barbie parties, played with both dolls, then rated both on a -3 (=Not At All Like Me) to +3 (=Just Like Me) scale. These data are in **Assignment 3-4 Barbie.xls**.

Do the 7-year-olds identify more strongly with the Duff Barbie in Stuff than the Bratz? (Please use 95% confidence.)

What evidence allows you to reach this conclusion?

### *CASE 3-1 Yankees v Marlins: The Value of a Yankee Uniform<sup>1</sup>*

The Marlins General Manager is disgruntled because two desirable rookies accepted offers from the Yankees instead of the Marlins. He believes that Yankee salaries must be noticeably higher—otherwise, the best players would join the Marlins organization. Is there a difference in salaries between the two teams? If the typical Yankee is better compensated, the General Manager is planning to chat with the Owners about sweetening the Marlins' offers. He suspects that the Owners will argue that the typical Yankee is older and more experienced, justifying some difference in salaries.

Data are in **Case 3-1 Yankees v Marlins Salaries.xls**.

Determine:

- whether or not Yankees earn more on average than Marlins, and
- whether or not players for the Yankees are older on average than players for the Marlins.

If you find a difference in either case, construct a *95% confidence interval* of the expected difference in any season.

Briefly summarize your results using language that the General Manager and Owners would understand, and illustrate with a column chart.

---

<sup>1</sup> This example is a hypothetical scenario using actual data.

### *CASE 3-2 Gender Pay*

The Human Resources manager of Slam's Club is shocked by the recent revelations of gender discrimination by WalMart ("How Corporate America is Betraying Women," Fortune, January 10, 2005), and wants to confirm the null hypothesis that there is no gender difference in average salaries in his firm. He also wants to know whether levels of responsibility (measured with the Position variable) and experience differ between men and women, since this could explain a difference in salaries.

**Case 3-2 GenderPay.xls** contains *salaries*, *positions*, and *experience* of men and women from a random sample of the company records.

Determine

- whether or not the sample supports a conclusion that men and women are paid equally,
- whether average level of *responsibility* differs across genders,
- whether average *experience* differs across genders.

If you find that the data support the alternate hypothesis that men are paid more, on average, construct a 95% confidence interval of the expected average difference.

If either average level of *responsibility* or average years of *experience* differs, construct 95% *confidence intervals* of the expected average difference.

Briefly summarize your results using language that a businessperson (who may not remember quantitative analysis) could understand.

Illustrate your results with column charts.

### *CASE 3-3 Polaski Vodka: Can a Polish Vodka Stand Up to the Russians?*

Seagrams management decided to enter the premium vodka market with a Polish vodka, suspecting that it would be difficult to compete with Stolichnaya, a Russian vodka and the leading premium brand. The product formulation and the package/brand impact on perceived taste were explored with experiments to decide whether the new brand was ready to launch.

**The taste.** First, Seagrams managers asked, “Could consumers distinguish between Stolichnaya and Seagrams’ Polish vodka in a *blind* taste test, where the impact of packaging and brand name were absent?”

Consultants designed an experiment to test the null and alternative hypotheses:

$H_0$ : The taste rating of Seagram’s Polish vodka is at least as high as the taste rating of Stolichnaya. The average difference between taste ratings of Stolichnaya and Seagrams’ Polish vodka does not exceed zero:

$$\mu_{STOLICHNAYA} - \mu_{POLISH} \leq 0$$

$H_1$ : The taste rating of Seagram’s Polish vodka is lower than the taste rating of Stolichnaya. The average difference between taste ratings of Stolichnaya and Seagram’s Polish vodka is positive:

$$\mu_{STOLICHNAYA} - \mu_{POLISH} > 0$$

In this first experiment, each participant tasted two unidentified vodka samples and rated the taste of each on a ten-point scale. Between tastes, participants cleansed palates with water. Experimenters flipped a coin to determine which product would be served first: if heads, Seagrams’ polish vodka was poured first; if tails, Stolichnaya was poured first. Both samples were poured from plain, clear beakers. The only difference between the two samples was the actual vodka.

These experimental data in **Case 3-3 Pulaski Taste.xls** are repeated measures. From each participant, we have two measures whose difference is the difference in taste between the Russian and Polish vodkas.

Test the difference between taste ratings of the two vodkas.

Construct a *95% confidence interval* of the difference in taste ratings.

Illustrate your results with a PivotChart and interpret your results for management.

**The brand & package.** Seagrams management proceeded to test the packaging and name, Polaski. The null hypothesis was:

$H_0$ : The taste rating of Polaski vodka poured from a Polaski bottle is at least as high as the taste rating of Polaski vodka poured from a Stolichnaya bottle. The mean difference between taste ratings of Polaski vodka poured from a Stolichnaya bottle and Polaski vodka poured from the Seagrams bottle bearing the Polaski brand name is not exceed zero.

Alternatively, if the leading brand name and distinctive bottle of the Russian vodka affected taste perceptions, the following could be true:

$H_1$ : The mean difference between taste ratings of Polaski vodka poured from Stolichnaya bottle and Polaski vodka poured from the Seagrams bottle bearing the Polaski brand name is positive.

In this second experiment, Polaski samples were presented to participants twice, once poured from a Stolichnaya bottle, and once poured from the Seagrams bottle, bearing the Polaski name. Any minute differences in the actual products were controlled for by using Polaski vodka in both samples. Differences in taste ratings would be attributable to the difference in packaging and brand name.

Thirty new participants again tasted two vodka samples, cleansing their palates with water between tastes. As before, a coin toss decided which bottle the first sample would be poured from: Stolichnaya if heads, Polaski if tails. Each participant rated the taste of the two samples on a ten-point scale.

These data are in **Case 3-3 Polaski Package.xls**.

Test the difference in ratings due to packaging.

Construct a *95% confidence interval* of the difference in ratings due to the packaging.

Illustrate your results with a PivotChart.

Interpret your results for management

### *CASE 3-4 American Girl in Starbucks*

Mattel and Warner Brothers are considering a partnership with Starbucks to promote their new American Girl movie. Starbucks previously backed Lionsgate's "Akeelah and the Bee," which earned \$19 million. In exchange for \$5 million, Starbucks would install signage and stickers in 6,800 of its stores, print American Girl-branded cup sleeves, sell plush American Girl pets and the picture's soundtrack. Materials for the movie would also appear on the company's website. Starbucks claims 44 million weekly customers in the 6,800 stores.

- a. In a pretest of the promotion during one week in one Starbucks store, **184 of the 924** Fast Card customers served that week agreed that they had heard of the movie when surveyed by phone the following week.

With 90% confidence, what *proportion* of Starbucks' customers can Mattel managers expect to become aware of the film from promotional materials in stores?

- b. Mattel managers believe that roughly **25%** of those who are aware of the movie will buy tickets, though this percent could be **as low as 10%** or possibly **as high as 60%**. Each movie-goer is expected to bring **2** family members or friends, on average, though the average number of guests could be **as low as 1.5**, or possibly **as high as 3.0**.
- c. Mattel would earn **\$1** royalty from each ticket sold.
- d. To justify the promotion, Mattel management wants to be sure that royalties from ticket sales are likely to **exceed \$5 million**.

What are the chances that *royalties* from ticket sales would **exceed \$5 million**?

- e. Mattel and Warner Brothers are also considering McDonalds as a potential promoter of the new movie. Mattel management suspects that Starbucks customers are wealthier than McDonalds customers. (Since wealthier families have the resources to buy American Girl products, this is the target market for the new movie audience, and Mattel would favor the sponsor with wealthier customers.)

Household income data from intercept interviews of thirty McDonalds customers and thirty Starbucks customers are in **Case 3-4 Starbucks vs McD.xls**.

Can Mattel managers conclude that Starbucks customers are wealthier than McDonalds customers? (Please use a 95% level of confidence.)

What evidence allows you to reach this conclusion?

Estimate the *income* difference between Starbucks and McDonalds customers using a *95% confidence interval*.

## 4

# Quantifying the Influence of Performance Drivers and Forecasting: Regression

Regression analysis is a powerful tool for quantifying the influence of a continuous, *independent, driver*  $X$  on a continuous *dependent, performance* variable  $Y$ . Often we are interested in both explaining how an independent decision variable  $X$  drives a dependent performance variable  $Y$  and also in predicting performance  $Y$  to compare the impact of alternate decision variable  $X$  values.  $X$  is also called a *predictor* since from  $X$  we can predict  $Y$ . Regression allows us to do both: quantify the nature and extent of influence of a performance driver and predict performance or response  $Y$  from knowledge of the driver  $X$ .

With regression analysis, we can statistically address these questions:

- Is variation in a dependent, performance, response variable  $Y$  influenced by variation in an independent variable  $X$ ?

If yes,  $X$  is a driver of  $Y$ , and, with regression, we can answer these questions:

- What percent of variation in performance  $Y$  can be accounted for with variation in driver  $X$ ?
- If driver  $X$  changes by one unit, what range of response can we expect in performance  $Y$ ?
- At a specified level of the driver  $X$ , what range of performance levels  $Y$  are expected?

In this chapter, simple linear regression is introduced, and we explore ways to address each of these questions linking a continuous driver, which may be a decision variable, to a continuous performance variable. We also explore the link between correlation and simple linear regression, since the two are closely related.

### 4.1 The Simple Linear Regression Equation Describes the Line Relating A Decision Variable to Performance

Regression gives us an equation for the line which best relates changes or differences in a continuous, dependent performance variable  $Y$  to changes or differences in a continuous, independent driver  $X$ . This line comes closest to each of the points in a scatterplot of  $Y$  and  $X$ :

$$\hat{Y} = b_0 + b_1 X$$

Where  $\hat{Y}$  is the expected value of the dependent performance, or response, variable, called “y-hat”,

$X$  is the value of an independent variable, decision variable, or driver,  
 $b_0$  is the intercept estimate, which is the expected value of  $\hat{Y}$  when  $X$  is zero,  
 $b_1$  is the estimated slope of the regression line, which indicates the expected change in performance  $\hat{Y}$  in response to a unit change from the driver's average  $\bar{X}$ .

**Example 4.1 HitFlix Movie Rentals.** An owner of a chain of movie rental kiosks is planning to add a new kiosk and needs to decide how large it will be. He is planning to add a kiosk of 100 square feet, but he thinks a larger store might generate more revenue, since footage may drive revenues. The null and alternate hypotheses which he would like to test are:

$H_0$ : Store footage  $X$  has no effect on movie rental revenues  $Y$ .

$H_1$ : Store footage  $X$  drives movie rental kiosk revenues  $Y$ .

Scatterplots of footage,  $X$ , and annual kiosk revenues,  $Y$ , for a random sample of fifty-two kiosks from the chain are shown in Figure 4.1 from Excel:

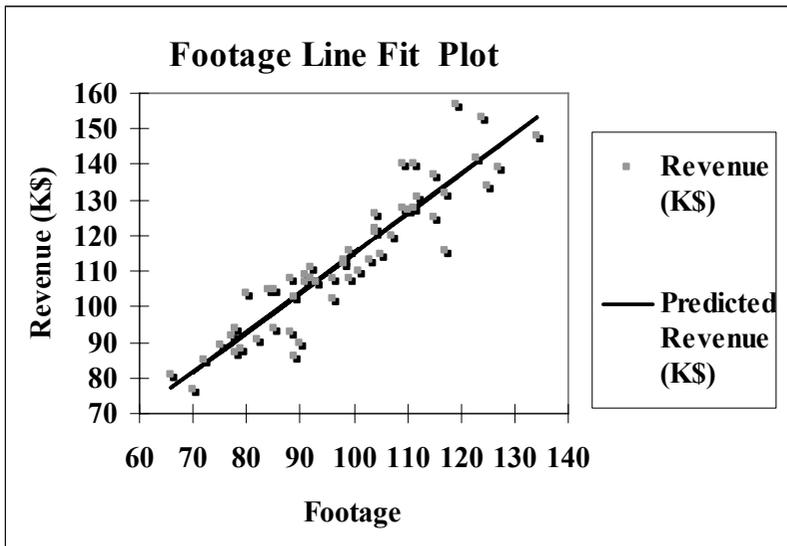


Figure 4.1 Store revenues by store footage

The scatterplot indicates that kiosk revenues may be a linear function of footage. For each additional foot of space, average annual revenues increase by about \$1.12K or \$1,120. The average difference in revenues between kiosks with 70 and 80 square feet, \$11,200 [= (80-70) x \$1,120] is identical to the average difference in revenues between kiosks with 120 and 130 square feet, \$11,200 [= (130-120) x \$1,120]. Expected revenues  $\hat{Y}$  increase at a constant rate of \$1,120 with each increase of one square foot. Because variation in revenues  $Y$  is related linearly to variation in footage  $X$ , the linear regression line is a good summary of the data:

$$\text{revenues}(K\$) = 3.43 + 1.12\text{Footage}$$

In this example, the intercept estimate  $b_0$  is 3.43. Were a kiosk to have zero square feet of space (which isn't possible), expected revenue would be \$3,430. The estimated slope  $b_1$  is 1.12, indicating that we expect an average change in revenue of \$1,120 in response to a change in kiosk space of one square foot.

### 4.2 *F* Tests the Significance of the Hypothesized Linear Relationship, *Rsquare* Summarizes Its Strength and Standard Error Reflects Forecasting Precision

Using the regression formula, we can predict the expected revenue  $\hat{Y}$  for any given size kiosk with square footage  $X$ . Table 4.1 contains predictions for five kiosks of different sizes:

<i>Footage</i> $X$	<i>Expected Revenue</i>		
	$b_0$	$+b_1X$	$= \hat{Y}$
70	3.43	1.12 (70)	82.6
80	3.43	1.12 (80)	93.7
90	3.43	1.12 (90)	104.9
110	3.43	1.12 (110)	126.2

Expected revenues are close to actual revenue for kiosks of these sizes, but not identical, since other factors also influence revenues.

Table 4.1 Expected revenue

The differences between expected and actual revenue are the *residuals* or errors. Errors from these four stores are shown in Table 4.2 and Figure 4.1.

<i>Square feet</i> $X$	<i>Expected Revenue</i> $\hat{Y}$	<i>Actual Revenue</i> $Y$	<i>Error</i> $e = Y - \hat{Y}$
70	82.5	76.6	-4.6
80	93.7	104.4	11.3
90	104.9	90.4	-13.9
<b>110</b>	126.2	126.7	.8

Table 4.2 Errors from the regression line

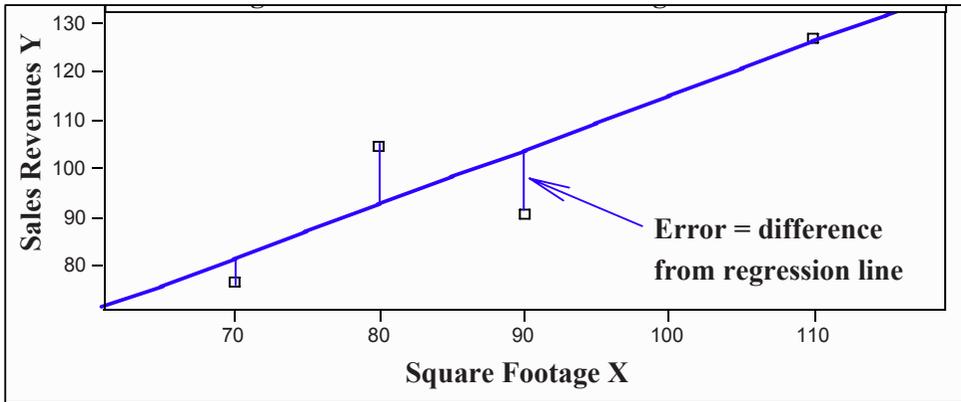


Figure 4.2 Four errors from the regression line

The *Sum of Squared Errors* in a sample,

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y})^2 = \sum (Y_i - b_0 - b_1 X_i)^2$$

is the portion of total variation in the dependent variable, SST, which remains unexplained after accounting for the impact of variation in  $X$ . The *Least Squares* regression line is the line with the smallest SSE of all possible lines relating  $X$  to  $Y$ .

The regression *standard error*, equal to the square root of SSE,

$$\text{standard error} = \sqrt{SSE}$$

reflects the precision of the regression equation. We expect forecasts to be within two standard errors of actual performance 95% of the time.

The difference,  $SST - SSE$ , called the *Regression Sum of Squares*,  $SSR$ , or *Model Sum of Squares*, is the portion of total variation in  $Y$  influenced by variation in  $X$ . To test the hypothesis that the independent variable influences the dependent variable in the population, we use our sample data to calculate the ratio of the mean variation explained by the regression  $MSR$  to mean unexplained variation  $MSE$ . This ratio is distributed as an  $F$  with 1 numerator (for the predictor) and  $(N-2)$  denominator degrees of freedom:

$$F_{1,(N-2)} = \frac{SSR/1}{SSE/(N-2)} = \frac{MSR}{MSE}$$

(We lose one degree of freedom from estimation of the dependent variable mean and one from estimation of the independent variable mean.) The percent of total variation in the

dependent, performance variable  $Y$  which can be accounted for by variation in the independent decision variable  $X$  is *RSquare*:

$$RSquare = SSR / SST$$

*RSquare* ranges between zero and one, or zero and one hundred percent. The greater the influence of  $X$  on  $Y$ , the closer *RSquare* is to one hundred percent, and the larger  $F$  is.

*RSquare* and the standard error appear in SUMMARY OUTPUT, which is followed by the ANOVA table in regression output. The SUMMARY OUTPUT and ANOVA tables from Excel for the **HitFlix Movie Rental** regression are shown in Table 4.3.

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.93				
R Square	0.86				
Adjusted R Square	0.86				
Standard Error	7.44				
Observations	52				
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	16,800	16763	303	.0000
Residual	50	2,800	55		
Total	51	19,500			

Table 4.3 Model summary of fit and ANOVA table

*RSquare*, the ratio of Regression Sum of Squares (16,800) to Total Sum of Squares (19,500), is .86, or 86%:

$$RSquare = \frac{\text{regressionSumofSquares}}{\text{TotalSumofSquares}} = \frac{16,800}{19,500} = .86$$

Variation in footage  $X$  accounts for 86% of the variation in revenues  $Y$ . Other factors account for the remaining 14%.

The regression standard error is 7.44(\$K): We can expect 95% of revenue forecasts for a kiosks of a specified size to be no further than twice this standard error, 14.8 (\$K), or \$14,800, from average revenues of all kiosks of that size.

The  $F_{1,50}$  statistic is 303. With a sample of size 52 and one independent variable, the significance of  $F$  is a very small number, less than .0001. There is less than a tenth of a percent chance that we would observe the sample data patterns, were footage not driving revenues.

Based on regression analysis of this sample, we have sufficient evidence to reject the null hypothesis:

$H_0$ : Store footage  $X$  has no effect on movie rental revenues  $Y$ .

And we then accept the alternate hypothesis:

$H_1$ : Store footage  $X$  drives movie rental revenues  $Y$ .

### 4.3 The Population Slope Is Tested And Inferred From Our Sample

Because the true impact  $\beta_1$  of a driver  $X$  on performance  $Y$  is unknown, this slope, or *coefficient*, is estimated from a sample. This estimate  $b_1$  and its sample standard error  $s_{b_1}$  are then used to test the hypothesis that  $X$  influences  $Y$ :

$H_0$ : The independent variable  $X$  has no influence on the dependent variable  $Y$ .

OR

$H_0$ : The regression slope is zero:  $\beta_1=0$ .

Alternatively,

$H_1$ : The *independent* variable  $X$  drives the dependent variable  $Y$ .

OR

$H_1$ : The regression slope is not zero:  $\beta_1 \neq 0$ .

In many instances, from experience or logic, we know the likely direction of influence. In those instances, the alternate hypothesis requires a one-tail test:

$H_1$ : The independent variable  $X$  positively influences the dependent variable  $Y$ .

OR

$H_1$ : The regression slope is greater than zero:  $\beta_1 > 0$ .

This one-sided alternate hypothesis describes an upward slope. A similar alternate hypothesis could be used when logic or experience suggests a downward slope.

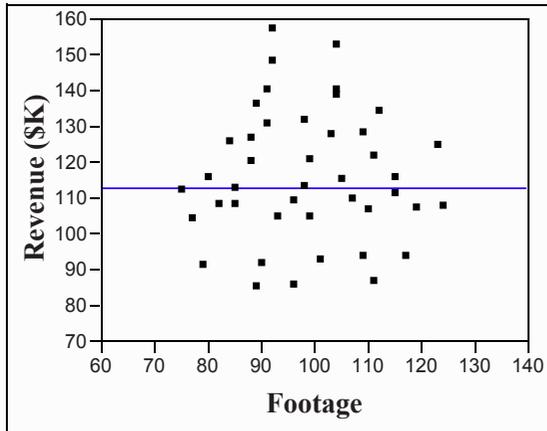


Figure 4.3  $X$  does not drive  $Y$  and the regression line slope is flat ( $b_1=0$ )

In our **Movie Rentals** example, if revenue did not depend on footage, the scatterplot would resemble a spherical cloud and the regression line would be flat at the dependent variable mean  $\bar{Y}$ , as in Figure 4.3. To form a conclusion about the significance of the slope, we calculate the number of standard errors which separate our estimate  $b_1$  from zero:

$$t_1 = b_1 / s_{b_1}$$

In **Movie Rentals**, the standard error of the slope estimate  $s_{b_1}$  is .064. The slope is more than seventeen standard errors from zero:

$$t_1 = 1.12 / .064 = 17.4,$$

At this  $t$  value, a two tail test has a  $p$  value of .0001. From both experience and logic, the kiosk chain owner had a good idea that footage has a positive impact on revenues, so his alternate hypothesis is that the slope is positive. Dividing the two tail  $p$  value by 2, the one tail  $p$  value is .00005. There is less than a twentieth of a percent chance that we would observe the sample data were footage not driving revenues. From our sample evidence, we reject the null hypothesis of a flat slope and accept the alternate hypothesis of a positive slope. Footage has a positive impact on revenues.

Excel does these calculations for us. The slope and intercept estimates are labeled *Coefficients* in Excel, shown in Table 4.4.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
<i>Intercept</i>	3.43	6.38	0.5	0.5931	-9.39	16.25
<i>Footage</i>	1.12	0.064	17.4	0.0000	0.99	1.24

Table 4.4 Coefficient estimates, standard errors and *t tests*

There is a 95% chance that the true population slope will fall within approximately two standard errors of our estimate:

$$b_1 - 2s_{b_1} < \beta_1 < b_1 + 2s_{b_1}$$

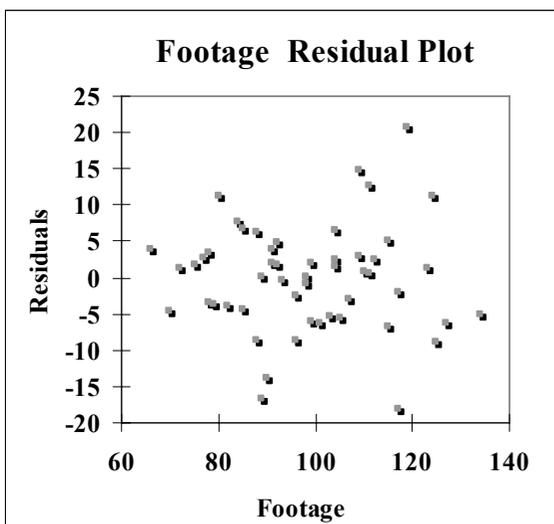
$$1.12 - 2(.064) < \beta_1 < 1.12 + 2(.064)$$

$$.99 < \beta_1 < 1.24$$

The impact of one additional square foot on kiosk revenue is within the range of .99 to 1.24 (\$K) or \$990 to \$1,240.

#### 4.4 Analyze Residuals To Learn Whether Assumptions Have Been Met

We assume when we use linear regression that the errors are uncorrelated with the independent variable. For example, we should be as good at our explanation and prediction of revenues for small kiosks, as we are for large kiosks. To confirm that this assumption is met, we look at a plot of the residuals by predicted values. We should see no pattern.



A plot of the residuals by predicted values, Figure 4.4, is not pattern-free. The residuals show more variation for larger kiosks. Within the range of existing sizes of kiosks, we can expect predictions for small kiosks to be more accurate than predictions for large kiosks. This situation, in which residual variation is nonconstant, is termed *heteroskedasticity*. A remedy may be rescaling either the dependent variable, the independent variable, or both, perhaps to natural logarithms.

Figure 4.4 Residuals by predicted values

Linear regression assumes that the residuals are *Normally* distributed.

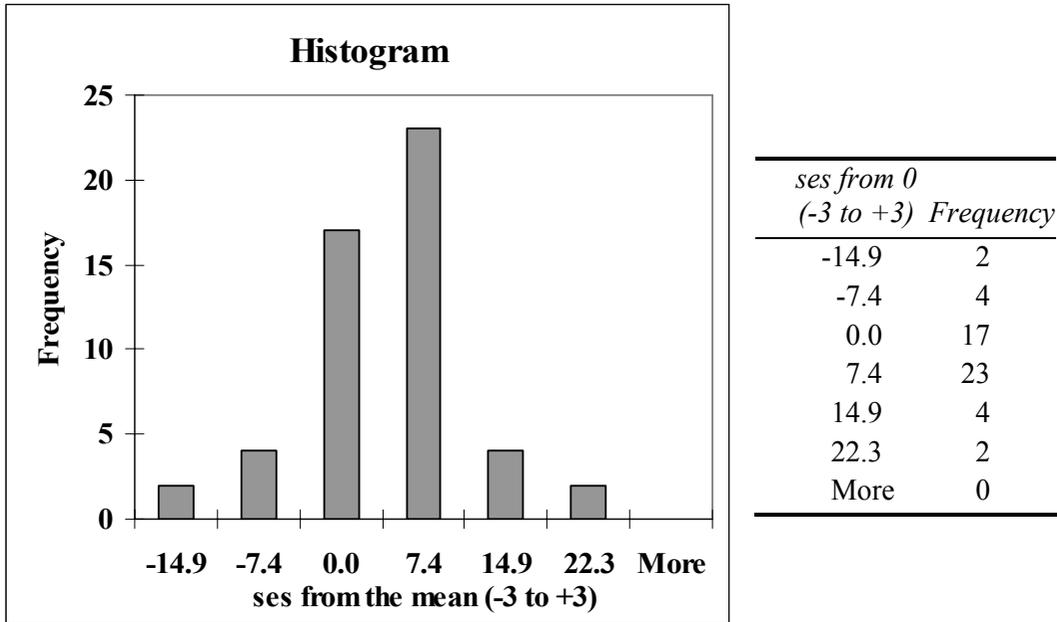


Figure 4.5 Slightly peaked residuals

The distribution of residuals, shown in Figure 4.5, is bell-shaped though slightly “peaked.” The distribution of residuals is more peaked than *Normal*. Too many residuals, 77%  $(=(17+23)/52)$  are within one standard deviation of the mean, which is more than the 67% expected from *Normally* distributed residuals. 92%  $(=(4+17+23+4)/52)$  of forecasts are within two standard errors, \$14.8 (000), of actual, and about eight percent are more than two standard errors, \$14,900, from actual, which is more than the 5% we expect from *Normally* distributed residuals.

#### 4.5 95% Prediction Intervals Acknowledge That Individual Elements Differ

Regression analysis can be used to forecast a 95% confidence interval for the value of the dependent variable  $Y$  given a specific value for the independent variable  $X$ . The standard error for this prediction  $s_y$ , depends on how much  $X$  influences  $Y$ , the sample size  $N$ , the standard deviation of  $X$ , and how far the particular, specific value of  $X$  is from the average  $\bar{X}$ . However, if the sample size is large, the standard prediction errors will be close to the regression *Standard Error* or *Root Mean Square Error*,  $s$ . As its name suggests, Root Mean Square Error  $s$  is the square root of  $SSE$ .

In **HitFlix Movie Rentals**,  $s$  is 7.44. This means that we expect forecasts for individual kiosks to be within approximately \$14,900 [=2 \* 7.44 (\$K)] of actual revenues. The prediction margin of error is approximately \$14,900. Approximate 95% prediction intervals for kiosks of several sizes are shown in Table 4.5 and Figure 4.6.

<i>footage</i>	<i>expected revenue (\$K)</i> $\hat{Y}$	<i>standard error</i> $s$	<i>approximate 95% prediction interval</i> $\hat{Y} \pm 2s$	
70	82	7.4	67	96
100	115	7.4	100	130
130	149	7.4	134	163

Table 4.5 Individual 95% prediction intervals

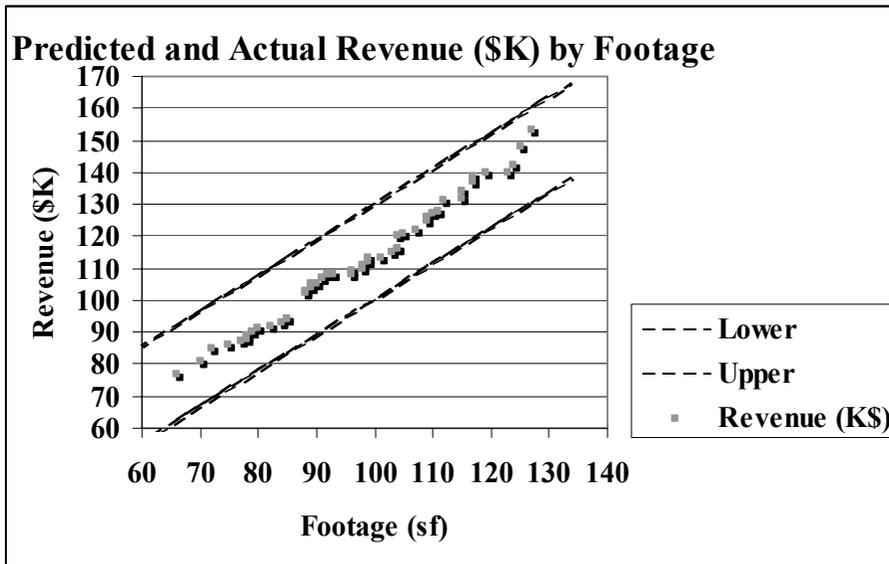


Figure 4.6 95% prediction intervals for individual kiosks

### 4.6 Use Sensitivity Analysis to Explore Alternative Scenarios

Comparing possible revenues from the planned kiosk of 100 square feet with a larger 130 square foot option, the HitFlix owner learns that the additional thirty square feet is expected to produce \$34,000 (= \$149,000 - \$115,000) additional revenue, though it could produce as little as \$4,000 (= \$134,000 - \$130,000) additional revenue, or as much as \$63,000 (= \$163,000 - \$100,000) more revenue. A kiosk with an additional thirty square feet (130, instead of 100) will generate \$4,000 to \$63,000 more revenue.

<i>Footage</i>	<i>predicted revenue (\$K)</i>	<i>lower 95% prediction</i>	<i>upper 95% prediction</i>
100	115	100	130
130	149	134	163

### 4.7 95% Conditional Mean Prediction Intervals Of Average Performance Gauge Average Performance Response To A Driver

If we are interested in estimating average population performance given a particular decision variable value  $X$ , our conditional mean prediction intervals will be narrower. In this case, we are incorporating only the model uncertainty and not the variation across individual stores of particular size. If, for example, the kiosk chain owner expected to add thirty new kiosks of the same size and wanted to know what average revenue to expect, he would ask for the 95% conditional mean prediction interval, given the planned kiosk size.

The formula for prediction error involves matrix algebra. However, we can calculate *approximate* standard prediction errors for conditional mean forecasts with this formula:

$$s_Y = s / \sqrt{N} ,$$

where  $s$  is the regression *standard error* and  $N$  is the sample size.

In **HitFlix Movie Rental Revenues**, the approximate standard error for mean predictions is \$1.03 (000) or \$1,030:

$$s_{Y_{approximate}} = 7.44 / \sqrt{52} = 1.03$$

We expect our forecasts to be within approximately \$2,060 [= 2 \* 1.03 (\$K)] of actual average revenues across kiosks with the same footage. The approximate forecast margin of error is \$2,060. 95% mean prediction intervals for average revenues at varying sizes from Excel are shown in Figure 4.7.

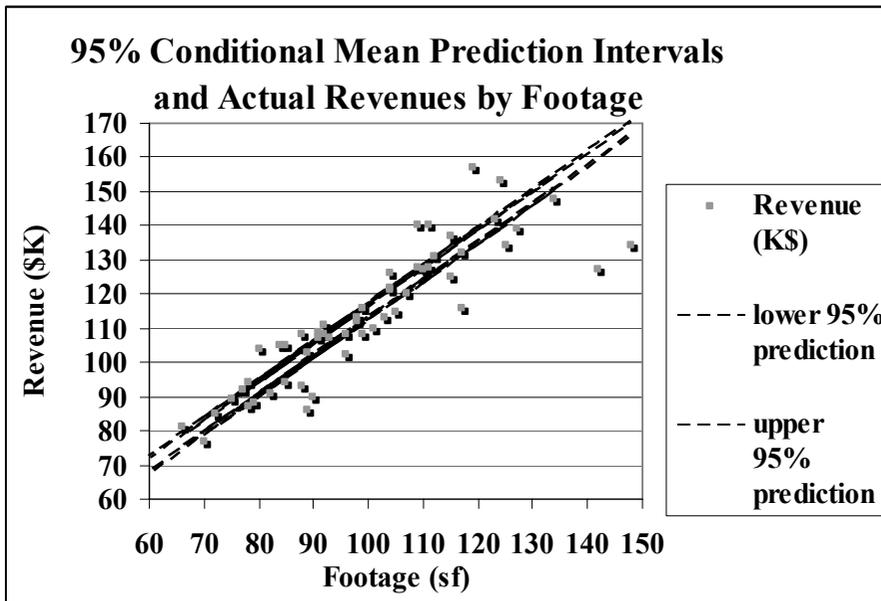


Figure 4.7 95% conditional mean prediction intervals for varying footage levels

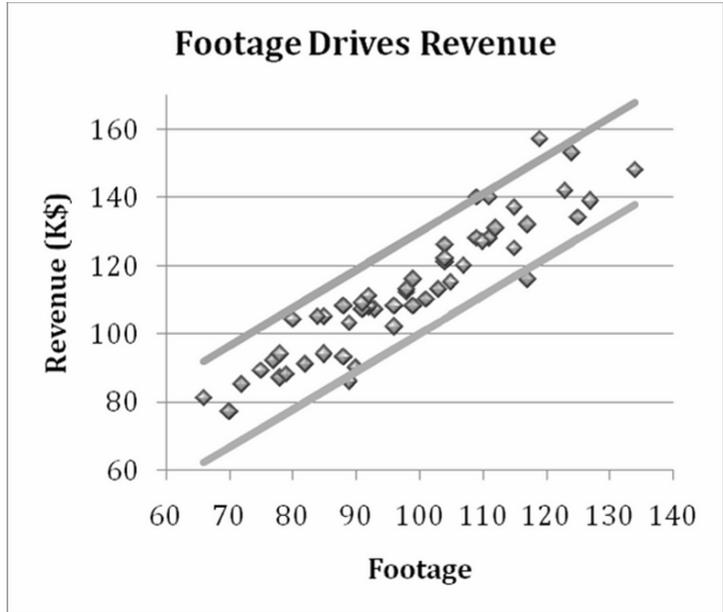
#### 4.8 Explanation And Prediction Create A Complete Picture

From the regression analysis, the **HitFlix Movie Rental** kiosk chain owner can now explain how footage drives revenues, and he is equipped to compare predicted revenues at alternate footage levels. In his presentation to management, he would conclude:

“From sample evidence, we conclude that kiosk footage drives kiosk revenues. Variation in footage accounts for 86% of the variation in revenues among a random sample of 52 stores.

With knowledge of square footage, revenue can be estimated with a margin of error of \$15,000.

For each square foot that a kiosk exceeds the average size of 100 square feet, we can expect an average increase in revenue of \$990 to \$1,240.



Comparing expected revenue from a new kiosk at 100 square feet and 130 square feet, the additional thirty feet is expected to generate \$34,000 more revenue, though this could be as little as \$4,000 and as large as \$63,000, as the table illustrates.

$$\hat{revenue}(K\$) = \$3,400^a + \$1,100^a \text{ Footage}$$

(\$6,400)      (\$64)

*RSquare: .86<sup>a</sup>*  
*<sup>a</sup>Significant at .0001*

<i>New Kiosk Footage</i>	<i>Expected Revenue</i>
100	\$100,000 to \$130,000
130	\$134,000 to \$163,000

#### 4.9 Present Regression Results In Concise Format

The HitFlix owner presented results of his regression analysis by illustrating the regression line with 95% confidence prediction intervals on top of the actual data. This demonstrates how well the model fits the data.

He included the regression equation in standard format, with the dependent variable on the left, standard errors under the parameter estimates, RSquare below the equation, and significance levels of the model and parameter estimates indicated with superscripts:

$$\hat{Y} = b_0^a + b_1^a X$$

$$\left( s_{b_0} \right) \left( s_{b_1} \right)$$

$$RSquare = \frac{\quad}{\quad}^a$$

$$^a \text{Significant at } \underline{\quad}.$$

Not everyone who reads his memo will understand these four lines. For the general business audience, the verbal description with graphical illustration conveys all of the important information. The four additional lines provide the information that statistically savvy readers will want in order to assess how well the model fits and which parameter estimates are significant.

#### 4.10 We Make Assumptions When We Use Linear Regression

Linear regression assumes that the dependent variable, which is often a performance variable, is related linearly to the independent variable, often a decision variable. In reality, few relationships are linear. More often, performance increases or decreases in response to increases in a decision variable, but at a diminishing rate. The dependent variable is often limited. Revenues, for example, are never negative and are limited (probably at some very high number) by the number of customers in a market. In these cases, linear regression doesn't fit the data perfectly. Extrapolation beyond the range of values within a sample can be risky if we assume constant response when response is actually diminishing or increasing. Though often not perfect reflections of reality, linear relationships can be useful approximations. In Chapter 11, we will explore simple remedies to improve linear models of nonlinear relationships by simply rescaling to square roots, logarithms or squares.

Linear regression of time series data assumes that the unexplained portion of a model, the residuals, are stable over time. Our predictions do not get better or worse with time. Patterns uncovered in the data are stable over time. Chapter 9 introduces diagnosis of and remedies for *autocorrelated* errors which break this assumption and vary with time.

If we attempt to explain or predict a dependent variable with an independent variable, but omit a third (or fourth) important influence, our results will be misleading. It will seem that the independent variable that we've chosen is more important than it actually is. Often a group of independent variables together jointly influence a dependent variable. If just one from the group is included in a regression, it may seem to be responsible for the joint impact of the group. Chapters 8 and 9 introduce diagnosis of *multicollinearity*, the situation in which predictors are correlated and jointly influence a dependent variable.

## 4.11 Correlation Is A Standardized Covariance

A correlation coefficient  $\rho_{XY}$  is a simple measure of the strength of the linear relationship between two continuous variables,  $X$  and  $Y$ . Our sample estimate of the population correlation coefficient  $\rho_{XY}$  is calculated by summing differences from the sample means  $\bar{X}$  and  $\bar{Y}$ , and standardizing those differences by the standard deviations  $s_X$  and  $s_Y$ :

$$r_{XY} = \frac{1}{(N-1)} \sum_i \frac{(x_i - \bar{X})}{s_X} \frac{(y_i - \bar{Y})}{s_Y},$$

Where  $x_i$  is the value of  $X$  for the  $i$ 'th sample element, and

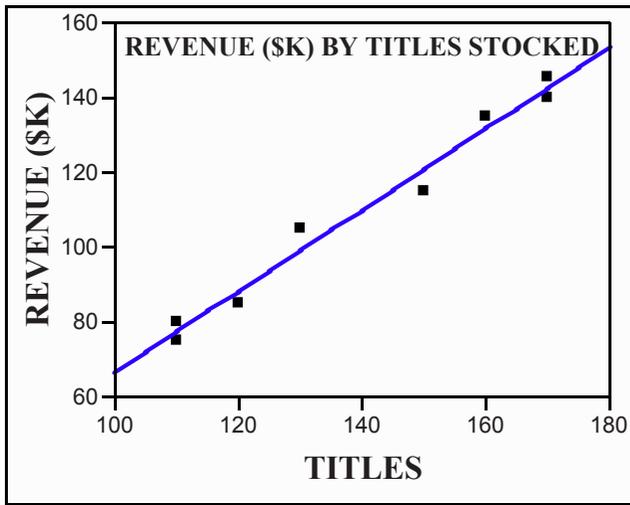
$y_i$  is the value of  $Y$  for the  $i$ 'th sample element.

When  $X$  and  $Y$  move together, they are positively correlated. When they move in opposite directions, they are negatively correlated.

*Example 4.2 HitFlix Movie Rentals.* Table 4.6 contains titles stocked  $X$  and revenues  $Y$  from a sample of eight movie rental kiosks:

<i>kiosk</i>	<i>titles stocked X</i>	<i>revenues (\$K) Y</i>
1	110	75
2	110	80
3	120	85
4	130	105
5	150	115
6	160	135
7	170	140
8	170	145
<i>sample mean</i>	<b>140</b>	<b>\$110</b>

Table 4.6 Titles stocked and revenues (\$K) for eight kiosks



A scatterplot in Figure 4.8 reveals that kiosks which stock more titles also have greater revenues.

Figure 4.8 Movie rental kiosk revenues (\$K) by titles stocked

Differences from the sample means and their products are shown in Table 4.7.

<i>Kiosk</i> $i$	<i>Titles Stocked</i>			<i>Revenues (\$K)</i>			
	$x_i$	$\bar{X}$	$x_i - \bar{X}$	$y_i$	$\bar{Y}$	$y_i - \bar{Y}$	$(x_i - \bar{X})(y_i - \bar{Y})$
1	110	140	-30	\$75	\$110	-\$35	1050
2	110	140	-30	80	110	-30	900
3	120	140	-20	85	110	-25	500
4	130	140	-10	105	110	-5	50
5	150	140	10	115	110	5	50
6	160	140	20	135	110	25	500
7	170	140	30	140	110	30	900
8	170	140	30	145	110	35	1050

Table 4.7 Differences from sample means and crossproducts

The sample standard deviations are  $s_X = 25.6$  square feet and  $s_Y = 28.2$  (\$K).

The correlation coefficient is:

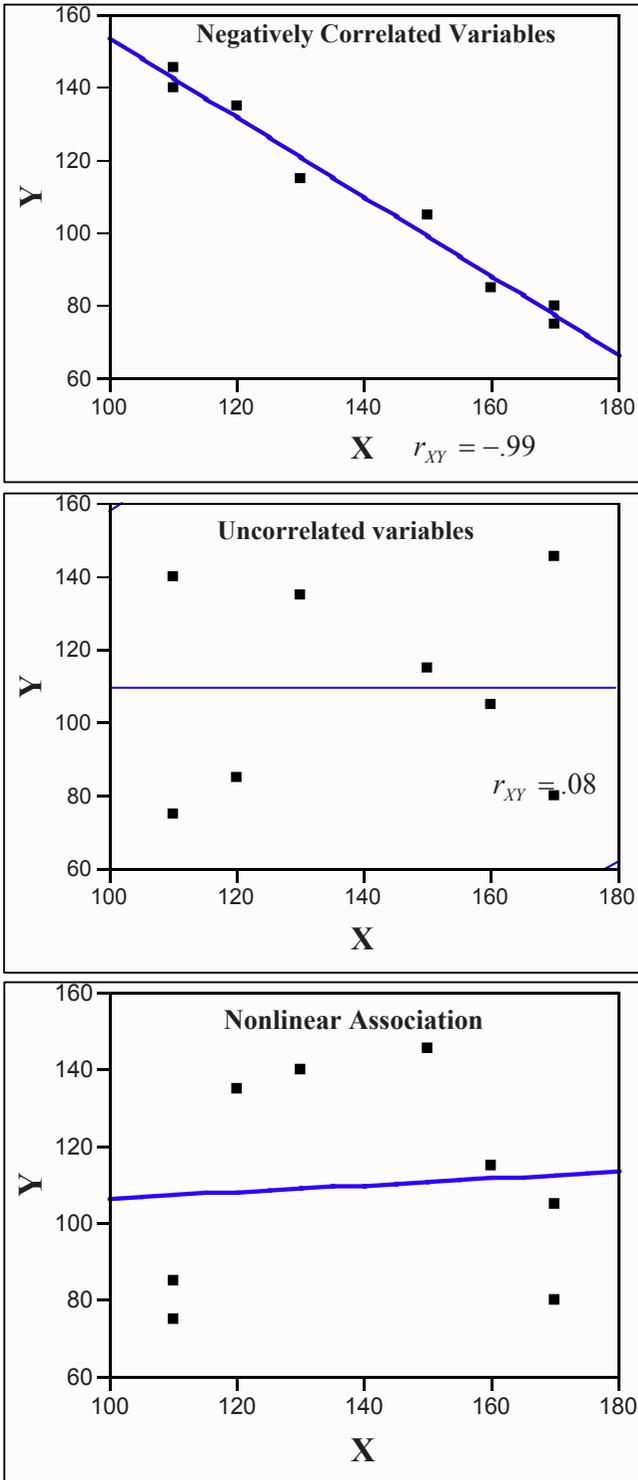
$$\begin{aligned} r_{XY} &= \frac{1}{(8-1)} \left[ \frac{1050+900+500+50+50+500+900+1050}{(25.6)(28.2)} \right] \\ &= \frac{1}{7} [5000/722] \\ &= .990 \end{aligned}$$

A correlation coefficient can be as large in absolute value as 1.00, if two variables were perfectly correlated. All of the points in the scatterplot would fall on top of the regression line in that case. *RSquare*, which is the squared correlation in a simple regression, would be 1.00, whether the correlation coefficient were -1.00 or +1.00.

In the **HitFlix Movie Rentals** example above, *RSquare* is

$$RSquare = r_{XY}^2 = .990^2 = .979$$

If two variables are strongly negatively correlated, their scatterplot looks like the top panel in Figure 4.9. Two scatterplots of uncorrelated variables are shown in the middle and lower panels.



Notice that while  $X$  and  $Y$  are not related linearly in the third panel, they are strongly related. There are situations, for example, where more is better up to a point and improves performance, then, *saturation* occurs and, beyond this point, response deteriorates.

- Without enough advertising, customers will be not aware of a new product. Spending more increases awareness and improves performance. Beyond some saturation point, customers grow weary of the advertising, decide that the company must be desperate to advertise so much, and switch to another brand, reducing performance.
- A factory with too few employees  $X$  to man all of the assembly positions would benefit from hiring. Adding employees increases productivity  $Y$  up to a point. Beyond some point, too many employees would crowd the facility and interfere with each other, reducing performance.

Figure 4.9 Negatively correlated and uncorrelated variables

### 4.12 Correlation Coefficients Are Key Components Of Regression Slopes

As you might suspect, correlation coefficients are closely related to regression slopes. If we know the correlation between  $X$  and  $Y$ , as well as their sample standard deviations  $s_X$  and  $s_Y$ , we can calculate the regression slope estimate:

$$b_1 = r_{XY} \frac{s_Y}{s_X}$$

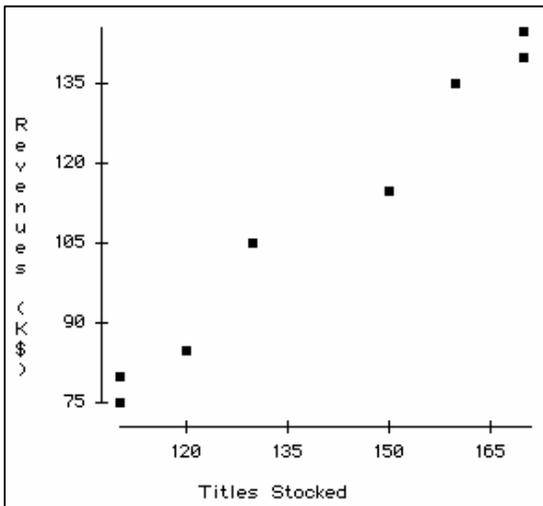
Similarly, if we know the regression slope estimate and sample standard deviations  $s_X$  and  $s_Y$ , we can calculate the correlation coefficient:

$$r_{XY} = b_1 \frac{s_X}{s_Y}$$

In the **HitFlix Movie Rentals** example, the correlation coefficient  $r_{XY} = .99$ , the sample standard errors are  $s_X = 26.5$  and  $s_Y = 28.2$ , so we can calculate the regression slope estimate:

$$b_1 = .99 \frac{28.2}{26.5} = 1.09$$

Correlation coefficients from Excel are shown in Figure 4.10.



<i>Correlation</i>	0.99
<i>t statistic</i>	16.82
<i>p value</i>	< .0001

Figure 4.10 Correlation between revenue and titles

Corresponding simple regression results are shown in Table 4.8.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.99					
R Square	0.98					
Adjusted R Square	0.98					
Standard Error	4.38					
Observations	8					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	5435	5435	283.0	3E-06	
Residual	6	115	19			
Total	7	5550				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-42.174	9.177	-4.6	0.0037	-64.630	-19.718
<i>Titles Stocked</i>	1.087	0.065	16.8	0.0000	0.929	1.245

Table 4.8 Regression of revenue by titles

*Example 4.3 Pampers.* Procter & Gamble hoped that targeted customers who value fit in a preemie diaper would use price as a quality of fit cue and prefer a higher-priced diaper. Ideally, fit importance would be negatively correlated with price responsiveness. In the concept test of the new preemie diaper using a sample of 97 preemie mothers, price responsiveness was measured as the difference between trial intentions at competitive and premium prices, each measured on a 5-point scale (1 = “Definitely Will Not Try” to 5 = “Definitely Will Try”). Fit importance was measured on a 9-point scale (1 = “Unimportant” to 9 = “Very Important”). The correlation between price responsiveness and fit importance from Excel are shown in Figure 4.11:

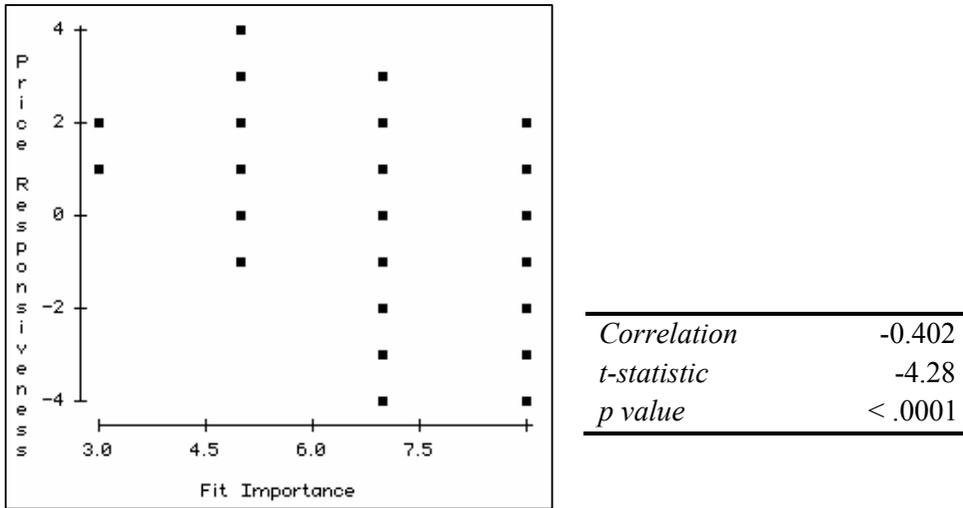


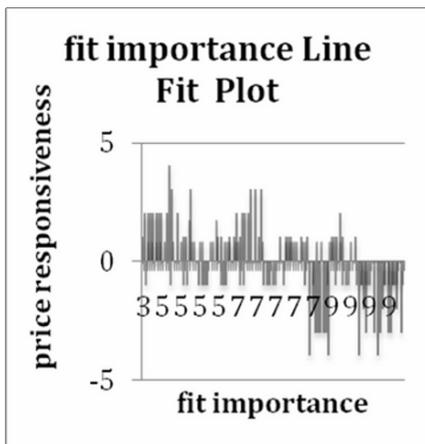
Figure 4.11 Correlation between price responsiveness and fit importance

The correlation between price responsiveness  $Y$  and fit importance  $X$  is moderately large and negative:

$$r_{XY} = -.40$$

The lower the importance of fit to a preemie mom, the greater her responsiveness to a price reduction.

Regression analysis from Excel, shown in Figure 4.12, quantifies this negative, linear relationship:



## SUMMARY OUTPUT

<i>Regression Statistics</i>						
Multiple R	0.402					
R Square	0.161					
Adjusted R Square	0.153					
Standard Error	1.704					
Observations	97					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	53.1	53.1	18.3	5E-05	
Residual	95	275.8	2.9			
Total	96	328.9				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	3.02	0.73	4.1	8E-05	1.56	4.48
<i>Fit Importance</i>	-0.45	0.10	-4.3	5E-05	-0.66	-0.24

Figure 4.12 Regression of price responsiveness by fit importance

From results of correlation and regression analysis, Procter & Gamble management concluded:

*“Price responsiveness is negatively correlated with fit importance of diapers to preemie mothers. Variation in fit importance accounts for 16% of the variation in price responsiveness. Though not a large influence on price responsiveness, fit importance does drive responsiveness, along with other factors. A difference between “Moderately Important” and “Important”, which is a two-point difference on the 9-point importance scale, reduces price responsiveness by about one (.5 to 1.3) scale point on a 11-point responsiveness scale.*

*It is likely that preemie mothers seeking a high quality diaper with superior fit find claims of superior fit at a lower price unbelievable. A higher price supports the higher quality, superior fit image.”*

### 4.13 Correlation Summarizes Linear Association

The correlation coefficient summarizes direction and strength of linear association between two continuous variables. Because it is a standardized measure, taking values between -1 and +1, it is readily interpretable. Unlike regression analysis, it is not necessary to designate a dependent and an independent variable to summarize association with correlation analysis. Later, in the context of multiple regression analysis, the correlations between independent variables will be an important focus in our diagnosis of multicollinearity, introduced in Chapters 8 and 9.

One must be careful to use correlation analysis together with visual inspection of data. It would be possible to overlook strong, nonlinear associations with small correlations. Inspection of a scatterplot will reveal whether or not association between two variables is linear.

Correlation is closely related to simple linear regression analysis:

- The squared correlation coefficient is *RSquare*, our measure of percent of variation in a dependent variable accounted for by an independent variable.
- The regression slope estimate is a product of the correlation coefficient and the ratio of the sample standard deviation of the dependent variable to sample standard deviation of the independent variable.
  - Slope estimates from simple linear regression are unstandardized correlation coefficients.
  - Correlation coefficients are standardized simple linear regression slope estimates.

### 4.14 Linear Regression Is Doubly Useful

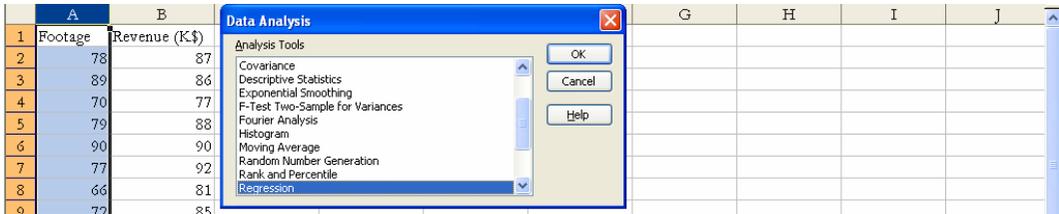
Linear regression handles two modeling jobs, quantification of a driver's influence and forecasting. We build regression models to quantify the direction and nature of influence of a driver on a response or performance variable. We also use regression models to construct forecasts and to compare decision alternatives. This latter use of regression to answer "what if" questions, *sensitivity analysis*, is an important tool for decision making.

## Excel 4.1 Fit a simple linear regression model

*Impact of Footage on HitFlix Movie Rental Revenues.* We will use regression analysis to explore the linear influence of *footage* differences on *revenue* (\$K) differences across a random sample of 52 movie rental kiosks.

Open **Excel 4.1 HitFlix Movie Rental Revenues.xls**.

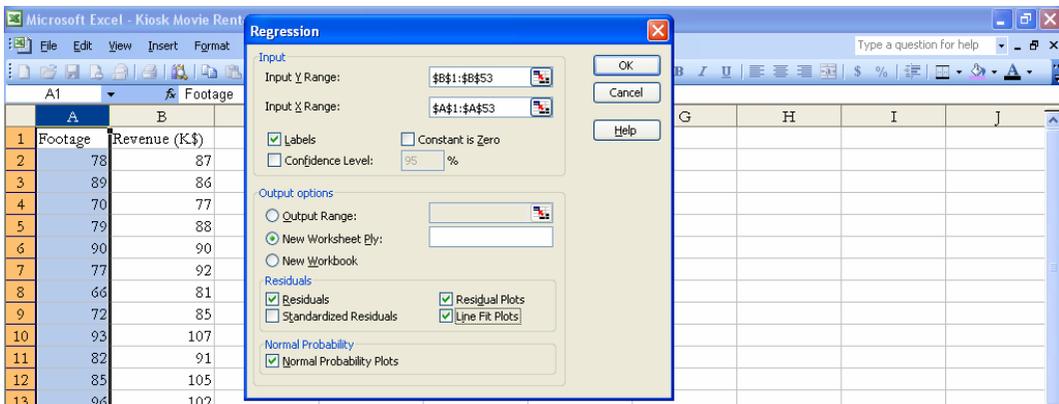
Use shortcuts to run regression: **Alt AY2, Regression, OK:**

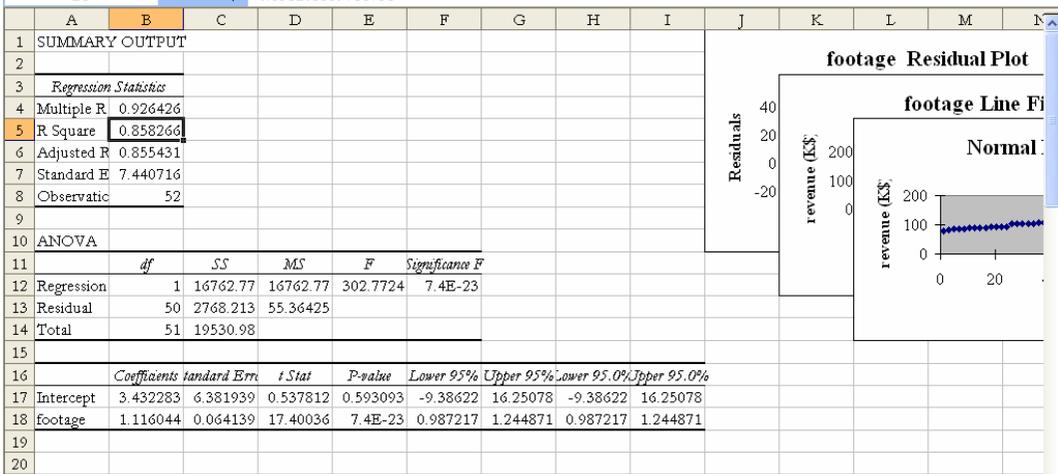


For **Input Y Range**, observations on the dependent variable, *revenues* (\$K), select **B1**, then use shortcuts to select the cells in **B: Cntl+Shift+down arrow to B53**.

For **Input X Range**, observations on the independent variable, *footage*, select **A1**, then use shortcuts to select the cells in **A: Cntl+Shift+down arrow through A53**.

Choose **Labels, Residuals, Residual Plots, and Line Fit Plots, OK:**





The Coefficients  $b_0$  and  $b_{Footage}$ , for the Intercept and the *footage* slope, and their Standard Errors,  $s_{b_1}$  and  $s_{b_{Footage}}$ , allow us to write the regression equation:

	Coefficients	Standard Error
Intercept	3.432283	6.381939
Footage	1.116044	0.064139

$$\text{revenues}(\$K) = 3.43 + 1.12^a \text{ Footage}$$

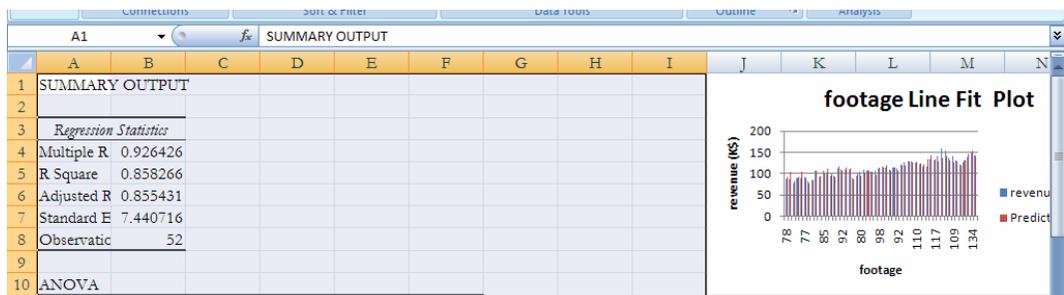
(6.38)    (.06)

R Square: .86

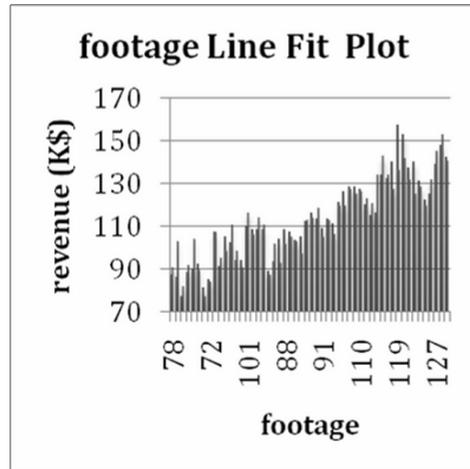
<sup>a</sup>Significant at .01.

In the population of HitFlix movie rental kiosks, the expected difference in *Revenues* due to a unit change of one square foot of *Footage* is in the range .99 to 1.25

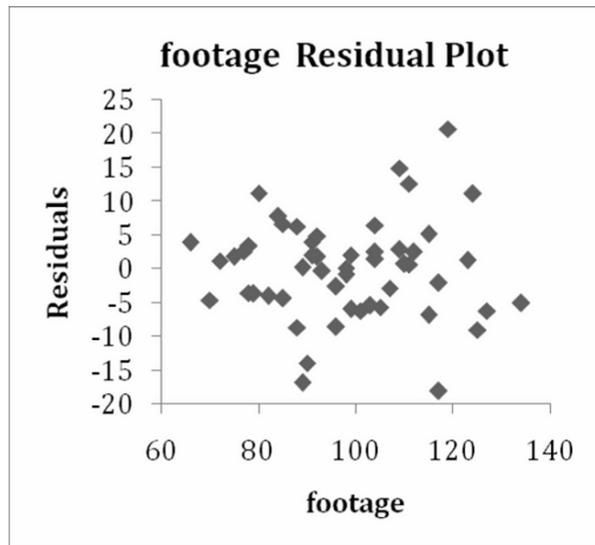
$$(b_{Footage} - t_{50}s_{b_{Footage}} = .99 \text{ and } (b_{Footage} + t_{50}s_{b_{Footage}} = 1.25) .$$



The Line Fit plot suggests that *Revenues (\$K)* do increase at a constant rate with each increase in *Footage*.



The plot of residuals by predicted values is not quite spherical and shows more variation among residuals of larger kiosks. This pattern of nonconstant residual variation, *heteroskedasticity*, may be reduced by rescaling one or both variables to natural logarithms. With heteroskedastic residuals, we expect predictions for smaller kiosks to be more accurate than predictions for larger kiosks.

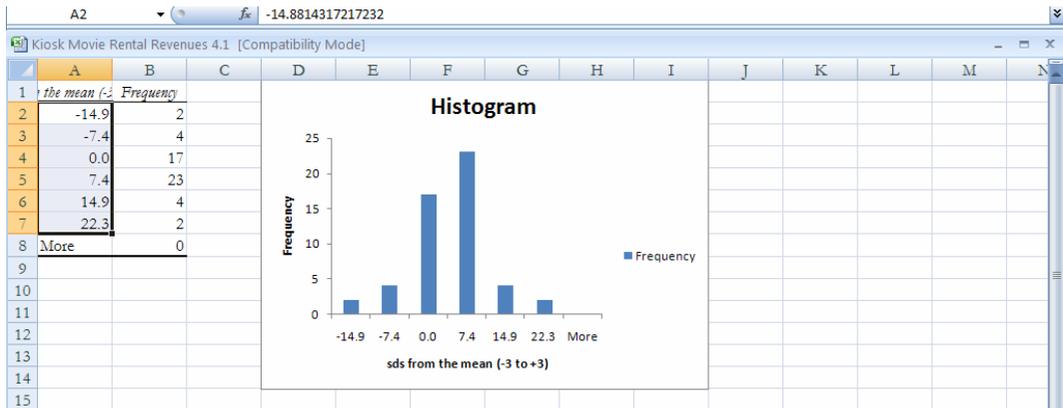


To see the distribution of residuals, copy and paste the **histogram bins.xls** formulas into **G24:I:30**, then replace the standard deviation with the residual standard deviation in **B7**:

In **H25** enter the standard error, or residual standard deviation =**B7** [Enter].

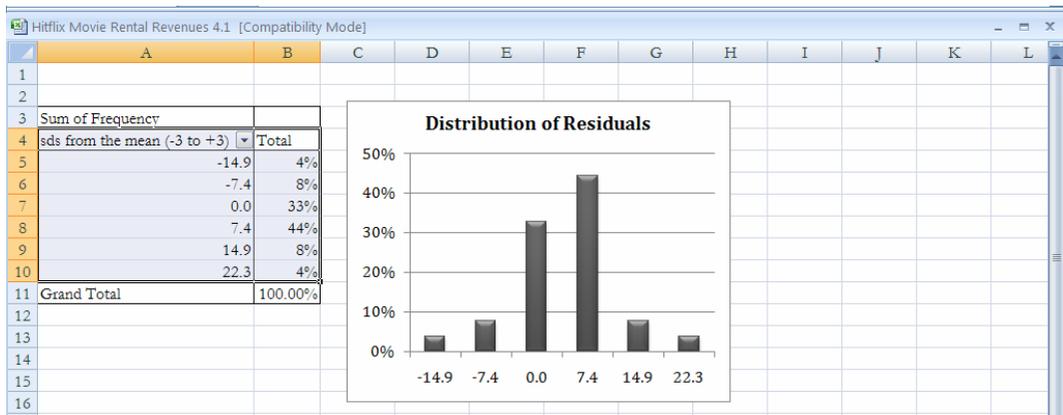
	RESIDUAL OUTPUT			PROBABILITY OUTPUT				
	Observation	revenue	Residuals	Percentile	revenue (K\$)	mean	standard deviation	sds from the mean (-3 to +3)
25	1	90.48372	-3.483724386	0.961538	77	5.46571E-16	7.440716	-14.8814
26	2	102.7602	-16.76020977	2.884615	81			-7.44072
27	3	81.55537	-4.555371381	4.807692	85			5.47E-16
28	4	91.59977	-3.599768512	6.730769	86			7.440716
29	5	103.8763	-13.87625389	8.653846	87			14.88143
30	6	89.36768	2.632319739	10.57692	88			22.32215

Make a histogram of the residuals:



The distribution of residuals is slightly more peaked than *Normal*.

To compare distribution percentages with *Normal* percentages, make a PivotTable and PivotChart:



Too many residuals, 77% (=33%+44%) are within one standard deviation of the mean, which is more than the 67% expected from *Normally* distributed residuals.

### Excel 4.2 Construct prediction and conditional mean prediction intervals

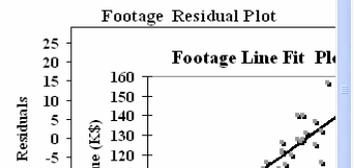
To see 95% prediction intervals for a particular kiosk of specific size, select *Predicted Revenues* in the column that begins in **B23**, copy and paste into column **C** of sheet 1:

RESIDUAL OUTPUT			
Observation	Predicted Revenue (K\$)	Residuals	
1	90.483724	-3.483724	
2	102.76021	-16.76021	
3	81.555371	-4.555371	

	A	B	C	D	E	F	G	H	I
1	footage	revenue (K\$)	Predicted Revenue (K\$)						
2	78	87	90.4837244						
3	89	86	102.76021						
4	70	77	81.5553714						
5	79	88	91.5997685						

Select and copy the *standard error* in **B7** of the regression sheet and paste into cell **D2** of sheet 1, adding the *standard error* label in **D1**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	SUMMARY OUTPUT													
2														
3	Regression Statistics													
4	Multiple R	0.926426												
5	R Square	0.858266												
6	Adjusted R	0.855431												
7	Standard Error	7.440716												
8	Observations	52												



	A	B	C	D	E	F	G	H	I	J	K	L
1	footage	revenue (K\$)	Predicted Revenue (K\$)	standard error								
2	78	87	90.4837244	7.44072								
3	89	86	102.76021									
4	70	77	81.5553714									

To make prediction intervals, we will need the *t* value which corresponds to a 95% confidence level ( $probability=.05$ ) and 50 ( $=N-2$ ) degrees of freedom.

In cell **E1**, enter the label *t*, then use the Excel function **TINV(probability, df)** to ask Excel to look up the *t* value. For **probability**, enter .05 for a 95% level of confidence, and for **df** enter 50 ( $=N-2$ ), the sample size minus two degrees of freedom lost from calculation of the intercept and the slope:

In **E2**, enter =TINV(.05,50) [Enter].

	A	B	C	D	E	F	G	H	I	J	K	L
1	footage	revenue (K\$)	Predicted Revenue (K\$)	standard error	t							
2		78	87	90.4837244	7.44072	2.01						
3		89	86	102.76021								
4		70	77	91.5553714								

Add 95% lower prediction and 95% upper prediction labels in **F1** and **G1**.

In **F2**, type in the formula for the 95% lower prediction bound, the prediction minus the prediction margin of error,

$$95\% \text{ Lower Prediction} = \text{Predicted Revenue}(\$K) - t_{.05,50} * s$$

by entering =C2-E2, press **F4**, enter \*D2, press **F4**, [Enter].

(Your formula will use the *Predicted Revenue* in each row with the *t* value and *standard error* in row 2, because you have locked the cell references for the latter by pressing **F4** to add dollar signs.)

Select the new cell, **F2**, grab the lower right corner, and drag down through row 53, filling in the column.

In **G2**, type in the formula for the 95% upper prediction bound, adding the prediction plus the prediction margin of error,

$$95\% \text{ Upper Prediction} = \text{Predicted Revenue}(\$K) + t_{.05,50} * s$$

by entering =C2+E2, press **F4**, enter\*D2, press **F4**, [Enter].

Select the **G2** cell, double click the lower right corner to fill in the column:

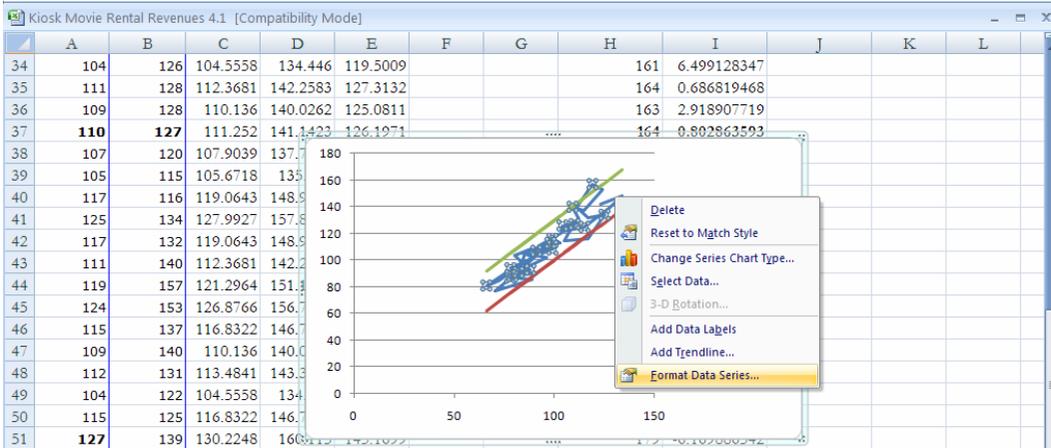
	A	B	C	D	E	F	G	H	I	J	K	L
1	footage	revenue (K\$)	Predicted Revenue (K\$)	standard error	t	95% lower prediction	95% upper prediction					
2		78	87	90.4837244	7.44072	2.01	75.538607	105.42884				
3		89	86	102.76021			87.8150924	117.70533				
4		70	77	81.5553714			66.610254	96.500489				
5		79	88	91.5997685			76.6546512	106.54489				
6		90	90	103.876254			88.9311365	118.82137				

Results from row 2 tell us that revenues for a kiosk with 78 square feet will fall within the interval \$76,000 to \$105,000 with 95 percent certainty.

To see the model fit and prediction intervals, first rearrange columns: Select columns **F** and **G**, use shortcuts to cut those columns, **Cntl+X**, and paste into columns **C** and **D** by selecting column **C**, then **Alt HIE**.

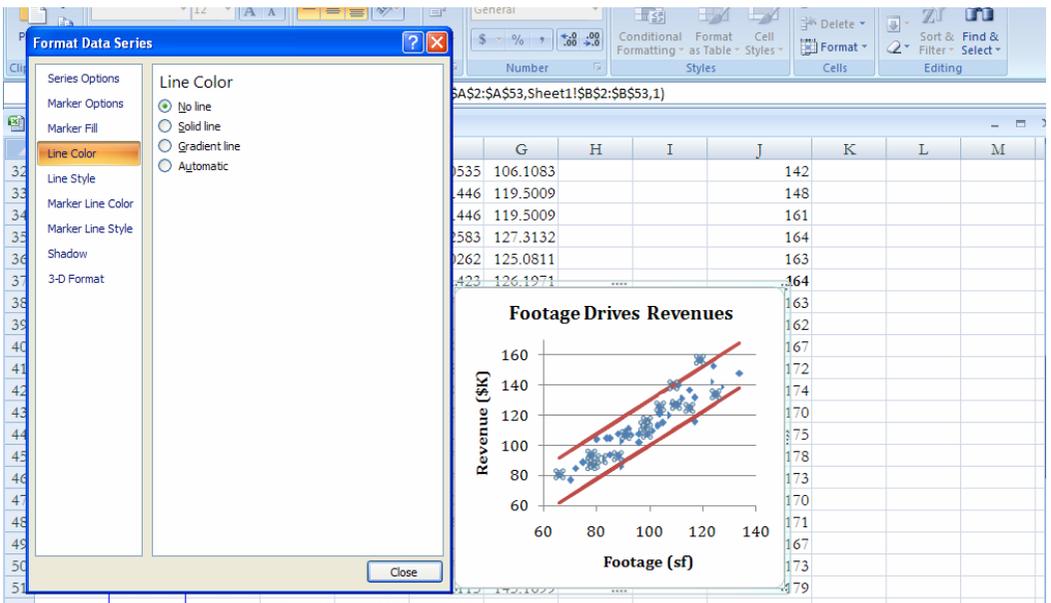
(**Cntl+X** cuts selected cells. **Alt HIE** selects the Home menu and Insert function and inserts cut or copied cells to the left of the selected column or cell.)

Select filled cells in columns **A** through **D**, *footage*, actual *Revenues*, and *95% lower* and *upper* prediction intervals and make a scatterplot:

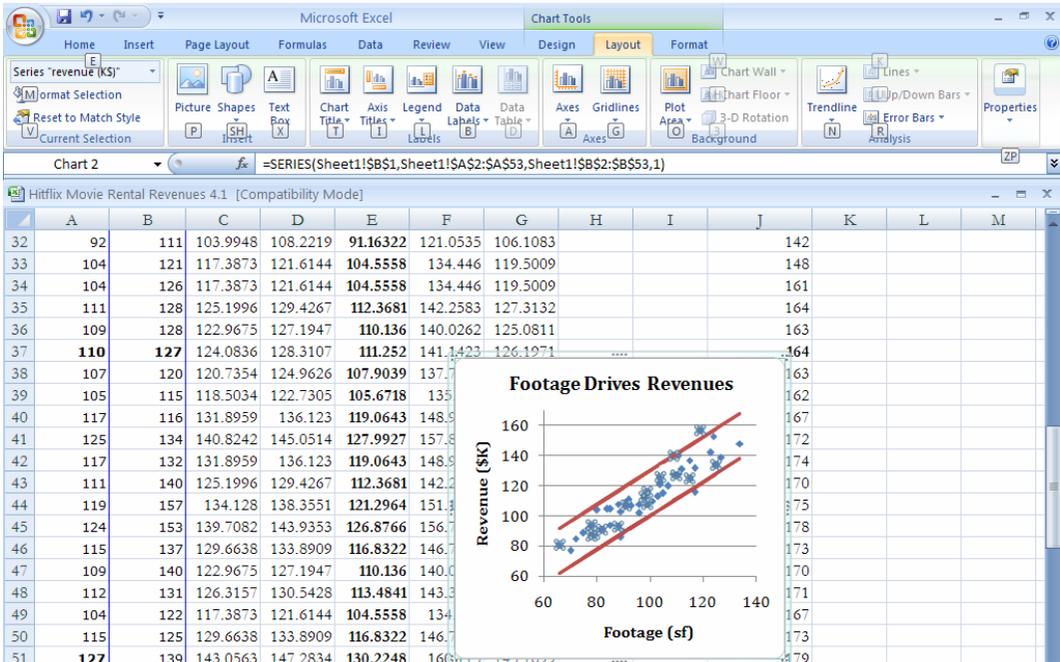


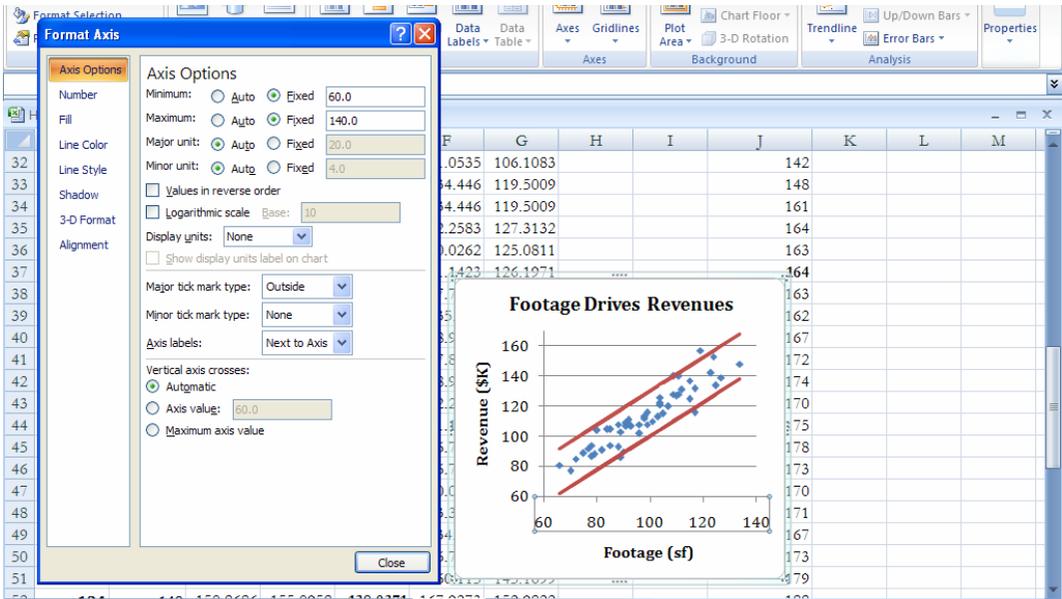
Click the *Revenue* points, right click, then **Format Data Series**.

Choose **Line Color, No Line and Marker Options, Built-In**.



Click one of the 95% confidence lines, right click, **Format Data Series, Line Color, Solid Color**, and recolor to match the other 95% confidence line. Add a title and axes labels using shortcuts: **Alt JAT** and **Alt JAI**.

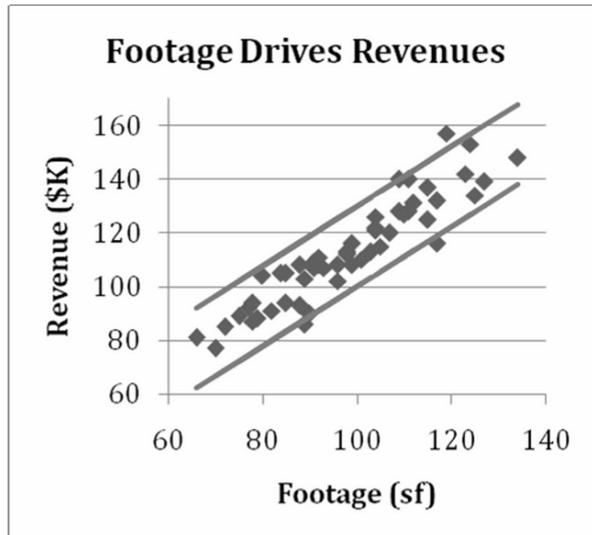




Click the vertical axis, then right click, **Format Axis**, with **Minimum, Fixed, 60** and **Maximum, 170**:

Click the legend and delete.

The model does a good job of predicting actual revenues. Actual revenues for 49 of the 52 kiosks in the sample fall within the 95% prediction intervals. Actual revenues are no further than two standard errors, \$15,000 ( $=2.01 * \$7,400$ ) in 92% ( $=48/52$ ) of the sample kiosks. The prediction margin of error is \$15,000.



To find the 95% conditional mean prediction intervals, add labels *95% conditional mean lower prediction* and *95% conditional mean upper prediction* in columns **H** and **I**.

In **H2** and **I2** enter the formula for the 95% conditional mean lower and upper bounds,

$$95\% \text{ Conditional Mean Lower} = \text{Predicted Revenue}(\$K) \pm t_{.05,50} * s / \sqrt{N}$$

In **H2**, enter **=E2-G2, press F4, \*F2, press F4, /Sqrt(50) [Enter]**.

In **I2**, enter **=E2+G2, press F4, \*F2, press F4, /Sqrt(50) [Enter]**.

Select the new cells **H2:I2**, grab the lower right corner, and drag through the rows to fill in columns.

	A	B	C	D	E	F	G	H	I	J	K	L
1		revenue (K\$)	95% lower prediction	95% upper prediction	Predicted Revenue (K\$)	standard error	t	95% conditional mean lower prediction	95% conditional mean upper prediction			
2	78	87	75.538607	105.42884	90.4837244	7.44072	2.01	88.37016562	92.59728315			
3	89	86	87.8150924	117.70533	102.76021			100.646651	104.8737685			
4	70	77	66.410734	87.550274	81.550274			78.11016562	82.489728315			

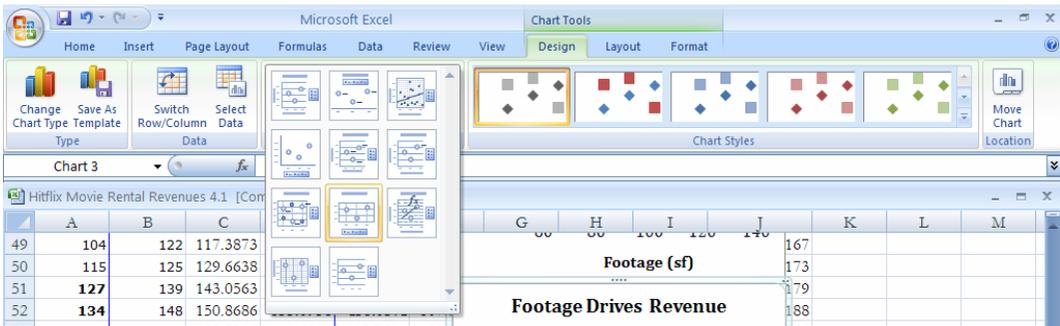
From row 2 we learn that across all kiosks with 78 feet, average revenues will fall between \$89,000 and \$93,000 with 95% certainty.

To see the 95% conditional mean predictions and actual Revenues (\$K) by Footage, rearrange columns:

Select H and I, then cut and paste into C and D.

Select footage, revenue, and 95% conditional mean lower and upper predictions in columns A through D, and insert a scatterplot.

Choose Design, Chart Layout 8:

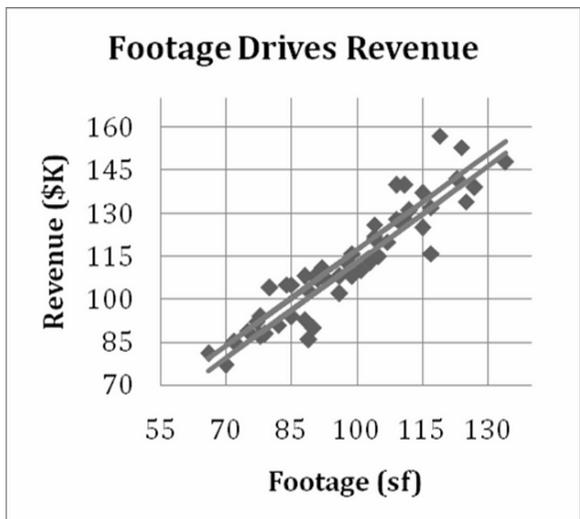


Choose markers for Revenue points and lines for 95% conditional forecasts, adjust both axes scales, and add chart and axis titles:

The HitFlix owner is considering a choice between larger 130 square foot kiosks and average size 100 square foot kiosks for thirty new locations.

We see from the scatterplot that average revenues for the larger size will fall within the interval \$146,000 to \$151,000, while average revenues for the standard size will fall within the interval \$113,000 to \$117,000.

The larger kiosks will most certainly produce higher revenues, though the incremental gain could be as little as \$29,000 (= \$146,000 - \$117,000) or as large as \$38,000 (= \$151,000 - \$113,000).



## Excel 4.3 Find correlations between variable pairs

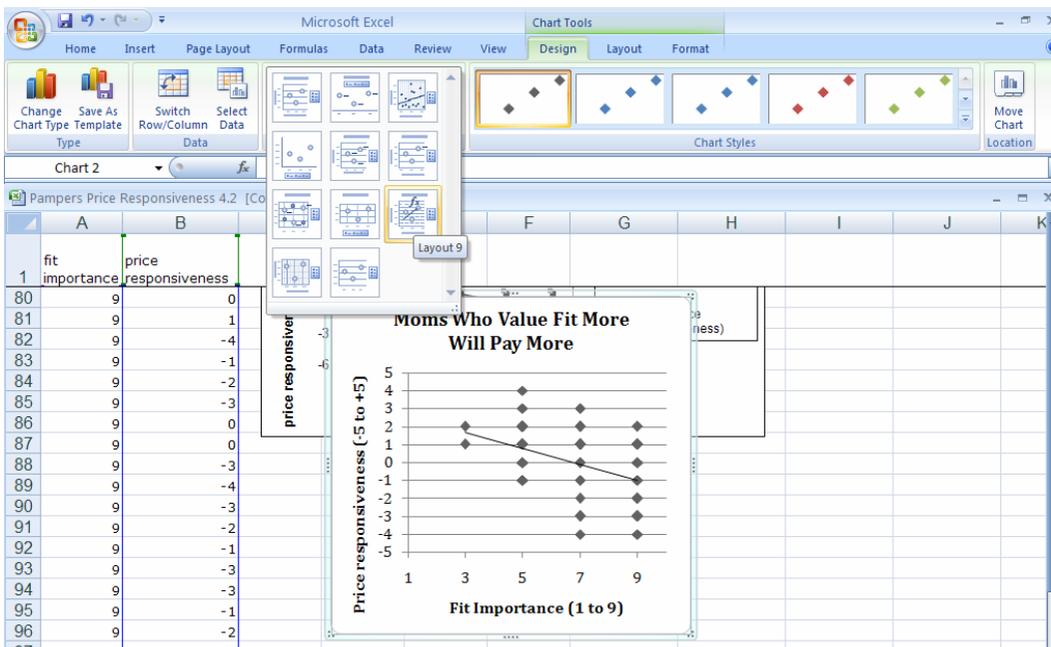
Management would like to know whether there is an association between the perceived importance of diaper fit and price responsiveness among preemie mothers.

*Fit importance* ratings and *price responsiveness* from a concept test sample of 97 preemie mothers are in **Excel 4.3 Pampers Price Responsiveness.xls**.

First plot the two variables with a scatterplot.

Select filled cells in **A** and **B**, then insert a scatterplot.

Select **Design, Chart Layout 9**, which will add the line of best fit, and enter chart and axes titles.



Customize background and markers.

Reformat the vertical axis, right click the axis, **Format axis**, and set **Minimum, Fixed, -5**, **Maximum, Fixed, 5**, **Major unit, Fixed, 1**.

At the bottom of the dataset, in **A99**, enter the label *correlation*, and use the Excel function **CORREL(array1,array2)** to find the correlation between *fit importance* rating and *price responsiveness*.

	A	B	C	D	E	F	G	H	I	J	K
1	fit importance	price responsiveness									
75	5	1									
76	9	-3									
77	9	0									
78	9	0									
79	5	0									
80	7	-4									
81	9	-3									
82	7	-2									
83	9	-4									
84	9	-3									
85	9	-2									
86	9	-1									
87	7	-3									
88	7	-3									
89	7	-3									
90	7	-3									
91	9	-3									
92	7	-3									
93	9	-3									
94	9	-1									
95	9	-2									
96	9	0									
97	7	-4									
98	9	-3									
99	correlation	-0.401832834									

In **B99** enter **=CORREL(A2:A98,B2:B98)[Enter]**:

## Excel Shortcuts at Your Fingertips

### By Shortcut Key

**Alt** activates the shortcuts menus, linking keyboard letters to Excel menus. Press **Alt**, then release and press letters linked to the menus you want.

The following are examples of shortcuts. Press **Alt**, then

**H 9** to select the Home menu and the reduce decimals function

**H DC** to select the Home menu and the Delete function to delete column(s)

**H IC** to select the Home menu and Insert function and to insert a column to the left of the selected cell or column

**HIE** selects the Home menu and Insert function and inserts cut or copied cells to the left of the selected column or cell

**AY2** to select the Data and Data Analysis menus

**AS** to select the Data and the Sort menus

**NC** to select the Insert function and to insert a column chart

**ND** to select the Insert function and to insert a scatterplot

**NE** to select the Insert function and to insert a pie chart

**NVT** to select the Insert function, the Pivot menu, and to insert a PivotTable

**NX** to select the Insert function and to insert a text box

**WFR** to select the View and Freeze panes menus, and to Freeze rows

**JAB** to select the Layout and Data Labels menus

**JARM** to select the Layout, the Error Bar, and the custom Error Bar menus

**JAT** to select the Layout and Title menus

**JAI** to select the Layout and Axis Labels menus

**Shift+arrow** selects cells scrolled over

**Cntl+C** to copy

**Cntl+X** cuts selected cells and places them on the clipboard.

**Cntl+down arrow** scrolls through all cells in the same column that contain data and stops at the last filled cell.

**Cntl+R** fills in values of empty cells using a formula from the first cell in a selected array

**Cntl+Shift+down arrow** selects all filled cells in the column.

## By Goal

If you want to

Activate shortcuts menus, press **Alt**, then release.

Add data labels in a column chart: select a column, then **Alt JAB**

Add error bars in a column chart: select a column, then **Alt JARM**

Add a title: **Alt JAT**

Add axis label: **Alt JAI**

Analyze data: **Alt AY2**

Copy cells: select the cells, then **Cntl+C**

Delete a column: **Alt HDC**

Freeze the top row: **Alt WFR**

Insert copied cells: **Alt HIE**

Insert a column: **Alt HIC**

Insert a column chart: **Alt NC**

Insert a pie chart: **Alt NE**

Insert a PivotTable: **Alt NVT**

Insert a row: **Alt HIR**

Insert a scatterplot: **Alt ND**

Insert a text box: **Alt NX**

Move cells or a column: select the cells or column, **Cntl+X**, then select the new location, **Alt HIE**

Move to the end of a column: **Cntl+down arrow**

Reduce decimals: **Alt H9**

Select all of the filled cells in a column: select the first cell in the column, then **Cntl+Shift+down arrow**

Sort data: **Alt AS**

## Lab 4 Regression

### Dell Slimmer PDA

Dell is considering the introduction of an ultraslim PDA which would fit in a shirt pocket, come in an array of colors and be sold in Wal-Mart's. Dell withdrew its Axim PDA after share fell to 3%. Developers want to be sure that the new PDA will offer the features most desired by the target segments of young, lower income high school students and service workers. Managers believe from past research that there are three PDA lifestyle segments.

- **Younger Players.** The **youngest** segment, **high school** students, who are fashion conscious and technically savvy. Some PDAs in this segment are provided by **higher income** parents. PDAs are primarily used to text message, play music and video games. Penetration in this segment is low.
- **Older Players.** **High school graduates** employed in service jobs. These users are the least technically savvy. PDAs are a luxury used to play music and video games. Penetration in this segment is the lowest.
- **Professionals and Soon to Be. College students and college graduates.** This segment is technically savvy and uses PDA software in classes or on the job. PC connectivity is important, though text messaging and music are also important. This market is saturated and most purchases are upgrades.

Palm and HP cater to the *Professionals and Soon to Be* segments.

**Dell is targeting *Younger and Older Players*, hoping to avoid competition.**

The new PDA would be ultra slim and also fit in a shirt pocket (unlike the withdrawn Axim).

Data from a concept test of 14 to 34 year olds in **Lab 4 Dell Slimmer.xls** include

- a measure of *the importance of thinness and ability to fit in a shirt pocket*, on a 1- to 9-point scale (1=unimportant . . . 9=extremely important)
- key demographics
  - *age*
  - *household income* (in thousands)
  - *years of education*

**Importance of thinness.** Use a one-sample *t test* to determine whether “*thinness*” is an important attribute of PDAs to potential customers like those surveyed. To qualify as an important attribute, average *importance* must be **greater than 5** on the 9-point scale.

A \_\_\_ one tail \_\_\_ two tail *t test* is required.

*p value*: \_\_\_\_\_

Management can conclude that 14 to 34 years olds rate *thinness* important (**at least 5** on a 9 point scale): \_\_\_Y \_\_\_N

Construct a *95% confidence interval* for the average *importance* of “*thinness*” in the **population** and illustrate your result with a clustered column chart.

Average importance of “*thinness*.” \_\_\_\_\_ to \_\_\_\_\_ on a 9-point scale.

**Demographics that drive *thinness importance*.** Use simple regression to identify demographics which drive the *importance* of “*thinness*.”

demographic	<i>p value</i>	drives <i>thinness importance</i>
<i>age</i>		Y or N
<i>education</i>		Y or N
<i>income</i>		Y or N

How much variation in the *importance of thinness* is explained by variation in each of the demographics?  
 (If one or more of the potential drivers is not significant, leave the corresponding row(s) blank.)

demographic	% variation in <i>thinness importance</i> explained
<i>age</i>	
<i>education</i>	
<i>income</i>	

Find the *95% confidence interval* for the difference in *thinness importance* associated with a unit difference in each demographic in the population.  
 (If one or more of the potential drivers is not significant, leave the corresponding row(s) blank.)

demographic	<i>95% lower bound</i>	<i>95% upper bound</i>
<i>age</i> (years)		
<i>education</i> (years)		
<i>income</i> (\$k)		

Illustrate *one* of the significant driver’s influence with a scatterplot showing **population** average difference in response to a unit difference in the driver by adding the line of fit with *95% conditional mean prediction intervals*.

### *CASE 4-1 GenderPay (B)*

The Human Resources manager of Slam's Club was shocked by the recent revelations of gender discrimination by Wal-Mart ("How Corporate America is Betraying Women," *Fortune*, January 10, 2005), but believes that the employee salaries in his company reflect levels of responsibility (and not gender). He has asked you to analyze this hypothetical link between level of responsibility and salary.

He would like to know whether or not *responsibility* drives *salaries*.

If level of *responsibility* drives *salaries*, he would like to know

- the percent of variation in *salaries* which can be accounted for by variation in level of *responsibility*
- the margin of error in forecasts of *salaries* from level of *responsibility*
- with 95% certainty, how much *expected salary in the population* changes with each additional *responsibility* level

The Human Resources manager noticed that many employees are working at *responsibility level 5*. He would like to know

- how much payroll might be reduced if a **level 5** employee were replaced with a new **level 1** employee with similar experience.

Include in your report the *95% prediction intervals* for *salaries* of new employees at both *responsibility* levels, **1** and **5**.

The Human Resources manager is statistically savvy and will want to see

- the regression equation in the standard format
- a scatterplot of *salaries* by level of *responsibility* with
  - the regression line
  - *95% individual prediction intervals*
  - *95% conditional mean prediction intervals*.

Your client is a busy executive and will have only enough time in the near future to read **a single page** of analysis, **single spaced**, in **12 pt font**.

**Case 4-1 GenderPay.xls** contains employee *salaries* and levels of *responsibility* from a random sample of employees.

## CASE 4-2 GM Revenue Forecast<sup>1</sup>

General Motors Management would estimate the percent of customers who will return to again choose a GM car. GM's award-winning customer Loyalty has been widely publicized.

The news release, below, describes their success: (Source: [www.polk.com/News/LatestNews/news\\_011905.htm](http://www.polk.com/News/LatestNews/news_011905.htm))

### Polk Announces Automotive Loyalty Award Winners Numerous New Winners Emerge Across Segment Level Categories for Model Year 2004

**SOUTHFIELD, Mich.** (Jan. 19, 2005) – R. L. Polk & Co., the automotive industry's premier tracker of consumer loyalty among new vehicles, presented the ninth annual Polk Automotive Loyalty Awards yesterday at the 2005 *Automotive News* World Congress.

Capturing loyalty honors for the 2004 model year, which ended September 30, 2004, are Buick, Cadillac, Chrysler, Ford Division, General Motors Corp., Jaguar, Land Rover, Mercury, Lexus, Saturn, Subaru and Toyota.

General Motors won for the fifth consecutive year in the Overall Manufacturer Category. "General Motors' success can be partially attributed to the wide range of vehicle offerings," said Stephen R. Polk, president and CEO of R. L. Polk & Co. "The more vehicle choices an automaker provides a returning customer, the more likely the customer will remain within the manufacturer family."

**TABLE: POLK AUTOMOTIVE LOYALTY AWARD WINNERS - 2004 MODEL YEAR**

Category	Winner (*2003 Winner)	Loyalty %	Avg. Loyalty % for Category
<b>Overall Awards:</b>			
Manufacturer Loyalty	General Motors*	65.1	54.6
Make Loyalty	Ford Division*	57.4	44.8
<b>Vehicle Segment Awards:</b>			
Small Car	Saturn ION	30.6	13.9
Midsize Car	Toyota Camry	28.0	19.2
Large Car	Buick LeSabre	41.2	29.3
Luxury Car	Cadillac DeVille	40.9	18.8
Prestige Luxury Car	Lexus LS 430	32.7	25.9
Sports Car	Ford Mustang	15.0	9.0
Prestige Sports Car	Jaguar XK	22.8	15.4
Minivan	Chrysler Town & Country*	25.8	14.0
Compact Pickup Truck	Ford Ranger*	18.2	13.2
Full-Size Pickup Truck	Ford F-Series*	42.7	36.7
Compact SUV	Subaru Forester*	25.2	16.3
Midsize SUV	Mercury Mountaineer*	32.9	18.1
Full-Size SUV	Ford Expedition*	27.1	22.1
Prestige SUV	Land Rover Range Rover	33.9	21.6

<sup>1</sup> This case is a hypothetical scenario using actual data.

**Case 4-2 General Motors Revenue.xls** contains five years of quarterly data, including:

*quarter,*  
*revenues, revenues*  
*revenues q-4, lagged revenues from four quarters ago,*

Build a simple regression model to estimate the impact of past year *revenues* on current *revenues*.

- a. Present your regression equation in standard format.
- b. What percent of variation in *revenues* can be accounted for by past *revenues*?
- c. How close to actual *revenues* could you expect a forecast to be 95% of the time?
- d. What range in percents of this quarter's GM *revenues* could management be 95% certain to expect will repeat next year?
- e. Present a scatterplot of 95% *individual prediction intervals* with *actual revenues* by *quarter*.

## *Assignment 4-1 Impact of Defense Spending on Economic Growth*

Some experts have suggested that the U.S. economy thrives when the Nation is involved in global conflict. **Assignment 4-1 Defense.xls** contains quarterly *GDP* and past quarter *Defense* spending in billion dollars.

Create a scatterplot and calculate the correlation coefficient to see whether or not *GDP* and *defense spending* are related linearly.

Fit a simple linear regression to estimate the impact on quarter *GDP* of changes in past quarter *defense spending*.

Analyze the residuals.

Are they

- homoskedastic?
- pattern-free?
- approximately *Normally* distributed?

Summarize your results, in a **single-spaced** report, **12 pt font**, with **one embedded figure** and your **regression equation in standard format**.

Choose a title which summarizes your conclusions (the “Bottom Line”).

Use language that policy-makers could easily understand, whether or not they have recently taken statistics.

Include in your report:

- whether or not past quarter *defense spending* is correlated with *GDP*
- the percent of variation in *GDP* that can be explained by variation in past quarter *defense spending*
- the margin of error in forecasts of *GDP* from past quarter *defense spending*,
- the expected range of possible impacts on *GDP* of a **\$1 billion increase** in past quarter *defense spending*

In a technical footnote, include your conclusions from your residual analysis. This should be **no longer than one to three sentences**.

## 5

# Marketing Segmentation with Descriptive Statistics, Inference, Hypothesis Tests and Regression

### *CASE 5-1 Segmentation of the Market for Premie Diapers*

Deb Henretta is about to commit substantial resources to launch *Pampers Premies*. The following article from the *Wall Street Journal* describes Procter & Gamble's involvement in the premie diaper market:

New York, N.Y.  
May 5, 2003

#### **P&G Targets the 'Very Pre-Term' Market** *Wall Street Journal*

*Copyright Dow Jones & Company Inc May 5, 2003*

THE TARGET MARKET for Procter & Gamble Co.'s newest diaper is small. Very small.

Of the nearly half a million infants born prematurely in the U.S. each year, roughly one in eight are deemed "very pre-term," and usually weigh between 500 grams and 1,500 grams (one to three pounds). Their skin is tissue-paper-thin, so any sharp edge or sticky surface can damage it, increasing the chance of infection. Their muscles are weak, and unlike full-term newborns, excessive handling can add more stress that in turn could endanger their health.

Tiny as they are, the number of premature infants is increasing – partly because of improved neonatal care: From 1985 to 2000, infant mortality rates for premature babies fell 45%, says the National Center for Health Statistics. Increasingly, such babies are being born to older or more affluent women, often users of fertility drugs, which have stimulated multiple births.

It's a testament to the competitiveness of the \$19 billion global diaper market that a behemoth like Procter & Gamble, a \$40 billion consumer-products company, now is focusing on a niche that brought in slightly more than \$1 million last year; just 1.6% of all births are very pre-term. But P&G sees birth as a "change point," at which consumers are more likely to try new brands and products. Introducing the brand in hospitals at an important time for parents could bring more Pampers customers, the company reasons.

P&G's Pampers, which is gaining ground on rival Kimberly-Clark, but still trails its Huggies brand, has made diapers for premature infants for years. (P&G introduced its first diaper for "pre-emies" in 1973; Kimberly-Clark in 1988), but neither group had come up with anything that worked well for the very smallest of these preemies.

The company that currently dominates the very-premature market is Children's Medical Ventures, Norwell, Mass., which typically sells about four million diapers a year for about 27 cents each. The unit of Respiroics Inc., Murrysville, Pa., has been making its "WeePee" product for more than a decade. But the company, which also makes incubator covers, feeding tubes and extra small bathtubs for preemies, hadn't developed certain features common in mass-market diapers, such as softer fabric coverings.

By contrast, P&G's preemie diapers, which it started distributing to hospitals in August, sell for about 36 cents each; about four cents more than P&G's conventional diapers. P&G's "Preemie Swaddler" fits in the palm of an adult's hand and has no adhesives or hard corners. It closes with mild velcro-like strips and is made of breathable fabric, not plastic. It has an extra layer of fabric close to the infant's skin to avoid irritation.

Children's Medical Ventures is coming out with another size of the WeePee, and plans to introduce velcro-like closures, a development the company says was in the works before P&G came out with a rival diaper. The new diapers won't cost any more, Children's Medical Ventures says.

P&G says the new diaper is the natural extension of its Baby Stages initiative, which took effect in February 2002 when P&G revamped its Pampers brand in the U.S. to cater to various stages of a baby's development. Working with very small preemies helps the company better understand infant development and become "more attuned to new products they might need," says Deb Henretta, president of P&G's global baby-care division.

But the marketing director for Children's Medical Ventures believes the increasing affluence of preemie parents is a greater inducement for big companies to enter the market. In the past, the typical mother of a preemie was poorer, often a teenager, but today more preemie "parents tend to be older, well-educated, and have money for things like fertility treatments," says Cathy Bush, marketing director for Children's Medical Ventures.

The competition may raise the bar for the quality of diapers for these smallest of preemies. P&G says the parents of premature babies are demanding better products. "They have much higher expectations than they did years ago," Ms. Henretta says.

Neonatal nurses have all sorts of opinions about the relative merits of Preemie Swaddlers and WeePees. Pat Hiniker, a nurse at the Carilion Roanoke Community Hospital in Virginia, says the new Pampers diaper, while absorbent, is too bulky for small infants. Allison Brooks of Alta Bates Hospital in Berkeley, Calif., says P&G's better absorbency made the babies less fidgety when they needed to be changed. "That sounds small, but you don't want them wasting their energy on squirming around," she says. "They need all their energy to grow."

In any case, if health professionals have their way, the very-premature market will shrink, or at least stop growing. The March of Dimes recently launched a \$75 million ad campaign aimed at stemming the rise of premature births. P&G is donating 50,000 diapers to the nonprofit organization.

Reproduced with permission of the copyright owner. Further reproduction or distribution is prohibited without permission.

Before resources are dedicated, Deb wants to confirm that preemie parents are attracted to the *Pampers Preemies* concept of superior comfort and fit. She has commissioned a concept test to assess consumers' intentions to try the product.

## The Market for Preemie Diapers

The market for preemie diapers is unusual in that the first diapers that a preemie baby wears are chosen by the hospital. Procter & Gamble is banking on positive experiences with *Pampers Preemies* in the hospital and consumer brand loyalty once baby goes home. If parents see *Pampers Preemies* in the hospital, are satisfied with their performance, and find them widely available at the right price, parents may adopt the *Pampers Preemies*

brand after the infant comes home. Satisfaction and brand loyalty to *Pampers Preemies* could then lead to choice of other Pampers products as the baby grows. If the concept test indicates that consumers' intentions to try are high, then the results will be included in promotional materials and selling efforts to hospital buyers.

## Preemie Parent Segments

Based on focus group interviews and market research, Deb's team has learned that there are five broad segments of preterm parents:

- Younger (14 to 19), unemployed mothers who live with their parents. These young mothers are inexperienced and their pregnancies are unplanned. They tend to differ widely in their attitudes and preferences, and so a further breakdown is necessary:
  - **Younger, Detached.** These young mothers are relatively unattached to their babies and relatively indifferent about the particular diapers they use. Their means are limited and they are highly responsive to low prices and price promotions.
  - **Younger, Committed.** These young mothers are attached to their babies and want the best diapers. They are inexperienced consumers and could be attracted by a premium diaper, though resources may limit their buying power. Brand name appears to be very important to these young women, and they believe that better mothers choose name brands seen on television.
- **Young** (20 to 35) mothers tend to be married and have adequate resources. Their pregnancies tend to be planned and this segment is virtually indistinguishable from the larger segment of disposable diaper users for full-term babies. This group has the fewest preterm births.
- **Older Victorious Over Biological Clocks** (35 to 39) and **Oldest** (40+) mothers tend to be wealthier, more highly educated professionals with higher incomes. A large proportion has no other children and has undergone fertility treatment. Multiple preemie births are more likely in this segment. Some of these mothers are single parents. This group is particularly concerned about functional diaper features and wants the best diaper their dollars can buy. They are willing to pay for a premium diaper perceived as the highest quality, offering superior fit and comfort.

## The Concept Test

A market research agency has conducted a concept test of *Pampers Preemies* to gauge interest among consumers in a variety of potential target markets. The ninety-seven mothers with preemies who had been born at two local hospitals were asked to fill out a survey about purchase intentions after trying the product on their babies. If that data

supports the launch, Deb will need to know which functional feature(s) to stress in advertising and the type of mother and family to feature in the ads. Therefore, questions regarding attribute importance and demographic information were also collected in the survey.

Data from the concept test is contained in **Case 5-1 Pampers Concept.xls**. Below is an overview of the questions asked in the survey, the manner in which they were coded, and the variable names contained in the dataset (which are in italics).

### 1) Trial Likelihood

Participants were asked, “How likely would you be to try *Pampers Premies* if they were available in the store where you normally buy diapers and were sold at a price of \$X.XX per diaper?”

The question was asked twice at two different price points; a “premium” price of \$0.36 (*premium intent*) and a “value” price of \$0.27 (*value intent*).

Responses were coded as follows:

Definitely Would Not Try	= .05
Probably Would Not Try	= .25
Maybe Would Try	= .5
Probably Would Try	= .75
Definitely Would Try	= .95

### 2) Attribute Importance

Participants were asked, “How important are each of the following attributes to you when choosing a diaper?” for the attributes:

- “brand name” (*brand importance*),
- “comfort/fit” (*fit importance*),
- “keeps baby dry/doesn’t leak” (*staysdry importance*) and
- “natural composition” (*natural importance*),

Responses were given on a scale from 1-9 where “1” = “Not Important at All” and “9” = “Extremely Important.”

### 3) Demographic Information

Consumers were asked to report their age (*age*), annual household income (*income*), family size including the new baby (*family size*), and the number of other children in the home (*other children*).

### Data Re-Coding

Some of the original variables were re-coded to make new variables for analysis.

*Likely and Unlikely Triers*

Two new variables, *premium trier* and *value trier*, were created from the intention to try questions (*premium intent* and *value intent*) to identify “likely triers” of the product at both price points tested. “Likely triers” were identified using a “Top-two-box rule” (i.e., those who indicated that they “Probably” or “Definitely” would try the product). Therefore, for  $premium\ intent \geq .75$ ,  $premium\ trier = 1$ ; otherwise  $premium\ trier = 0$ . Likewise, for  $value\ intent \geq .75$ ,  $value\ trier = 1$ ; otherwise  $value\ trier = 0$ .

**Information Needed**

Deb’s team needs an estimate of revenue potential, plus additional information in four areas.

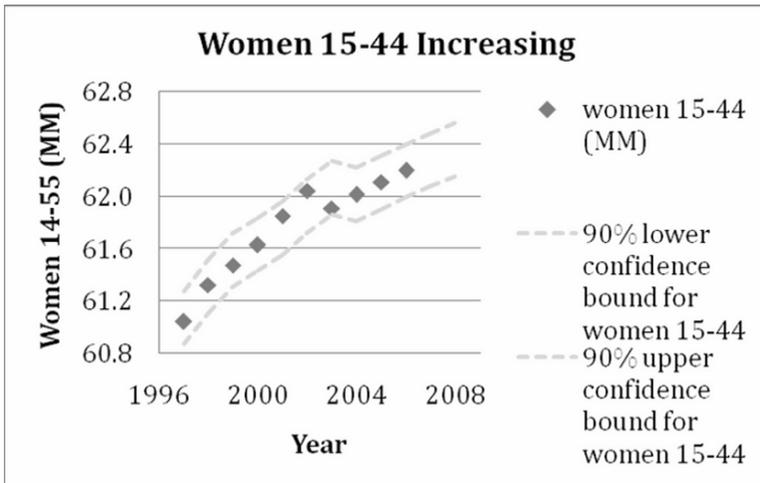
**I. Revenue Potential**

Deb’s team has constructed a spreadsheet, **revenue simulation**, in **Case 5-1 pampers concept test.xls** which links demographic factors to expected revenues in 2008. The logic behind the spreadsheet is explained below.

**Logic behind the Revenue Spreadsheet**

The potential market for Pampers Preemies depends on several key demographic factors. *Births* in a year are a product of *women 15-44* who could have babies and the *birthrate* among those women:

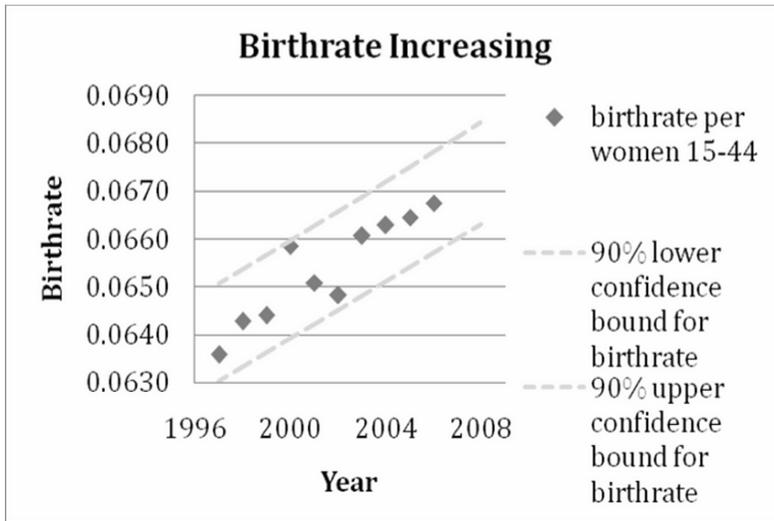
$$births_t = women\ 15-44_t \times birthrate_t$$



The number of women of child-bearing age has been increasing and is expected to lie within the 62.1 to 62.6 million range in 2008.

Greater growth to 62.6 million is more likely, since immigration has been linked to faster growth.

Management believes that the number of women of childbearing age is unlikely to fall below the 2006 level of 62.2 million.



Medical advances and changing demographics, including immigration, have led to an increasing birthrate among women of child-bearing age.

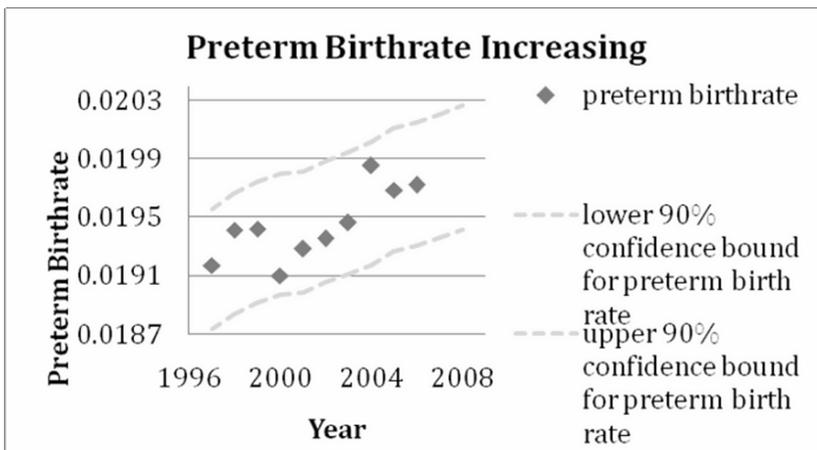
The *birthrate* is expected to lie within the 6.63% to 6.84% range in 2008.

Greater growth (to 6.84%) is more likely.

Management expects the birthrate in 2008 is unlikely to be less than the 2006 birthrate of 6.68%.

The number of *very preterm births* in a year is the product of number of *births* and the chance that a newborn will be very preterm, (*very preterm birthrate*):

$$\text{very preterm births}_t = \text{births}_t \times \text{very preterm birthrate}_t$$



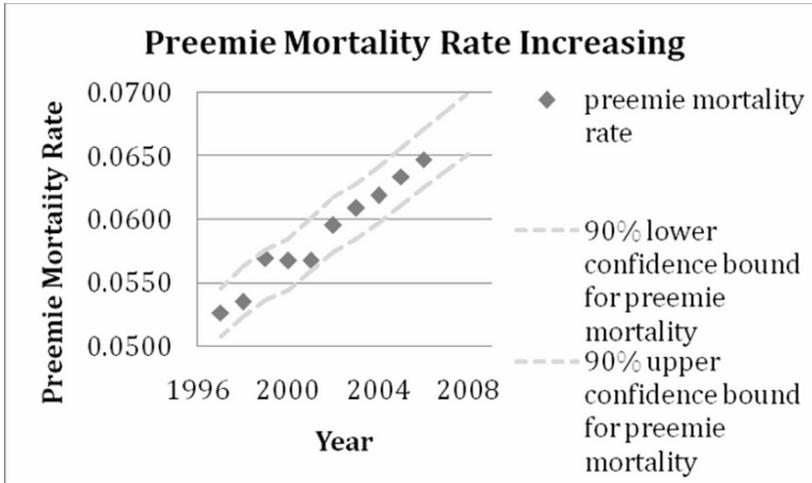
Advances in infertility treatments have led to more births by older, high-risk mothers.

Immigration has led to more births by the youngest mothers, many with little information about prenatal care.

The percent of babies born very preterm has been increasing and is expected to be within the range 1.93% to 2.04% in 2008.

The number of *surviving very preterm babies* is the product of *very preterm births* and the survival rate, which is  $(1 - \text{preterm mortality rate})$ :

$$\text{surviving very preterm babies}_t = \text{very preterm births}_t \times (1 - \text{preterm mortality rate}_t)$$



With the increase in high-risk preterm births, the preterm mortality rate has been increasing and is expected to reach 6.52% to 7.01% in 2008.

The preemie diaper *market* is a product of *surviving very preterm babies*, the average number of days a very preterm baby remains very preterm, approximately 30, and the average number of diapers used per day, approximately 9:

$$\text{market}_t = 30 \times 9 \times \text{surviving very preterm babies}_t$$

Procter & Gamble *revenues* depend on *price*, *market share* (which is expected to vary with *price*), and *market size*:

$$\text{revenue}_t = \text{price} \times \text{market share} \times \text{market}_t$$

From past experience, Procter & Gamble managers have learned that 75% of the proportion of *Likely Triers*, the *trial rate*, become loyal customers in the first year.

$$\text{market share}_t = .75 \text{ trial rate}$$

To be a viable investment, revenue following commercialization of Pampers Preemies must be greater than \$3 MM (million).

1. Estimate of target market segment proportions that are *price responsive* and *not price responsive*.
  - a. Infer the **population proportions** who are *price responsive* and *not price responsive* from *changes in trial intention* in the sample due to change in price from the *premium price* to the discounted, *value price*.

- b. Using *change in likely trial due to discount* in the sample in **Case 5-1 pampers concept test.xls**, compare the *expected population proportions*
- i. *less likely to try (-1) who are Likely Triers at the premium price who become Unlikely Triers at the value price,*
  - ii. *equally likely to try (0) at premium and value prices,*
  - iii. *more likely to try (+1) who are Unlikely Triers at the premium price who become Likely Triers at the value price.*

Illustrate the impact of a price discount with a pie chart of the expected population proportions, noting the *conservative approximate margin of error* in your estimates.

2. Find the chance that *revenues will exceed \$3MM at the premium price* in 2008.
  - a. Infer the *trial rate* (proportion who are *Likely Triers*) in the population from the sample proportion who would try Pampers Preemies at the *premium price* (*Premium Trier=1*).
  - b. Find the *standard error of proportion of Likely Triers*, then calculate the approximate 90% margin of error by multiplying the standard error of the proportion by **1.64**.  
(Note that we are using a *90% confidence interval* so that results can be used in Crystal Ball.)  
Subtract and add the approximate margin of error from the expected *trial proportion* to find the *upper 90%* and *lower 90%* confidence bounds.
  - c. Input the premium price, \$.36, the *lower 90%, expected,* and *upper 90% Likely Trier proportion* into the **revenue simulation** spreadsheet.  
Run a simulation to find the chance of revenues greater than \$3MM in 2008 at the premium price.
3. Find the chance of *revenues greater than \$3MM at the value price*.
  - a. Infer the *expected market share proportion* of the population from the sample proportion who would try Pampers Preemies at the *value price* (*Value Trier=1*).
  - b. Find the *standard error of the market share proportion*, then calculate the approximate 90% margin of error by multiplying the standard error of the proportion by **1.64**. Subtract and add the margin of error from the expected *trial proportion* to find the *upper 90%* and *lower 90%* confidence bounds.
  - c. In the spreadsheet, change the price to the *value price*, \$.27, and change the *lower 90%, expected,* and *upper 90% market share* to reflect the value price.  
run a simulation to find the chance of revenues greater than \$3MM at the value price.

Illustrate the distributions of forecast revenues at *premium* and *value* prices with output from Crystal Ball.

## II. Additional Information Needed

4. Demographic differences between Likely and Unlikely Triers and identification of lifestyle segments most likely to try.
  - a. Test suspected population differences between Likely and Unlikely Triers (premium trier) using a two sample *t test* along each of the following demographics.
    - *Age*
    - *Income*
    - *Family size*
    - *Number of other Children*
  - b. For each significant demographic difference between Trier segments, estimate the extent of difference between Likely and Unlikely Trier segments in the population. The **Q2 Likely v Unlikely** worksheet in **Case 5-1 pampers concept.xls** has been sorted by trier segment for these tests. Illustrate significant differences with a column chart.
  - c. From differences in a), identify the lifestyle segments which you believe will be most attracted to the concept (*younger detached, younger committed, young older victorious over biological clock, oldest*)
5. Identification of attributes likely to be considered important by Likely Triers.

The worksheet page, **Q3 likely triers only**, of **Case 5-1 pampers concept test.xls** contains importance ratings from the segment of *Likely Triers* only.

- a. Determine which attributes are likely to be considered important to the segment of *Likely Triers* (*premium trier=1*) from sample ratings of:
  - *brand importance*,
  - *fit importance*,
  - *staysdry importance*,
  - *natural importance*,

To qualify as an important attribute, the average importance rating for that attribute by the *Likely Trier* segment would be significantly greater than 5 on a 9-point scale.

Illustrate your results with a clustered column chart of the 95% lower and upper confidence interval bounds.

6. Demographics which drive the *importance of "fit"* and lifestyle segment(s) that consider "fit" important.
  - a. Identify those demographics which likely drive the *importance of "fit"* in the **population** of preemie mothers.

- *Age*
- *Income*
- *Famsize*
- *Number of other children*

b. For significant driver:

- How much variation in the *importance of fit* is explained by variation in that driver?
- What **population** average change in *fit importance* is associated with a unit change in that driver?  
(Estimate with *95% confidence intervals* for that driver's coefficient.)

c. Illustrate each significant driver with a scatterplot showing **population** average difference in response to a unit difference in the driver by adding the line of fit with *95% prediction intervals*.

d. From the set of significant drivers, identify the particular lifestyle segments that you believe probably regard *fit* as important (*younger detached, younger committed, young, older victorious over biological clock, oldest*).

## Team Assignment

To prepare for the case discussion, your Team should estimate revenue potential and find the additional information needed by Deb Henretta, listed above.

Each Team is responsible for the presentation of estimated *trial proportion*, *revenue* forecast, and information in one the three additional information areas.

- To facilitate your presentation of the informational item, construct no more than six PowerPoint slides that illustrate your key results, using the guidelines in that follow.
  - Slide 1 introducing your team
  - Slides 2 and 3 summarizing your revenue forecasts
  - Slides 4 through 6 additional information
- Each Team is also responsible for creating a single-spaced memo, using **12 pt font, no longer than two pages**, presenting your estimate of *trial proportion*, *revenue* forecast, and key results from the additional information item assigned. Each page should include **one embedded figure** and follow the format suggested in Chapter 5. You may attach a third page with exhibits if needed.
- Each Team's memo should be accompanied by annotated printout showing that the correct analysis was used and identifying the relevant statistics which led to the results and conclusions.

## 5.1 Guide to Effective PowerPoint Presentations and Writing Memos that your Audience will Read

### Six simple PowerPoint guidelines will enhance your presentations

Your PowerPoint presentations will be more effective if you follow six simple guidelines for content and design.

#### Content

1. ***“Bottom line first, at the top, in the title.”*** Your audience has seconds to digest your slide. The conclusion from your results should be first, at the top, in the title. The “bottom line” is the conclusion that you draw from your results. This will help the audience see what you see right away.
2. ***Limited text.*** Your slide illustrates the story you tell verbally. The focus should be on graphics and key words. Help your audience remember with these. *What you will say during a presentation does not need to appear on the slide.*
3. ***Use graphics, instead of tables.*** Your audience has seconds to process each slide. A figure with key statistics will be more easily understood and remembered.

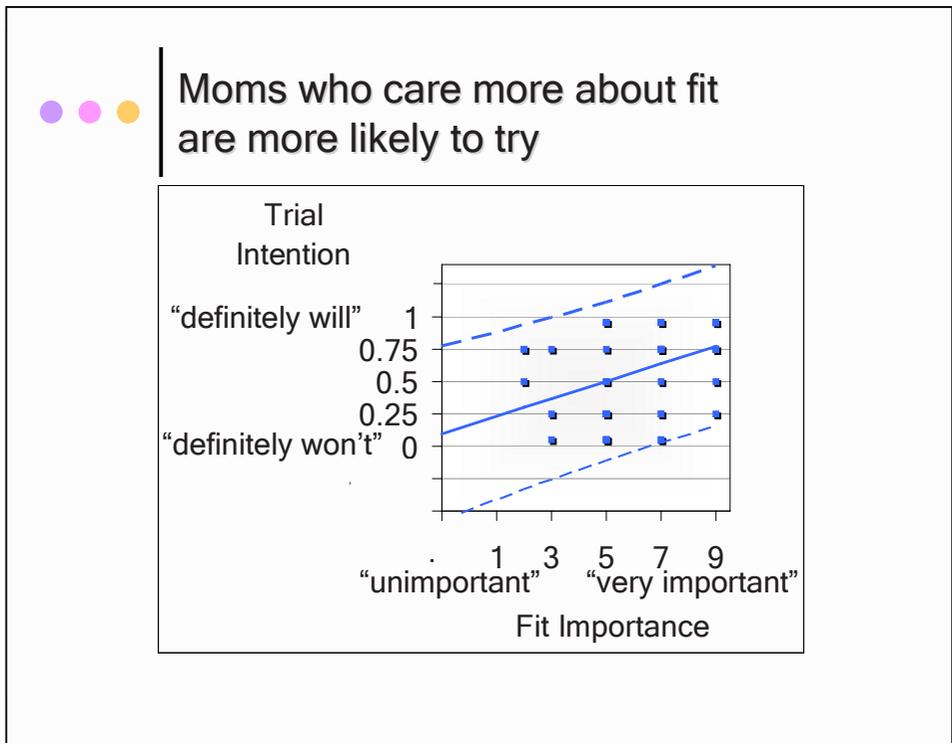
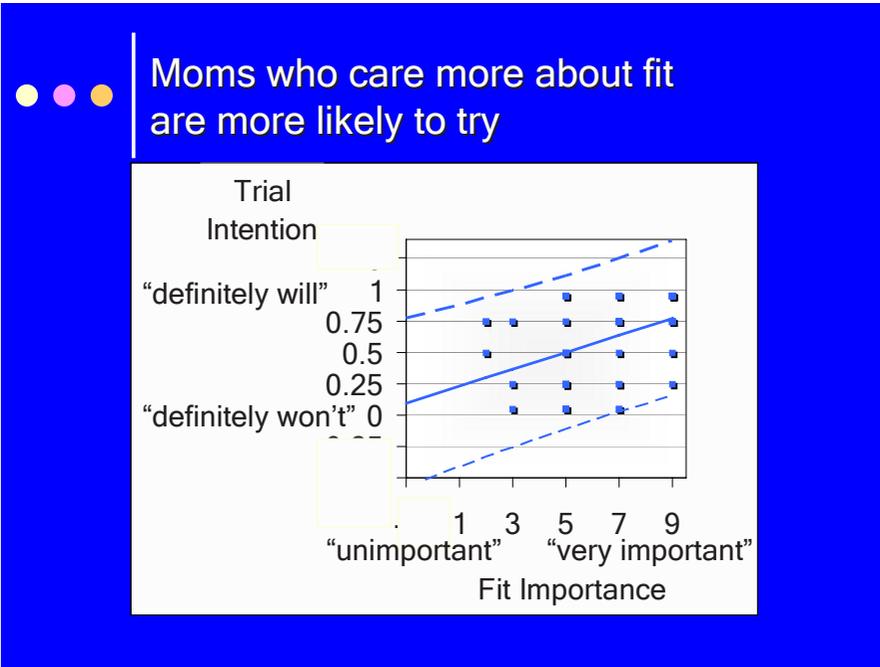
#### Design

##### Font

4. Use ***at least 24 pt. sans serif*** font (Ariel, Lucida, or Garamond). Be sure your title and key words can be easily read by everyone, including those in the back row. *San serif* characters, without extra lines, are clearer in slides. If you have any doubts about readability, test your slides in a room similar in size and shape to the presentation location.

##### Colors

5. ***Background darker, but not dark.*** The background should be darker than the title and key words (*and not white*). It shouldn't be so dark that the audience begins thinking of nighttime and a nap. Choose a medium or darker blue, gray, green or purple background, with complementary, contrasting, lighter text, such as yellow, lighter blue, lighter gray or white. Students (and faculty) are accustomed to writing reports with black text on white backgrounds. This combination looks great in hardcopy but is difficult to see in a PowerPoint slide. View the two PowerPoint slides in **PowerPoint Design 5.1.ppt** to compare the same slide with light on dark and dark on light:



6. **Limit the number of colors** on a slide. If we see more than five colors on a slide (including text), our brains overload and we have difficulty processing the message and remembering it.

Following these six simple guidelines will help you to produce professional PowerPoint slides that command attention, help you deliver your message effectively, and encourage audience members to remember that message.

## 5.2 Write Memos that Encourage Your Audience to Read and Use Results

Memos are the standard for communication in business. They are short and concise, which encourages the intended audience to read them right away. Memos which present statistical analysis to decision makers

- feature the bottom line in the subject line,
- quantify how the bottom line result influences decisions
- are ideally confined to one single-spaced page
- include an attractive, embedded graphic which illustrates the key result.

Many novice analysts copy and paste pages of output. The output is for consumption by analysts, whose job it is to condense and translate output into general business language for decision makers. Decision makers need to be able to easily find the bottom line results without referring to a statistics textbook to interpret results. It is our job to explain in easily understood language how the bottom line result influences decisions. For the quantitative members of the audience, key statistics are included.

On the following page is an example of a memo which might have been written by the quantitative analysis team at Procter & Gamble to present a key result of a concept test of Pampers Preemies to brand management.

Notice that

- the subject line contains the bottom line result,
- the regression analysis tables are omitted,
- results are illustrated with a scatterplot of the fit and
- described in general business English.
- The regression equation is visible for the quantitatively adept, who are assumed to be a minority proportion of the readers.

Description of the concept test and results are condensed and translated. Brand management learns from reading the memo what was done, who was involved, what results were, and what implications are for decision making.

# MEMO

**Re: Importance of Fit Drives Trial Intention**  
**To: Pampers Preemies Management**  
**From: Procter & Gamble Quantitative Analysis Team**  
**Date: October 2007**

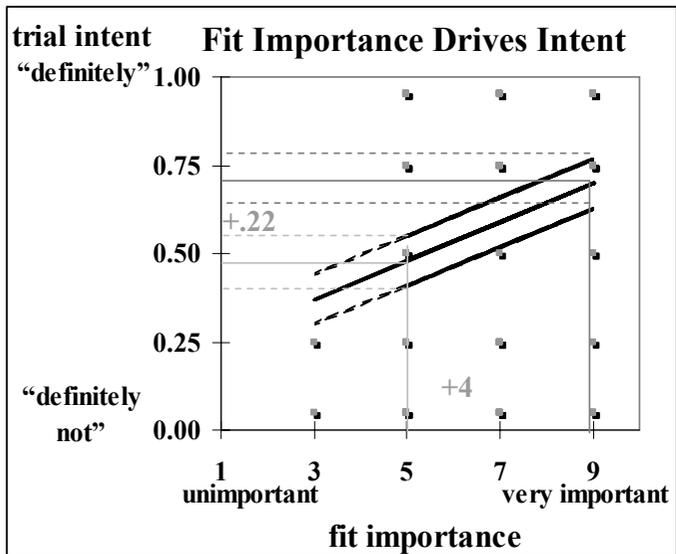
**summary:** Results of a concept test of the Pampers Preemies suggest that the *Importance of fit* drives trial intentions, supporting the expected market salience of superior diaper fit.  
 what was done & learned

**The Concept Test Sample.** The Preemies concept with premium price was described to a convenience sample of 60 preemie mothers in three hospitals in Cincinnati during the week of August 10-14, 2007. Demographics of this sample mirror national demographics of preemie mothers, suggesting that results are representative of all preemie mothers.  
 where the data came from: sample & method

**Concept Test Measures.** The mothers indicated intent to purchase on a five-point scale (.05 = “Definitely Won’t Try” . . . .95 = “Definitely Will Try”) and rated the importances of diaper attributes, including fit, brand, capability to protect from insults, and natural composition on balanced 9-point scales (1 = “Unimportant” . . . .9 = “Very Important”).  
 data & scales

**Concept Test Results.** Differences in fit importance account for 6% of the differences in trial intention.  
 results in English

Comparing mothers who rate fit moderately important (5 on the 9-point scale) with those who rate fit very important (9), the difference in intention could be as low as .08 or as high as .36, and is expected to be .22, which translates into the difference between mothers who “might try” and “probably will try.”



$$Int\acute{e}nt\acute{o}Try = .21 + .054^a \text{ fit importance}$$

(.15) (.021)

*a*Significant at .01      *RSquare: 6%<sup>a</sup>*

**Conclusions.** Targeting mothers who value fit can increase the proportion of triers noticeably. Offering exceptional fit promises to deliver a salient feature.  
 conclusions

**Other Potential Drivers.** Other attributes, including brand, composition, capability to keep baby dry, and price, probably also affect intent. Demographics are likely to affect diaper attitudes, as well as intent to try Pampers Preemies.  
 what else might matter

## 6

### Finance Application: Portfolio Analysis with a Market Index as a Leading Indicator in Simple Linear Regression

We can use simple linear regression of stock rates of return with a Market index to estimate *betas*, measures of risk, which are central to finance investment theory.

Investors are interested in both the mean and the variability in stock price growth rates. Preferred stocks have higher expected growth—expected *rates of return*—shown by larger percentage price increases over time. Preferred stocks also show predictable growth—low variation—which makes them less risky to own. A portfolio of stocks is assembled to diversify risk, and we can use our estimates of portfolio beta to estimate risk.

#### 6.1 Rates of Return Reflect Expected Growth of Stock Prices

*Example 6.1 Goldman Sachs and Yahoo Returns.* Figure 6.1 contains plots of share prices of two well-known companies, Goldman Sachs and Yahoo, over a 58-month period, January 2002 to September 2006. To each graph, the value of a *risk-free* investment has been added. Investment in Treasury bills guarantees a 5% annual return. Their monthly return is certain, and hence, *risk-free*. Had an investor invested \$81 in Treasury bills in January 2002, instead of one share of Goldman Sachs stock, the value of the risk-free investment would be guaranteed to increase 5% annually, or about .4% each month. The risk-free investment value [approximately \$81  $(1.004)^{\text{MONTHS SINCE } 1/02}$  if compounded monthly] allows an investor to see the expected gain from purchase of more variable, risky stock.

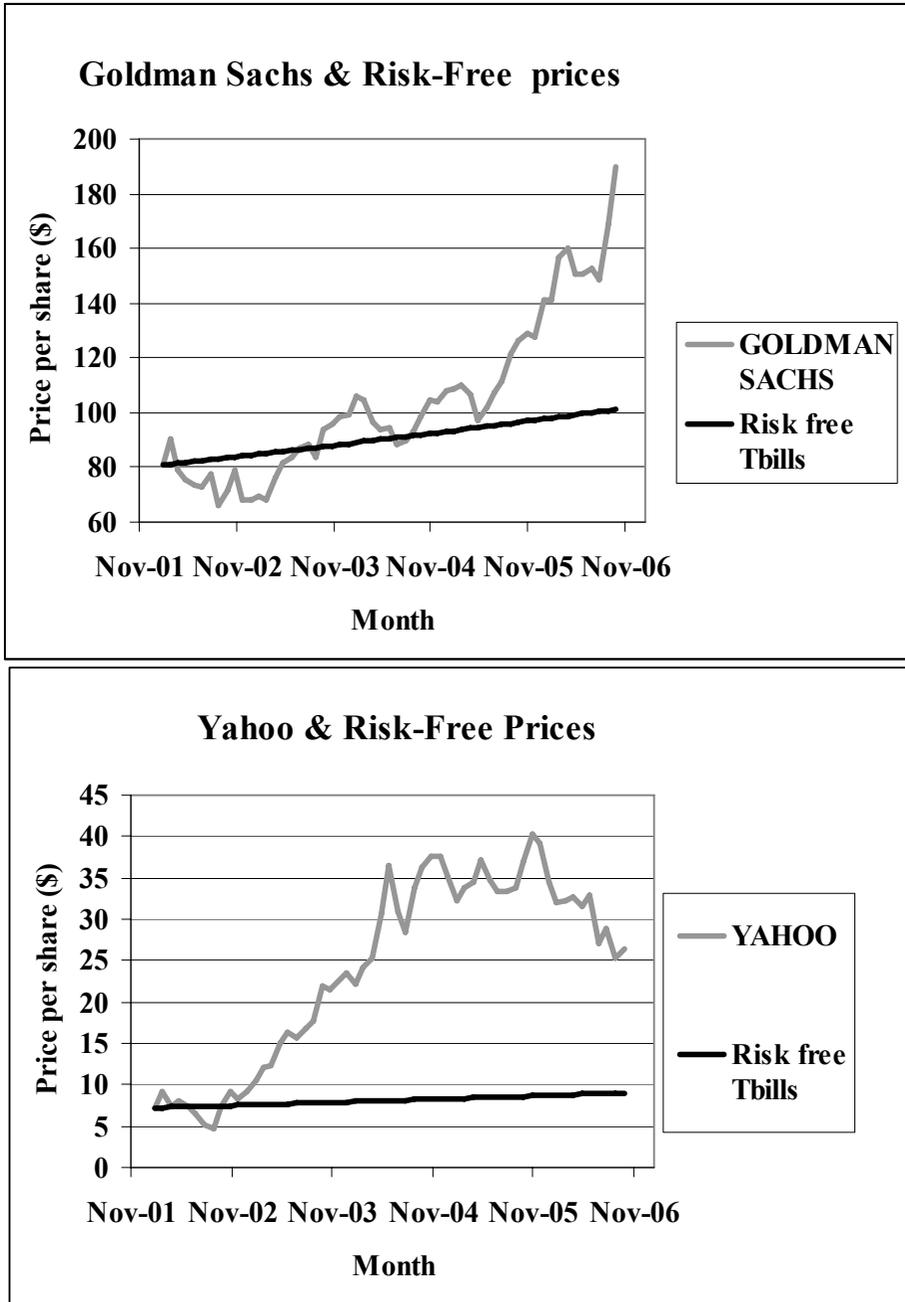


Figure 6.1 Monthly share prices of Goldman Sachs and Yahoo, January 2002 to September 2006

These scatterplots suggest that over the five year period, both stocks offer higher expected returns than the risk-free rate. An investor would have earned more by purchasing a share of either stock instead of buying a risk-free 5% bond, though she

would have to worry about a potential drop in the price of the stock and consequent loss of value in her investment.

It is important to note that although prices in some months were statistical outliers, those unusual months were not excluded. We would mislead a potential investor were we to ignore unusually high or low prices. Extreme values are expected and included, since they influence our conclusions about the appeal of each potential investment. The larger the number of unusual months, the greater the dispersion in a stock price, and the riskier the investment.

To find the growth rate in each of the stock investments, we calculate the monthly percent change in price, or *rate of return*,  $RR$ :

$$RR_{stock,t} = \frac{(price_{stock,t} - price_{stock,t-1})}{price_{stock,t-1}}$$

where  $t$  is time period (month).

Investors seek stocks with higher average rates of return and lower standard deviations. They would prefer to invest in stocks that exhibit higher expected, average growth and less volatility or risk. The standard deviation in the rate of return captures risk. If a stock price shows little variability, it is a less risky investment.

The scatterplot Goldman Sachs, Yahoo, and risk-free bond monthly rates of return in Figure 6.2 illustrates trends over the five year period:

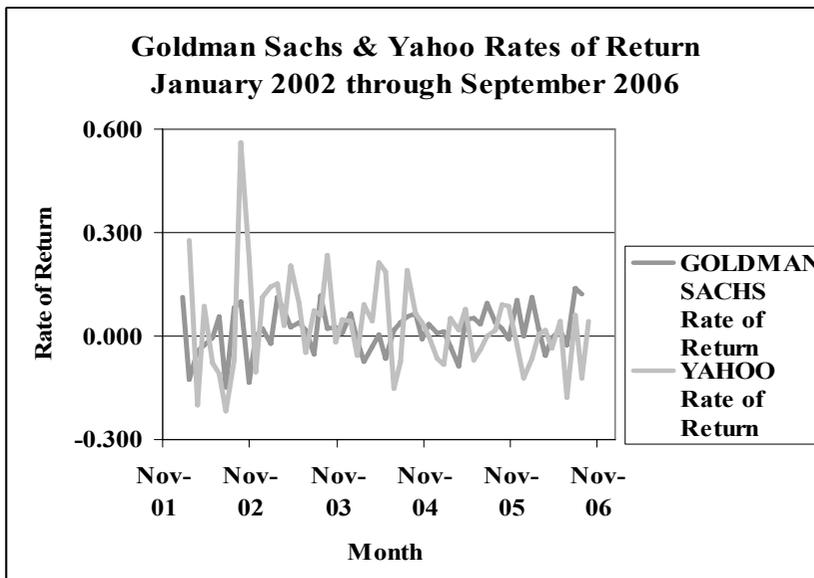


Figure 6.2 Monthly rates of return of Goldman Sachs and Yahoo, January 2002 to September 2006

<b>GOLDMAN SACHS</b>		<b>YAHOO</b>	
<b>Rate of Return</b>		<b>Rate of Return</b>	
<i>Mean</i>	0.017	<i>Mean</i>	0.031
<i>Standard Deviation</i>	0.064	<i>Standard Deviation</i>	0.133
<i>Minimum</i>	-0.146	<i>Minimum</i>	-0.219
<i>Maximum</i>	0.138	<i>Maximum</i>	0.559

Table 6.1 Monthly Rates of Return of Goldman Sachs and Yahoo Stock, January 2002 to September 2006

From Table 6.1, we see that Yahoo's mean monthly rate of return of 3.1% exceeds Goldman Sach's mean monthly rate of return of 1.7%, though Yahoo stock prices are more volatile: the standard deviation in monthly rates of return is .13, compared with Goldman Sach's standard deviation of .064. The greater expected return from Yahoo comes at the cost of added risk. Expected rates of return of both stocks greatly exceed the risk free rate of 5% (which is .41% per month).

We would report to a potential investor:

- *Over the 58 months examined, Yahoo offers a greater expected rate of return of 3.1%, relative to Goldman Sach's expected monthly return of 1.7%, but at higher risk with standard deviation in return (.13 versus .064).*
- *Both Goldman Sachs and Yahoo stocks promise higher rates of return than risk-free investments over the 58 month period examined.*

## 6.2 Investors Trade Off Risk And Return

Investors seek stocks which offer higher expected rates of return  $RR$  and lower risk. Relative to a Market index, such as the S&P 500, a composite of 500 individual stocks, many individual stocks offer higher expected returns, but at greater risk. Market indices are weighted averages of individual stocks. Like other weighted averages, a Market index has an expected rate of return in the middle of the expected returns of the individual stocks making up the index. An investor attempts to choose stocks with higher-than-average expected returns and lower risk.

## 6.3 Beta Measures Risk

A Market index reflects the state of the economy. When we regress a time series of an individual stock's rates of return against a Market index, the simple linear regression slope  $\beta_i$  indicates the expected change in a stock's rate of return in response to a unit change in the Market rate of return. We estimate  $\beta_i$  with  $b_i$  using a sample of stock prices:

$$\hat{RR}_{stock_i,t} = b_0 + b_1 RR_{Market,t}$$

Where  $RR_{stock_i,t}$  is the estimated rate of return of a stock  $i$  in month  $t$ , and

$RR_{Market,t}$  is the rate of return of a Market index in month  $t$ .

In this specific case, the simple linear regression slope estimate  $b_1$  is called *beta*. If, in response to a unit change in the Market rate of return, the expected change in a stock's rate of return  $b_1$  is greater than one, the stock is more volatile, and exaggerates Market movements. A one percent increase in the Market value is associated with an expected change in the stock's price of more than one percent change. Conversely, if the expected change in a stock's rate of return  $b_1$  is less than one, the stock dampens Market fluctuations and is less risky. A one percent change in the Market's value is associated with an expected change in the stock's price of less than one percent. Beta reflects the amount of risk a stock contributes to a well-diversified portfolio.

We know from Chapter 4 that the sample correlation coefficient between two variables  $r_{XY}$  and their sample covariance  $cov_{XY}$  are closely related to the simple regression slope estimate  $b_1$ :

$$b_1 = r_{XY} \frac{s_Y}{s_X} = cov_{XY} / s_Y^2$$

In a Leading Indicator model of an individual stock's rate of return against a Market index, our estimate of beta is directly related to the sample correlation and sample covariance between the individual stock's rate of return and the Market rate of return:

$$beta_{stock_i} = b_{stock_i} = r_{stock_i,Market} \frac{s_{stock_i}}{s_{Market}} = cov_{stock_i,Market} / s_{Market}^2$$

Our estimate of beta is a direct function of the sample covariance between an individual stock's rate of return and the Market rate of return, as well as Market sample variance. Stocks with rates of return that are more strongly correlated with the Market rate of return and those with larger standard deviations have larger betas.

*Example 6.2 Four diverse stocks.* To illustrate the relationship between individual stocks' covariances and correlations with the Market and their betas, monthly rates of return for Lockheed Martin, General Electric, Apple and IBM are plotted in Figure 6.3 with monthly S&P500 rates of return over the two year period from September 2003 through October 2005.

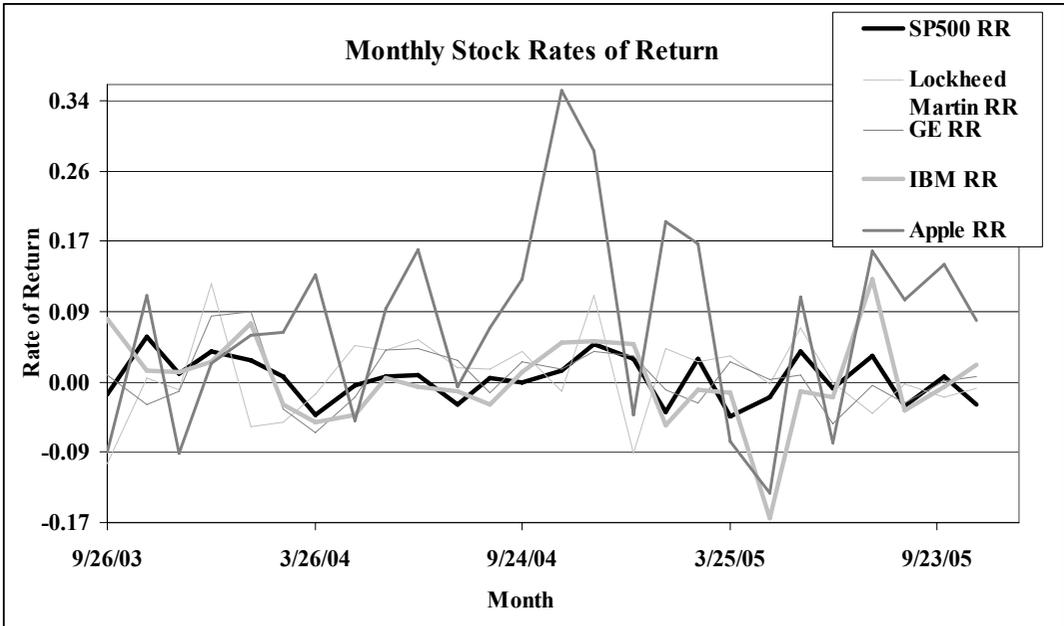


Figure 6.3 Monthly rates of return of four diverse stocks and S&P500 November 2000 – October 2005

Lockheed Martin and General Electric (fainter) in have smaller variances than the computer stocks (thicker). Lockheed Martin and General Electric are less risky investments. We also see that Lockheed Martin (fainter light) moves independently of the Market (black), while the other three tend to move with the Market.

We would expect Lockheed Martin to be relatively immune to economic swings, since much of their business consists of government contracts. We would also expect the two computer stocks to be riskier than General Electric, since the computers (MP3 players, software) are relatively expensive, luxury items. In boom cycles, the computer companies do more business. General Electric sells many necessities, including appliances and light bulbs. The demand for these products is affected less by economic swings, making GE stock relatively less correlated with Market swings, and, hence, less risky.

Only Lockheed Martin returns move opposite the Market and are negative in about a third of the months when the Market is gaining. Market returns never exceed ten percent, while individual stocks sometimes gain as much as thirty-four percent. Market losses are never greater than ten percent, while individual losses are as great as seventeen percent.

Table 6.2 contains sample correlation coefficients, covariances, and betas for each of the four stocks using five years of monthly data (December 2000 through October 2005).

	<i>correlation with the Market</i> $r_{stock,Market}$	<i>standard deviation</i>	<i>covariance with the Market</i> $COV_{stock,Market}$	<i>beta</i> $b_{stock}$
<i>SP500 RR</i>		0.047		
<i>Lockheed Martin RR</i>	-0.13	0.064	-0.00038	-0.18
<i>GE RR</i>	0.407 <sup>a</sup>	0.064	0.00119	0.55 <sup>a,b</sup>
<i>Apple RR</i>	0.416 <sup>a</sup>	0.138	0.00265	1.22 <sup>a</sup>
<i>IBM RR</i>	0.681 <sup>a</sup>	0.100	0.00313	1.45 <sup>a,c</sup>

<sup>a</sup>Significant at .01.

<sup>b</sup>Significantly less than 1.0 at a 95% confidence level.

<sup>c</sup>Significantly greater than 1.0 at a 95% confidence level.

Table 6.2 Correlations, Standard Deviations, Covariances and Betas for Four Stocks November 2000 to October 2005

The correlation between Lockheed Martin's monthly rate of return and the Market monthly rate of return does not differ from zero, confirming that Lockheed Martin's returns move independently of the Market. Correlations between each of the other three stocks' returns and the Market are significantly greater than zero, indicating that they do move with the Market. IBM, with its large correlation of .68, magnifies Market movement.

General Electric's and Apple's returns are both moderately correlated with the Market index returns ( $r_{GeneralElectric,Market} \cong r_{Apple,Market} \cong .4$ ). However, Apple returns are considerably more volatile ( $s_{Apple} = .14 < s_{GeneralElectric} = .06$ ). GE returns dampen market returns more, as we see in a comparison of the covariances between the two stocks' returns and Market returns ( $cov_{GE,Market} = .0012 < cov_{Apple,Market} = .0027$ ). Because Apple rates of return are more volatile than General Electric, Apple will also have a larger beta than General Electric.

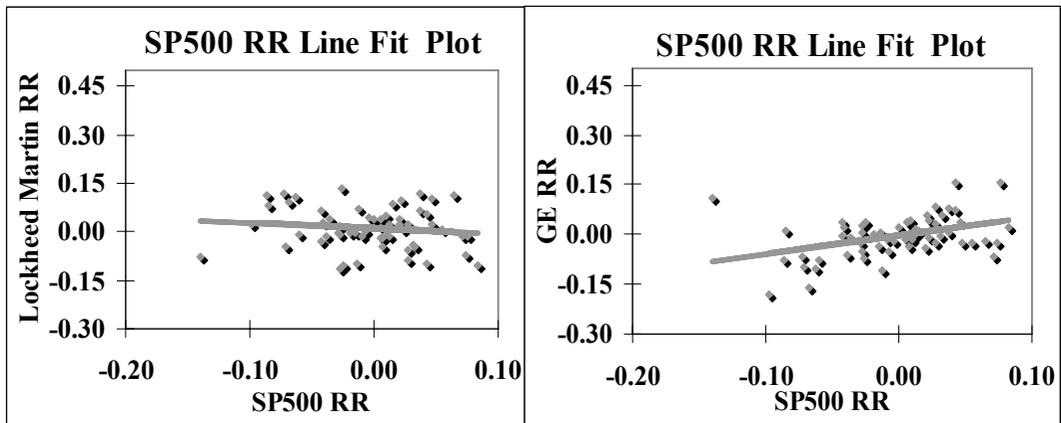
Betas  $b_{stocki}$  are shown in the last column of Table 6.3. A percent increase in the Market produces

- a zero expected change in Lockheed Martin's price,
- less than one percent expected increase in General Electric's price,
- a one percent expected increase in Apple's price, and
- more than one percent expected increase in IBM's price.

Beta estimates are shown in Table 6.3 and Figure 6.4.

<i>Regression Statistics: Lockheed Martin</i>						
Multiple R	0.130					
R Square	0.017					
Standard Error	0.064					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	0.004	0.004	0.97	0.3280	
Residual	57	0.234	0.004			
Total	58	0.238				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.011	0.008	1.3	0.1839	-0.005	0.028
SP500 RR	<b>-0.177</b>	0.179	-1.0	0.3280	-0.536	0.182

Table 6.3 Estimates of betas for four diverse stocks



$$\hat{R}R_{LMt} = .011 - .177SP500_t$$

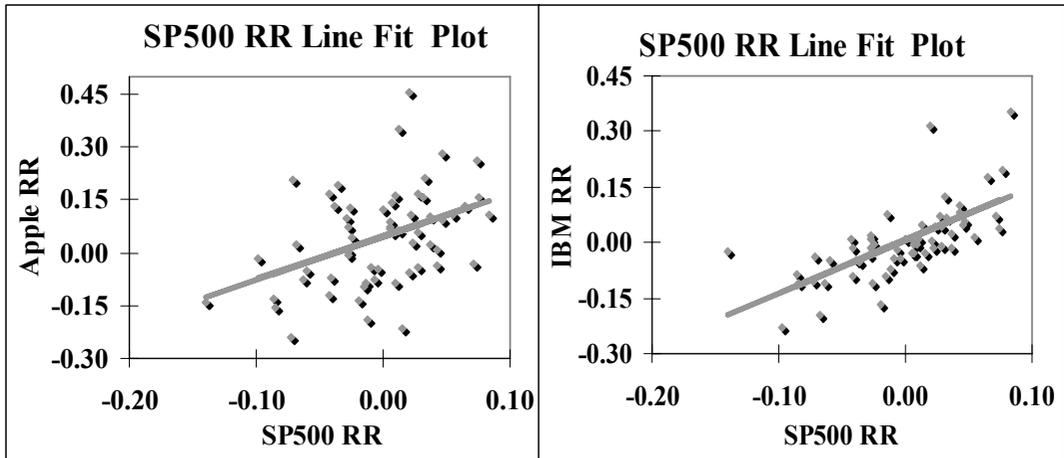
(.008)(.179)

*RSquare: .02*

$$\hat{R}R_{GEt} = -.003 + .55^a SP500_t$$

(.008)(.16)

*RSquare: .17<sup>a</sup>*  
*<sup>a</sup>Significant at .01*



$$\hat{RR}_{Apple_t} = .046^a + 1.22^a SP500_t$$

(.017) (.35)

RSquare: .17<sup>a</sup>

<sup>a</sup>Significant at .01

$$\hat{RR}_{IBM_t} = .007 + 1.45^a SP500_t$$

(.010) (.21)

RSquare: .46<sup>a</sup>

<sup>a</sup>Significant at .01

Figure 6.4 Response of four diverse stocks to The Market

A potential investor would conclude:

*“Lockheed Martin, with an estimated beta of zero, is the least risky stock of the four. LM returns are relatively invulnerable to Market swings. A change in the Market return is not associated with change in LM’s price.*

*General Electric, with an estimated beta less than one ( $b_{GE}=.55$ ), is a low risk investment. GE returns dampen Market swings. With a percent increase in the Market, we expect to see an average increase of .55% in GE’s price.*

*Apple stock, with an estimated beta of one ( $b_{Apple}=1.22$ ) is riskier than LM or GE, and mirrors Market movement. With a percent increase in the Market, we expect to see an average increase of about one percent, 1.22%, in Apple’s price.*

*IBM is the riskiest investment of the four, with an estimated beta greater than one ( $b_{IBM}=1.45$ ). IBM returns exaggerate Market swings. With a percent increase in the Market, we expect to see an average increase of 1.45% in IBM’s price.”*

### 6.4 A Portfolio’s Expected Return, Risk and Beta Are Weighted Averages of Individual Stocks

An investor is really interested in the expected return and risk of her portfolio of stocks. These are weighted averages of the expected returns and betas of the individual stocks in a portfolio:

$$E(RR_p) = \sum_i w_i E(RR_i)$$

$$b_p = \sum_i w_i b_i$$

Where  $E(RR_p)$  is the expected portfolio rate of return,  
 $w_i$  is the percent of investment in the  $i$ 'th stock,  
 $E(RR_i)$  is the expected rate of return of the  $i$ 'th stock,  
 $b_p$  is the portfolio beta estimate,  
 $b_i$  is the beta estimate of the  $i$ th stock,

*Example 6.3 Four Alternate Portfolios.* An Investment Manager has been asked to suggest a portfolio of three stocks from four being considered by a client: Lockheed Martin, General Electric, Apple and IBM. The prospective investor wanted to include computer stock in his portfolio and had heard that IBM was a desirable “Blue Chip.” He suspected that holding both Apple and IBM stocks might be risky, were the computer industry to falter.

To confidently advise her client, the Investment Manager compared four portfolios of three equally weighted stocks from the four requested options. Individual stock weights in each portfolio equal one third. Table 6.4 contains the expected portfolio rates of return and betas for the four possible combinations:

Portfolio	Expected Portfolio Return		Portfolio Beta Estimate	
	$\sum E(RR_i)/3$	$E(RR_p)$	$\sum b_i / 3$	$b_p$
<b>LM+GE+Apple</b>	$(.012 -.004 +.042)/3$	0.017	$(-.18+ .55+1.22)/3$	<b>0.53</b>
LM+GE+IBM	$(.012 -.004 +.002)/3$	0.003	$(-.18+ .55+1.45)/3$	0.61
<b>LM+Apple+IBM</b>	$(.012+.042+.002)/3$	<b>0.019</b>	$(-.18+ 1.22+1.45)/3$	0.83
GE+Apple+IBM	$(-.004+.042+.002)/3$	0.013	$(.55+1.22+1.45)/3$	1.07

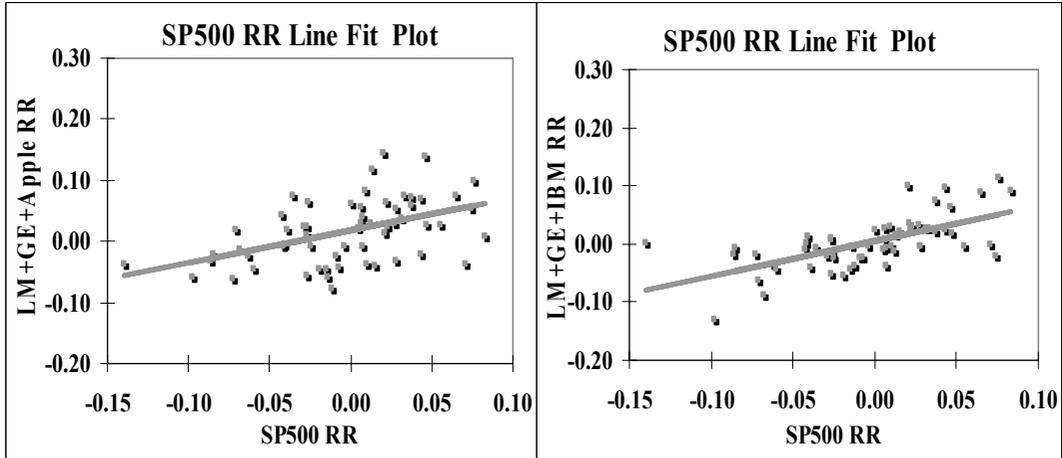
Table 6.4 Expected portfolio returns and beta estimates

Alternatively, she could find expected portfolio returns and betas with software, and this would be the practical way to compare more than a few portfolios. Table 6.5 and Figure 6.5 show expected (mean) rates of return and regression beta estimates for the four portfolios from Excel:

	<i>LM+GE+Apple</i>	<i>LM+GE+IBM</i>	<i>LM+Apple+IBM</i>	<i>GE+Apple+IBM</i>
Mean	0.017	0.003	<b>0.019</b>	0.013

Table 6.5 Expected rates of return of four alternate portfolios from descriptive statistics

<i>Regression Statistic: Lockheed Martin+General Electric+Apple</i>						
Multiple R	0.480					
R Square	0.230					
Standard Error	0.046					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	0.036	0.036	17.1	0.0001	
Residual	57	0.121	0.002			
Total	58	0.157				
	<i>Standard</i>					
	<i>Coefficients</i>	<i>Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.018	0.006	3.0	0.0036	0.006	0.030
<i>SP500 RR</i>	<b>0.533</b>	0.129	4.1	0.0001	0.274	0.791

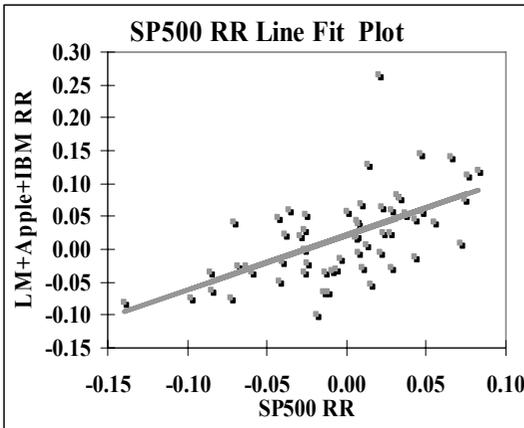


$$\hat{R}R_{LM+GE+Apple_t} = .018 + .533^{a,b} SP500RR_t$$

*RSquare: .23<sup>a</sup>*  
<sup>a</sup>Significant at .01  
<sup>b</sup>Significantly less than 1.

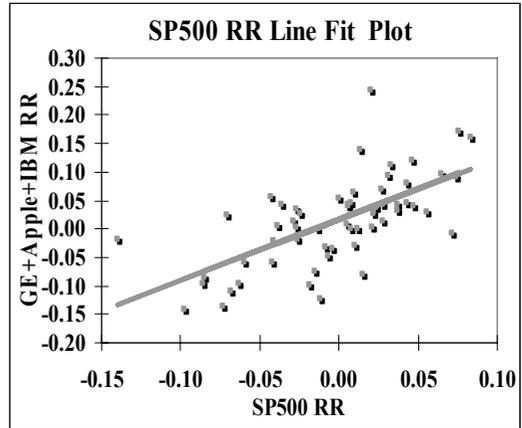
$$\hat{R}R_{LM+GE+IBM_t} = .005 + .607^{a,b} SP500RR_t$$

*RSquare: .41<sup>a</sup>*  
<sup>a</sup>Significant at .01  
<sup>b</sup>Significantly less than 1



$$\hat{R}R_{LM+Apple+IBM_t} = .0195 + 1.093^a SP500RR_t$$

*RSquare: .426<sup>a</sup>*  
<sup>a</sup>Significant at .01.



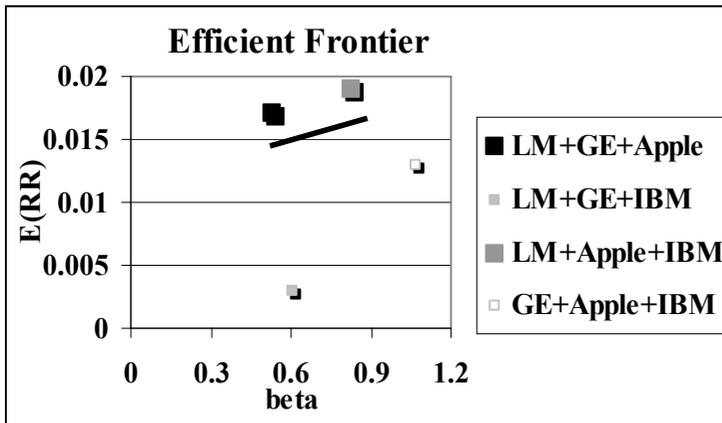
$$\hat{R}R_{GE+Apple+IBM_t} = .017 + 1.08^a SP500RR_t$$

*RSquare: .42<sup>a</sup>*  
<sup>a</sup>Significant at .01

Figure 6.5 Beta estimates of four alternate portfolios

## 6.5 Better Portfolios Define The Efficient Frontier

In the comparison of alternative portfolios, the Investment Manager wanted to identify alternatives which promised greater expected return without greater risk—or, alternatively, those which reduced risk without reducing return. Better portfolios, which promise the highest return for a given level of risk, define the *Efficient Frontier*. To see the Efficient Frontier, she made a scatterplot of portfolio expected rate of return by portfolio risk. Those relatively efficient portfolios lie in the upper left:



Comparing portfolios in Figure 6.6, the Investment Manager found that the portfolio which contains Lockheed Martin, Apple and GE (see the large, black marker) offers both a higher expected rate of return and lower risk than the two portfolios which lack the Lockheed Martin+Apple combination.

Figure 6.6 Relatively Efficient Portfolios Offer Greater Expected Return and Lower Risk

$$E(RR_{LM+GE+APPLE}) = .017 > E(RR_{GE+IBM+APPLE}) = .013 > E(RR_{LM+GE+IBM}) = .003$$

$$b_{LM+GE+APPLE} = .533 < b_{LM+GE+IBM} = .607 < b_{GE+IBM+APPLE} = 1.076$$

Adding IBM instead of GE to the Lockheed Martin+Apple combination (see the large, grey marker) increases both the expected return and the risk:

$$E(RR_{LM+APPLE+IBM}) = .019 > E(RR_{LM+GE+APPLE}) = .017$$

$$b_{LM+APPLE+IBM} = .83 > b_{LM+GE+APPLE} = .53$$

These two portfolios with the Lockheed Martin+Apple combination dominate the two portfolios without the combination. However, the choice between the two, with GE (black) or with IBM (grey), will depend upon the prospective investor's risk preference.

The Investment Manager presented results of her analysis with recommendations in this memo to her client:



## 6.6 Portfolio Risk Depends On the Covariances between Individual Stocks' Rates of Return and The Market Rate Of Return

Both the expected rate of return of a portfolio and its risk, measured by its beta, depend on the expected rates of return and betas of the individual stocks in the portfolio. Individual stock betas are direct functions of

- the correlation between a stock's rate of return and the Market index rate of return, and
- the standard deviation of a stock's rate of return

We estimate beta for a stock or a portfolio by regressing the stock or portfolio monthly rates of return against monthly Market rates of return. The resulting simple linear regression slopes are estimates of the stock or portfolio beta.

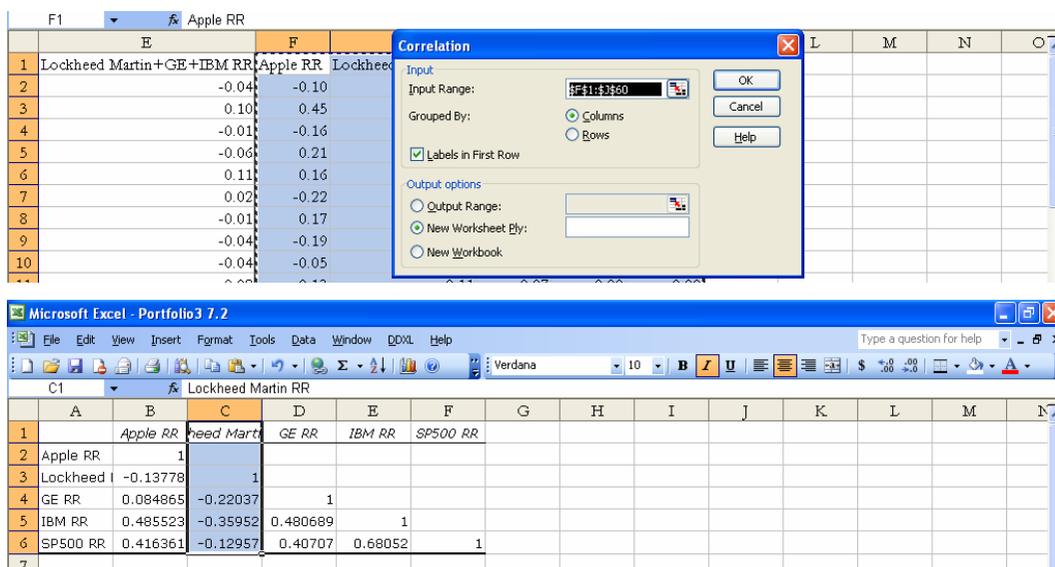
### Excel 6.1 Estimate portfolio expected rate of return and risk

*Four Portfolios with Lockheed Martin, GE, IBM and Apple.* Monthly rates of return for each of the four stocks and the S&P500 index of the Market, adjusted for inflation are in **Excel 6.1 Portfolio3.xls**.

**Correlations between stocks and The Market.** Correlations between rates of return of pairs of stocks and the Market sometimes suggest combinations which might reduce risk through diversification.

To see the pairwise correlations, **Alt AY2, Correlation, OK.**

For **Input Range**, use shortcuts to select the rates of return of the four stocks and the S&P500 in columns **F1** through **J60**: Select **F1**, **Cntl+Shift** right and down through **J60**. Choose **Labels**, **OK**:



Lockheed Martin adds diversification and reduces risk in portfolios with the other three stocks.

**Monthly portfolio returns formula.** We will make a new column for each portfolio's monthly rate of return, which will be the average of rates of return of each of the three stocks in each portfolio.

In **B1**, type in a label for the first portfolio with equally weighted investments in Apple, Lockheed Martin, and GE, *Apple+Lockheed Martin+GE RR*:

In **B2**, enter a formula for the average of the three stocks **=AVERAGE(F2,G2,H2)** [Enter].

Select the new cell and double click the lower right corner to fill in the monthly rates of return for this portfolio:

B2		=AVERAGE(F2,G2,H2)									
1	Month	Apple+Lockheed Martin+GE RR	Apple+Lockheed Martin+IBM RR	Apple+GE+IBM RR	Lockheed Martin+GE+IBM RR	Apple RR					
2	12/29/00	-0.04		-0.06	-0.07	-0.04					
3	1/26/01	0.15		0.26	0.24	0.10					
4	2/23/01	-0.02		-0.06	-0.08	-0.01					
5	3/30/01	0.02		0.04	0.02	-0.06					
6	4/27/01	0.10		0.11	0.17	0.11					
7	5/25/01	-0.04		-0.05	-0.08	0.02					
8	6/29/01	0.04		0.05	0.06	-0.01					

Monthly rates of return for the other three-stock portfolios have been calculated similarly in C through E.

**Expected monthly rates of return.** We will find the expected monthly return for the four portfolios in **Portfolio3.6.2.xls**.

Enter the label  $E(RR)$  in **A62**, then use the Excel function **AVERAGE(array)** to find the expected portfolio returns.

In **B62**, enter **=AVERAGE(B2:B60)** [CR].

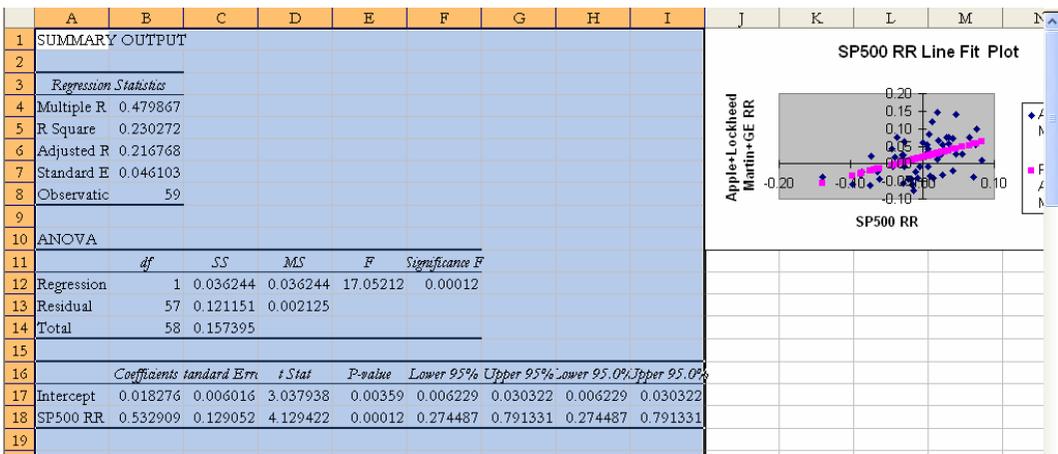
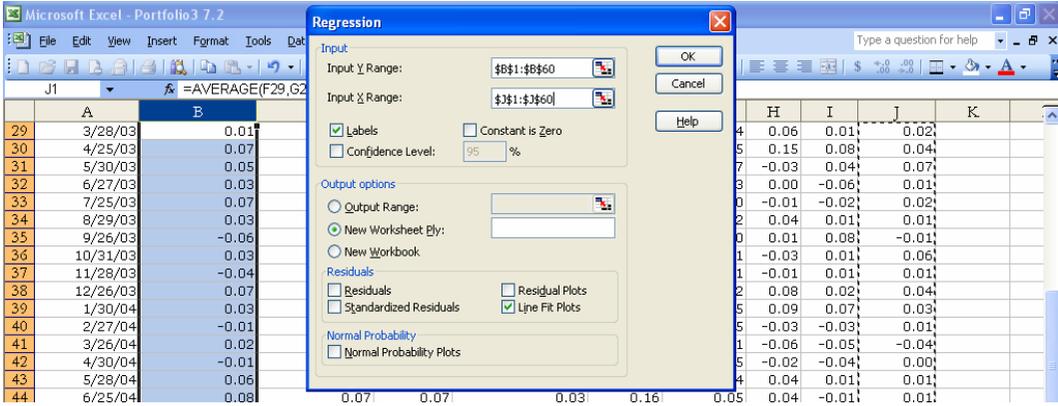
Use shortcuts to fill in the remaining expected expected portfolio returns:

Select **B62**, **Shift+→**through **E62**, **Ctrl+R**.

B62		=AVERAGE(B2:B60)									
1	Month	Apple+Lockheed Martin+GE RR	Apple+Lockheed Martin+IBM RR	Apple+GE+IBM RR	Lockheed Martin+GE+IBM RR	Apple RR	Lockheed Martin RR	GE RR	IBM RR	SP500 RR	
53	3/25/05	-0.005	-0.018	-0.020	0.014	-0.07	0.03	0.02	-0.01	-0.04	
54	4/29/05	-0.044	-0.100	-0.098	-0.054	-0.14	0.00	0.00	-0.16	-0.02	
55	5/27/05	0.059	0.052	0.033	0.021	0.10	0.07	0.01	-0.01	0.04	
56	6/24/05	-0.041	-0.031	-0.047	-0.023	-0.07	0.00	-0.05	-0.02	-0.01	
57	7/29/05	0.039	0.082	0.093	0.028	0.16	-0.04	0.00	0.13	0.03	
58	8/26/05	0.023	0.021	0.013	-0.021	0.10	0.00	-0.03	-0.03	-0.03	
59	9/30/05	0.042	0.040	0.047	-0.007	0.14	-0.02	0.00	-0.01	0.01	
60	10/28/05	0.024	0.029	0.034	0.007	0.07	-0.01	0.01	0.02	-0.03	
61											
62	e(RR)	0.017	0.019	0.013	0.003						

**Estimated betas from simple regression.** To find the Market-related risk,  $\beta$ , we will request simple regression slope of each portfolio rate of return with  $SP500 RR$ .

For the first portfolio, *Apple+Lockheed Martin+GE*, run regression with *Apple+Lockheed Martin+GE RR* in column **B2:B60** in the **Input Y Range**, and *SP500 RR* in **J2:J60** in the **Input X Range**:



From the *Lower 95%* ( $=b_{SP500RR} - t_{57}s_{b_{SP500RR}}$ ) and *Upper 95%* ( $=b_{SP500RR} + t_{57}s_{b_{SP500RR}}$ )

confidence interval bounds for coefficient, .27 and .79, we see that one lies outside this interval. The portfolio beta is less than one, meaning that the *Apple+Lockheed Martin+GE* combination dampens Market fluctuations and is a conservative portfolio. We expect that in months when the Market gains one percentage point, the portfolio will gain about half a percentage point (0.27% to 0.79%).

### Excel 6.2 Plot return by risk to identify dominant portfolios and the Efficient Frontier

To compare the expected rates of return and estimated risk of the four portfolios, we will plot the portfolio rates of return against their betas to see the Efficient Frontier.

Enter eight variables in a new worksheet:

*Apple+LM+GE beta* and *Apple+LM+GE E(RR)* in columns **A** and **B**,  
*Apple+LM+IBM beta* and *Apple+LM+IBM E(RR)* in columns **C** and **D**,  
*Apple+GE+IBM beta* and *Apple+GE+IBM E(RR)* in columns **E** and **F**,  
*LM+GE+IBM beta* and *LM+GE+IBM E(RR)* in columns **G** and **H**.

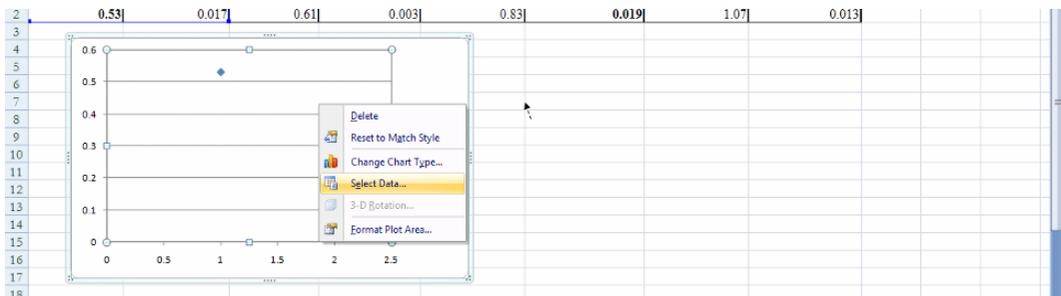
Enter the beta estimates in **B18** from regression output worksheets in row **2**, columns **A**, **C**, **E**, and **G** of the new worksheet.

Enter expected rates of return *E(RR)* from **B62:E62** of the original worksheet into row **2** of columns **B**, **D**, **F**, and **H**.

	A	B	C	D	E	F	G	H	I	J	K
1	LM+GE+Apple beta	E(LM+GE+Apple RR)	LM+GE+IBM beta	E(LM+GE+IBM RR)	LM+Apple+IBM beta	E(LM+Apple+IBM RR)	GE+Apple+IBM beta	E(GE+Apple+IBM RR)			
2		0.53	0.017	0.61	0.003	0.83	0.019	1.07	0.013		
3											

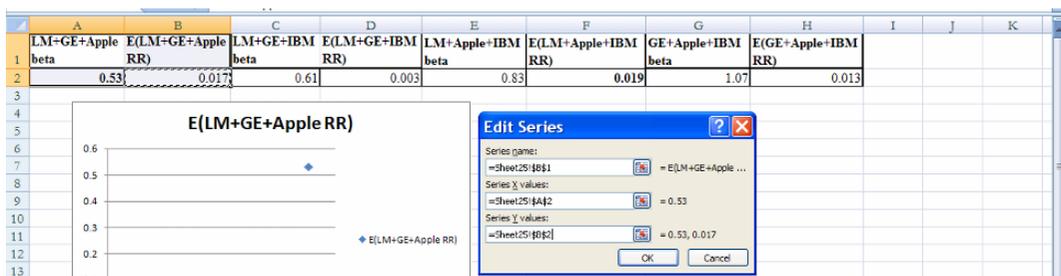
Select columns **A** and **B**, and insert a scatterplot, choosing the chart type with markers only.

Right click inside the scatterplot and choose **Select Data**:

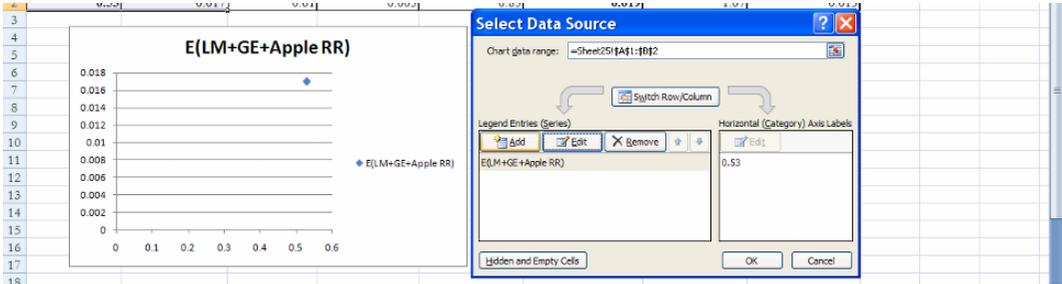


Select and **Edit** the series.

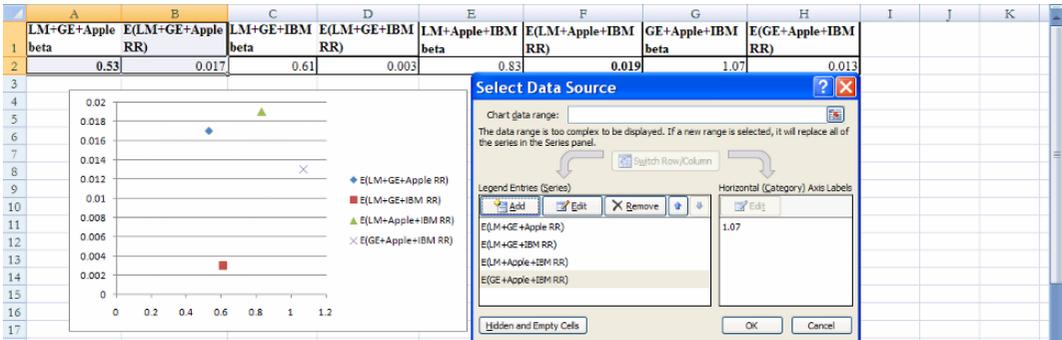
For **Series name** select **B1**, for **Series X values** select the portfolio *beta* in **A2**, and for **Series Y values** select the portfolio *E(RR)* in **B2**, **Ok**.



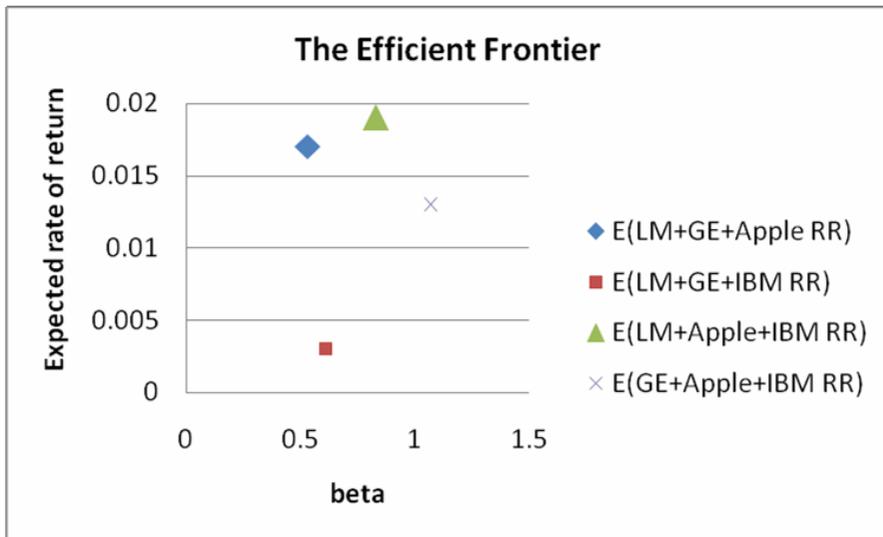
Add the second portfolio point, choosing **D1** for **Series name**, the *beta* in **C2** for **X values**, and the *E(RR)* in **D2** for **Y values**.



Add the third and fourth portfolio points:



Add a title and axes titles:



Portfolios that are higher offer higher expected returns. The *LM+Apple+IBM* portfolio, with large triangular marker, has the highest expected rate of return. Those that are more left are less risky than those to the right. The *LM+GE+Apple* portfolio, with large diamond marker, has the lowest risk. These two dominate the remaining two, offering either higher rate of return or lower risk, or both. An investor would prefer one of the two dominant portfolios, and choice between the two would depend on her risk taking.

### *Assignment 6-1 Individual Stocks' Beta Estimates*

Use logic to choose two stocks to analyze from **Assignment 6-1 Stock RR.xls**.

Choose a stock which you would expect to have a **beta less than one**, and a stock which you expect to have a **beta more than one**.

Be prepared to explain the logic of your choices.

The **Assignment 6-1 Stock.xls** dataset contains five years of monthly *rates of return* from November 2000 to October 2005, for seventeen individual stocks, as well as monthly rates of return for a Market index, the S&P500.

Stock *rates of return* included in the dataset are:

<i>Northrop Grumman</i>	<i>Procter &amp; Gamble</i>	<i>Microsoft</i>
<i>NUCOR Steel</i>	<i>WalMart</i>	<i>Goldman Sachs</i>
<i>US Steel</i>	<i>Disney</i>	<i>Merrill Lynch</i>
<i>Boeing</i>	<i>Starbucks</i>	<i>Nanogen</i>
<i>Merck</i>	<i>Whole Foods</i>	<i>Nanophase</i>
<i>Johnson &amp; Johnson</i>	<i>Yahoo</i>	

- Plot rates of return for both stocks and the S&P500 return across the 60 months in a scatterplot overlay.  
Do the stocks track the Market?  
Do they dampen or exaggerate Market swings?
- Conduct two simple linear regressions to estimate the betas of the two stocks which you chose.  
(The two dependent variables will be the monthly *rates of return* of the two stocks and the independent variable will be monthly *S&P 500 rates of return*, S&P500RR.)  
Record the beta estimates which you find to share with the class.

### *Assignment 6-2 Expected Returns and Beta Estimates of Alternate Portfolios*

A potential investor has asked you to recommend two stocks which together would produce a desirable portfolio. He expects to invest **half** in each stock.

Choose three stocks from the set of seventeen in **Assignment 6-2 Stock RR.xls** to potentially combine.

Compare the *expected return* and *risk (beta)* of the **three** portfolios from **all possible pairs** and make a recommendation to the investor.

To assess the three alternative portfolios, you will need to

- make three new portfolio variables equal to averages of each of the stock pairs' *rates of return*, then find the average sample portfolio return, which is the *expected portfolio return*, and
- run simple regressions of the portfolio monthly *rates of return* against the *Market rate of return* to find portfolio *betas*

### *Assignment 6-3 Portfolio Comparison*

An investor would like to construct a portfolio with three stocks, each weighted equally. She is considering General Motors, Kellogg, Toyota, and Yahoo.

**Assignment 6-3 Portfolio4.xls** contains five years of monthly data on:

*SP500*, the rate of return of the S&P500 Market index, adjusted for inflation, *GM*, the rate of return of GM stock,

*KELLOGG*, the rate of return of Kellogg stock,

*TOYOTA*, the rate of return of Toyota stock,

*YAHOO*, the rate of return of Yahoo stock,

$G+K+T$ , the rate of return of a portfolio of GM, Kellogg and Toyota stocks,

$G+K+Y$ , the rate of return of a portfolio of GM, Kellogg and Yahoo stocks,

$G+T+Y$ , the rate of return of a portfolio of GM, Toyota, and Yahoo stocks,

$K+T+Y$ , the rate of return of a portfolio of Kellogg, Toyota, and Yahoo stocks

- Find each of the four individual beta estimates and assign each individual stock to the group it belongs with:
  - **lower risk** and **uncorrelated** with the Market,
  - **lower risk** and **dampens** Market movement,
  - **reflects** Market movement, and
  - **higher risk** and **exaggerates** Market movement.

Explain, using logic, why each of the four stocks belong in their group, above,

- What are the *expected rates of return* of each of the four portfolios?
- What percent increase in each portfolio value is expected for a one percent increase in the Market's value?
- Construct a chart of the Efficient Frontier and offer your investment recommendation to the potential investor, based on comparison of *expected rates of return* and estimated estimated *risk*.

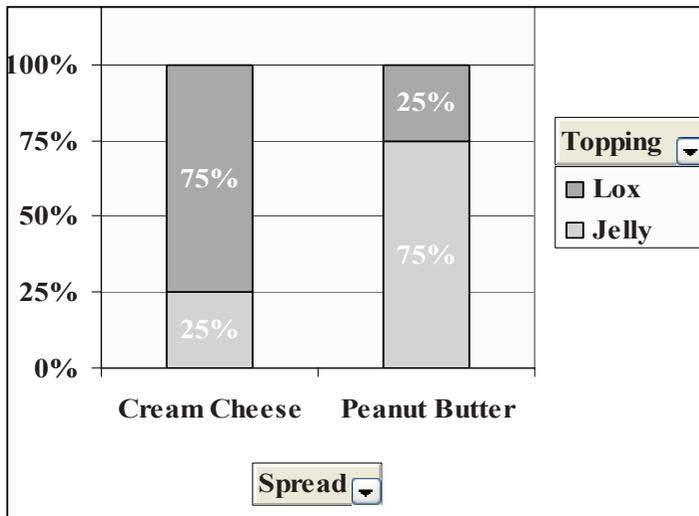
# 7

## Association between Two Categorical Variables: Contingency Analysis with Chi Square

Categorical variables, including nominal and ordinal variables, are described by tabulating their frequencies or probability. If two variables are associated, the probability of one will depend on the probability of the other. Chi square tests the hypothesized association between two categorical variables and contingency analysis allows us to quantify their association.

### 7.1 When Conditional Probabilities Differ From Joint Probabilities, There Is Evidence of Association

Contingency analysis begins with the crosstabulation of frequencies of two categorical variables. Figure 7.1 shows a crosstabulation of sandwich spreads and topping combinations chosen by forty students:



<i>Counts</i>	<i>JELLY</i>	<i>LOX</i>	<i>total</i>	<i>Percent of Row</i>	<i>JELLY</i>	<i>LOX</i>	<i>total</i>
	<i>Cream Cheese</i>	5	15		20	<i>Cream Cheese</i>	25
<i>Peanut Butter</i>	15	5	20	<i>Peanut Butter</i>	75	25	100
<i>total</i>	20	20	40	<i>total</i>	50	50	100

Figure 7.1 Crosstabulation: Sandwich topping depends on spread

To gauge association, the conditional probability of each category of the first variable, given each category of the second variable, is compared to the unconditional, row probabilities of the first variable. If these differ, we have evidence of association.

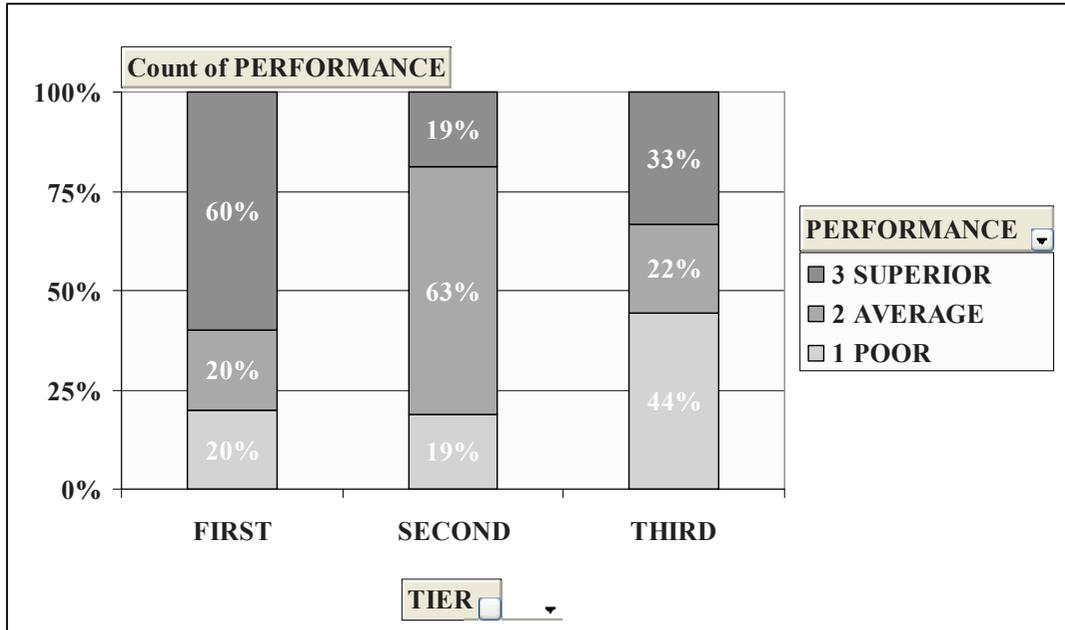
In this sandwich example, jelly topping was chosen by half the students, making its unconditional probability .5. If a student chose cream cheese spread, the conditional probability of jelly topping was lower (.25). If a student chose peanut butter spread, jelly was the more likely topping choice (.75).

*Example 7.1 Recruiting Stars.* Human Resource managers are hoping to improve the odds of hiring outstanding performers and to reduce the odds of hiring poor performers by targeting recruiting efforts. Management believes that recruiting at the Top Twenty Undergraduate Programs, identified each year by U.S. News & World Report, might improve the odds of hiring a star. Removing the lowest ranked programs from the recruiting list might reduce the number of lackluster performers. Management's hypotheses are:

$H_0$ : Job performance is not associated with undergraduate program quality.

$H_1$ : Job performance is associated with undergraduate program quality.

To test these hypotheses, department supervisors throughout the firm sorted a sample of forty recent hires into three categories based on job performance: poor, average, and outstanding. The sample employees were also categorized by the undergraduate program they had completed: Top, Middle, and Bottom. Undergraduate programs ranked in the Top Twenty by U.S. News & World Report were classified as "Top," those ranked 21<sup>st</sup> through 99<sup>th</sup> were classified as "Second Tier", and those ranked 100<sup>th</sup> through 200<sup>th</sup> were classified as "Third Tier." These cross-tabulations are shown in the PivotChart and PivotTable in Figure 7.2.



Count	Performance			
Program	Poor	Average	Outstanding	Total
First	3	3	9	15
Second	2	10	3	15
Third	5	2	3	10
Total	10	15	15	40
% of Row	Performance			
Program	Poor	Average	Outstanding	Total
First	20%	20%	60%	100%
Second	13%	67%	20%	100%
Third	50%	20%	30%	100%
Total	25%	38%	38%	100%

$$\chi^2_4 \quad 12.3p \text{ value} \quad .02$$

Figure 7.2 Job Performance Depends on Program Quality

The crosstabs indicate that a quarter of the firm’s new employees are *Poor* performers, about forty percent are *Average* performers, and about forty percent are *Outstanding* performers. From the PivotChart we see that more than a quarter of employees from *Third* Tier programs are *Poor* performers, and more than forty percent of employees from *First* Tier programs are *Outstanding* performers. Were program rank and performance

not associated, a quarter of the recruits from each type of program would be *Poor* performers. We would, for example, expect a quarter of ten employees recruited from *Third Tier* programs to be *Poor* performers, or 2.5 ( $=.25(10)$ ). Instead, there are actually five (*Third, Poor*) employees. There is a greater chance, 50%, of *Poor* performance, given *Third Tier*, rather than *Second* or *First* Tiers. Ignoring program quality, the probability of poor performance is .25; acknowledging program quality, this probability of poor performance varies from .13 (*Second*) to .50 (*Third*). These differences in row percentages suggest an association between program rank and performance.

## 7.2 Chi Square Tests Association between Two Categorical Variables

The chi square ( $\chi^2$ ) statistic tests the significance of the association between performance and program quality, by comparing expected cell counts with actual cell counts, squaring the differences, and weighting each cell by the inverse of expected cell frequency.

$$\chi^2_{(R-1),(C-1)} = \sum_{ij}^{RC} (e_{ij} - n_{ij})^2 / e_{ij},$$

Where  $R$  is the number of row categories,

$C$  is the number of column categories,

$n$  is the number in the  $i$ 'th row and  $j$ 'th column,

$e$  is the number expected in the  $i$ 'th row and  $j$ 'th column.

$\chi^2$  gives more weight to the least likely cells. In the **Recruiting Stars** example, **Figure 7.2**, Pearson Chi square,  $\chi^2$ , is 12.3, which can be verified using the formula:

$$\begin{aligned} \chi^2 &= (3.75 - 3)^2/3.75 + (5.625 - 3)^2/5.625 + (5.625 - 9)^2/5.625 \\ &+ (3.75 - 2)^2/3.75 + (5.625 - 10)^2/5.625 + (5.625 - 3)^2/5.625 \\ &+ (2.5 - 5)^2/2.5 + (3.75 - 2)^2/3.75 + (3.75 - 3)^2/3.75 \\ &= .15 \qquad + 1.23 \qquad + 2.03 \\ &+ .817 \qquad + 3.40 \qquad + 1.23 \\ &+ 2.5 \qquad + .82 \qquad + .15 = 12.3 \end{aligned}$$

From a table of  $\chi^2$  distributions, we find that for a crosstabulation of this size, with three rows and three columns, ( $df=(Rows-1) \times (Columns - 1)=2 \times 2 = 4$ ),  $\chi^2_4 = 12.3$  indicates that the p-value is .02. We reject the null hypothesis and accept the alternate hypothesis of association.

Those cells which contribute more to chi square indicate the nature of association. In this example, we see in Table 7.1 that these are the (*First, Outstanding*), (*Second, Average*), and (*Third, Poor*) cells:

$$\begin{aligned} \chi^2 &= .15 + 1.23 + \mathbf{2.03} \\ &+ .82 + \mathbf{3.40} + 1.23 \\ &+ \mathbf{2.5} + .82 + .15 = 12.3 \end{aligned}$$

	<i>Poor</i>	<i>Average</i>	<i>Outstanding</i>
<i>First</i>	.15	1.23	<b>2.03</b>
<i>Second</i>	.82	<b>3.40</b>	1.23
<i>Third</i>	<b>2.5</b>	.82	.15

Table 7.1 Contribution to chi square by cell

*Outstanding* performance is more likely if a new employee came from a *First* Tier program, *Average* performance is more likely if a new employee came from a *Second* Tier program, and *Poor* performance is more likely if a new employee came from a *Third* Tier program. Job performance is associated with program quality.

### 7.3 Chi Square Is Unreliable If Cell Counts Are Sparse

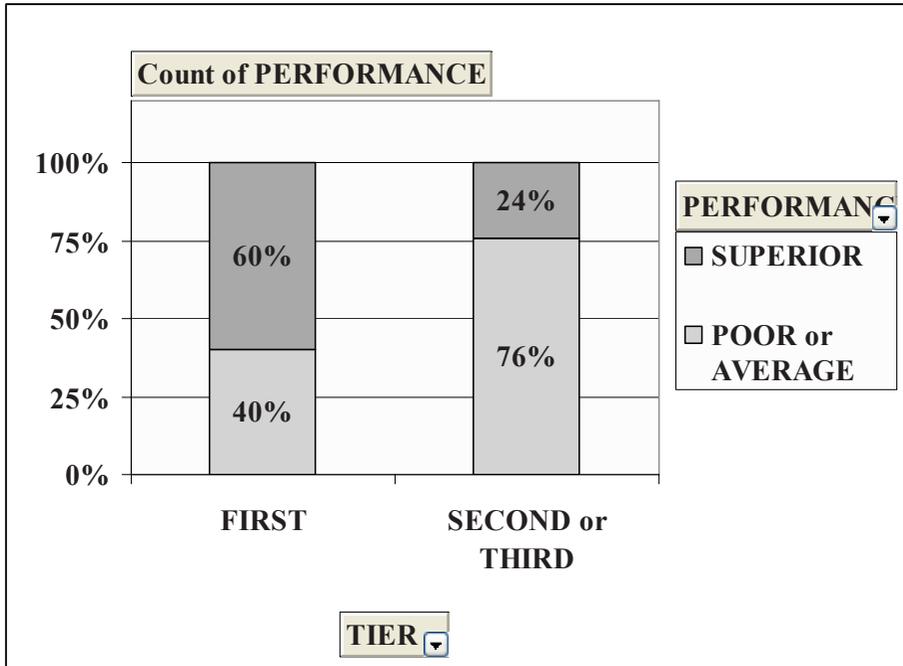
There are two possible reasons why the chi square statistic is large and apparently significant. The first reason is the likely actual association between program quality and performance. The second reason is that there are few (less than five) expected employees in five of the nine cells, shown in Table 7.2.

	<i>Poor</i>	<i>Average</i>	<i>Outstanding</i>
<i>First</i>	<b>3.75</b>	5.63	5.63
<i>Second</i>	<b>3.75</b>	5.63	5.63
<i>Third</i>	<b>2.5</b>	<b>3.75</b>	<b>3.75</b>

Table 7.2 Expected counts by cell

Since the chi square components include expected cell counts in the denominator, *sparse* (with expected counts less than five) cells inflate chi square. When sparse cells exist, we combine categories.

In the **Recruiting Stars** example, management was most interested in increasing the chances of hiring *Outstanding* performers. Since some believed that *Outstanding* performers were recruited from *First* Tier programs, these categories were preserved. *Second* and *Third* Tier program ranks were combined. *Poor* and *Average* performance categories were combined. We are left with a 2 x 2 contingency analysis, Figure 7.3.



Count	Performance			% Row	Performance		
	Poor/Average	Out-standing	Total		Poor/Average	Out-standing	Total
Top	6	9	15	Top	40%	60%	100%
Bottom/Middle	19	6	25	Bottom/Middle	76%	24%	100%
Total	25	15	40	Total	63%	38%	100%

Chi Square	5.18
df	1
p value	.0228

Figure 7.3 PivotChart of performance by program quality with fewer categories

With fewer categories, all expected cell counts are now greater than five, providing a reliable  $\chi^2_1 = 5.2$ , which remains significant at a 98% level of confidence ( $p\text{ value}=.02$ ). The PivotChart continues to suggest that the incidence of *Outstanding* performance is greater among employees recruited from *First* Tier programs. The impact of program Tier on *Poor* performance is unknown, since *Poor* and *Average* categories were combined. Also unknown is the difference between employees from *Second* and *Third* Tier programs, since these categories were likewise combined.

Recruiters would conclude:

*“We conclude that job performance of newly hired employees is associated with undergraduate program quality rank. Twenty-four percent of our new employees recruited from Second or Third Tier undergraduate programs have been identified as Outstanding performers. Within the group recruited from First Tier undergraduate programs, more than twice this percentage, 60%, are Outstanding performers, a significant difference. Results suggest that in order to achieve a larger percent of Outstanding performers, recruiting should be focused on First Tier programs.”*

## 7.4 Simpson's Paradox Can Mislead

Using contingency analysis to study the association between two variables can be potentially misleading, since we are ignoring all other related variables. If a third variable is related to the two that we're analyzing, contingency analysis may indicate that they are associated, when they may not actually be. Two variables may appear to be associated because they are both related to a third, ignored variable.

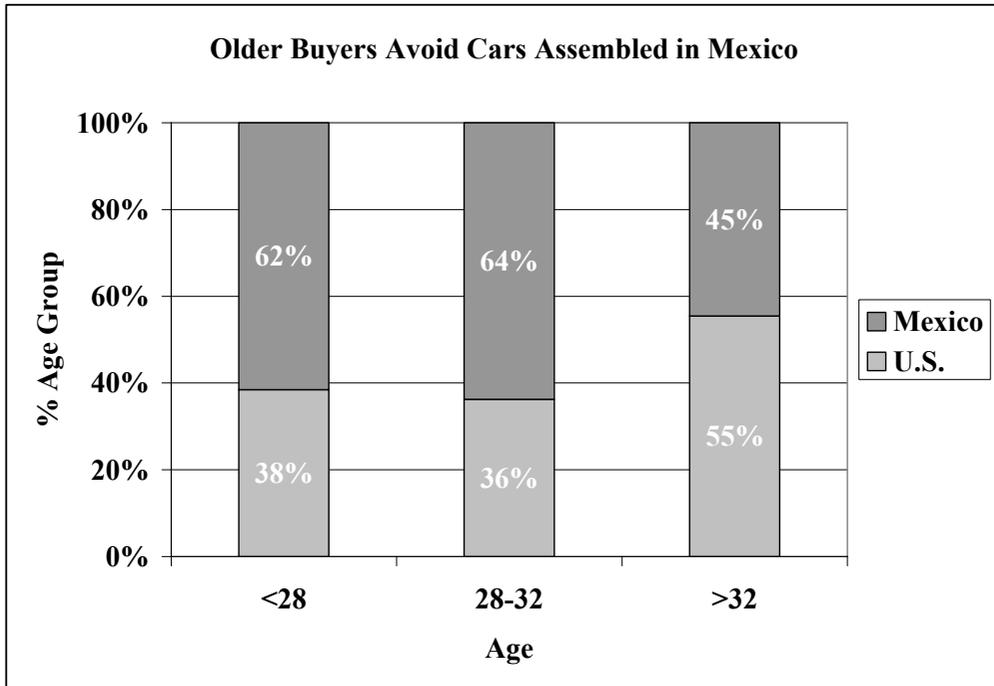
*Example 7.2 American Cars.* The CEO of American Car Company was concerned that the oldest segments of car buyers were avoiding cars that his firm assembles in Mexico. Production and labor costs are much cheaper in Mexico, and his long term plan was to shift production of all models to Mexico. If older, more educated and more experienced buyers avoid cars produced in Mexico, American Car stood to lose a major market segment unless production remained in The States.

The CEO's hypotheses were:

$H_0$ : Choice between cars assembled in the U.S. and cars assembled in Mexico is not associated with age category.

$H_1$ : Choice between cars assembled in the U.S. and cars assembled in Mexico is associated with age category.

He asked Travis Henderson, Director of Quantitative Analysis, to analyze the association between age category and choice of U.S.-made versus Mexican-made cars. The research staff drew a random sample of 263 recent car buyers, identified by age category. After preliminary analysis, age categories were combined to insure that all expected cell counts in an [Age Category x Origin Choice] crosstabulation were each at least five. Contingency analysis is shown in the PivotChart and Pivot Tables in Figure 7.4.



<i>Count</i>	<i>Assembled in</i>		<i>% Rows</i>	<i>Assembled in</i>		<i>Total</i>	
	<i>U.S.</i>	<i>Mexico</i>		<i>U.S.</i>	<i>Mexico</i>		
<i>Age</i>			<i>Age</i>				
<i>Under 28</i>	35	56	91	<i>Under 28</i>	38%	62%	100%
<i>28 to 32</i>	29	51	80	<i>28 to 32</i>	36%	64%	100%
<i>33 Plus</i>	51	41	92	<i>33 Plus</i>	55%	45%	100%
<i>Total</i>	115	148	263	<i>Total</i>	44%	56%	100%

<i>Chi Square</i>	7.968
<i>df</i>	2
<i>p value</i>	0.02

Figure 7.4 Contingency analysis of U.S.- vs. Mexican-made car choices by age

A glimpse of the PivotChart confirmed suspicions that older buyers did seem to be, rejecting cars assembled in Mexico. The p-value for chi square was .02, indicating that the null hypothesis, lack of association, ought to be rejected. Choice between U.S.- and Mexican-made cars was associated with age category. Fifty-six percent of the entire sample across all ages chose cars assembled in Mexico. Within the oldest segment, however, the Mexican-assembled car share was lower: 45%. While nearly two-thirds of the younger segments chose cars assembled in Mexico, less than half of the oldest buyers chose Mexican-made cars.

The CEO was alarmed with these results. His company could lose the business of older, more experienced buyers markets if production were shifted South of the Border. Brand managers were about to begin planning “Made in the U.S.A.” promotional campaigns targeted at the oldest car buyers. Emily Ernst, the Director of Strategy and Planning, suggested that age was probably not the correct basis for segmentation. She explained that the older buyers shop for a particular *type* of car—a family sedan or station wagon—and few family sedans or wagons were being assembled in Mexico. Models assembled at home in the U.S. tended to be large sedans and station wagons—styles sought by older buyers. She proposed that it was *style* that influenced the U.S.- versus Mexican-assembled choice, and not age, and that it was *style* that was dependent on age. Her hypotheses were:

$H_0$  : Choice of car style is not associated with age category.

$H_1$  : Choice of car style is associated with age category.

To explore this alternate hypothesis, the research team ran contingency analysis of style choice (SUV, Sedan/Wagon and Coupe) by age category, Figure 7.5 and Table 7.3.

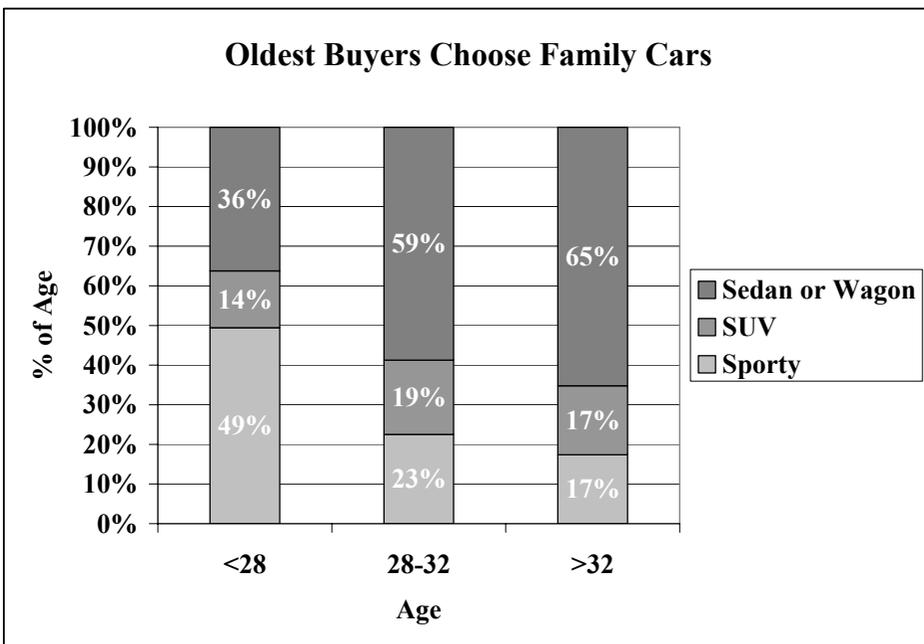


Figure 7.5 Contingency analysis of car style choice by age category

<i>Count</i>	<i>Style</i>			
<i>Age</i>	<i>sedan/wagon</i>	<i>coupe</i>	<i>SUV</i>	<i>Total</i>
< 28	33	45	13	91
28 to 32	47	18	15	80
33+	60	16	16	92
<i>Total</i>	140	79	44	263

<i>Row%</i>	<i>Style</i>			
<i>Age</i>	<i>sedan/wagon</i>	<i>coupe</i>	<i>SUV</i>	<i>Total</i>
< 28	36%	49%	14%	100%
28 to 32	59%	23%	19%	100%
33+	65%	17%	17%	100%
<i>Total</i>	53%	30%	17%	100%

$\chi^2_4$	<b>26.2p value</b>	<b>.0000</b>
------------	--------------------	--------------

Table 7.3 Contingency analysis of car style by age

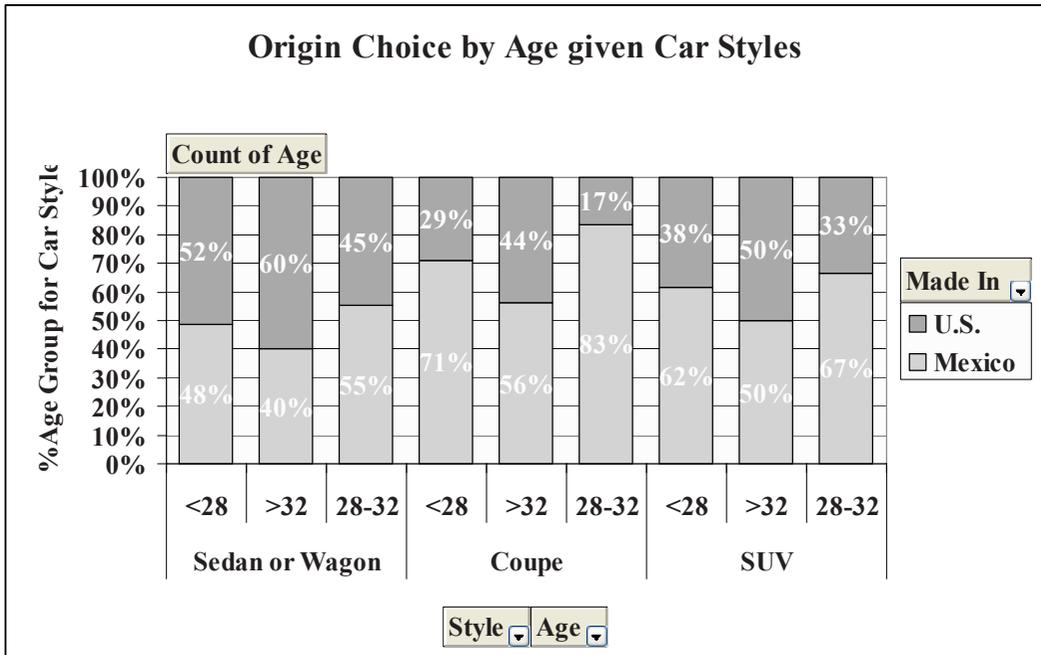
Contingency analysis of this sample indicates that choice of style is associated with age category. More than half (53%) of the car buyers chose a sedan or wagon, though only about a third (36%) of the younger buyers chose a sedan or wagon, and nearly twice as many (65%) older buyers chose a sedan or wagon. Thirty percent of the sample bought a coupe, and just nearly half (49%) of the younger buyers chose a coupe. Only 17% of the oldest buyers bought a coupe. These are significant differences supporting the conclusion that style of car chosen is associated with age category.

This is the news that the CEO was looking for. If older car buyers are choosing U.S.-made cars because they desire family styles, sedans and wagons, which tend to be assembled in the U.S., then perhaps these older buyers aren't shunning Mexican-made cars. His hypotheses were:

$H_0$ : Given choice of a sedan or wagon, choice of U.S.- versus Mexican- assembled is not associated with age category.

$H_1$ : Given choice of a sedan or wagon, choice of U.S.- versus Mexican-assembled is associated with age category.

To test these hypotheses, the analysis team conducted three contingency analyses of origin choice (U.S.- versus Mexican-assembled ) by age category, looking at each style separately in Figure 7.6.



%Age given Style		Made In:			$\chi^2$	df	p value
Style	Age	Mexico	U.S.	Total			
sedan or wagon	under 28	48%	52%	100%			
	28 to 32	40%	60%	100%			
	33 plus	55%	45%	100%			
total		47%	53%	100%	2.5	2	.29
coupe	under 28	71%	29%	100%			
	28 to 32	56%	44%	100%			
	33 plus	83%	17%	100%			
total		71%	29%	100%	3.0	2	.22
SUV	under 28	62%	38%	100%			
	28 to 32	50%	50%	100%			
	33 plus	67%	33%	100%			
total		59%	41%	100%	.9	2	.63
Grand Total		56%	44%	100%			

Figure 7.6 Contingency analysis: Origin choice by age given style

Controlling for style of car by looking at each style separately reveals lack of association between origin preference for U.S.- versus Mexican-made cars and age category. Across all three car styles, *p values* are greater than .05. There is not sufficient evidence in this sample to reject the null hypothesis. We conclude from this sample that the U.S.- versus Mexican-assembled choice is not associated with age category. The domestic automobile manufacturer should therefore not alter plans to move production South.

*Simpson's Paradox* describes the situation where two variables appear to be associated only because of their mutual association with a third variable. If the third variable is ignored, results are misleading. Because contingency analysis focuses upon just two variables at a time, analysts should be aware that apparent associations may come from confounding variables, as the **American Cars** example illustrates.

The Research Team summarized these results in this memo:

## MEMO

**Re: Country of Manufacture Does Not Affect Older Buyers’ Choices**

**To:** CEO, American Car Company

Emily Ernst, Director of Planning and Strategy

Brand Management

**From:** Travis Hendershott, Director of Quantitative Analysis

Analysis of a sample of new car buyers reveals that styles of car drive brand choices of distinct age segments. Brand choices of all ages of buyers are independent of country of manufacture.

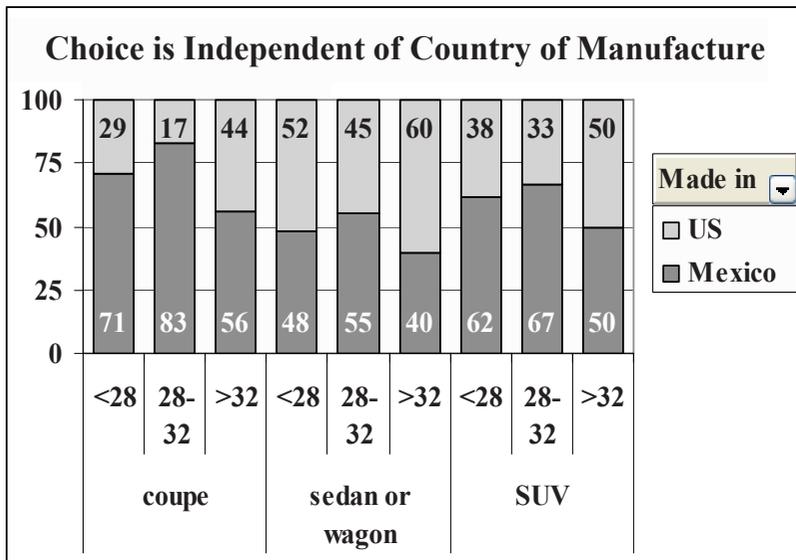
**Contingency Analysis.** Brand choices of 263 new car buyers were analyzed to assess the dependence of choice on country of manufacture, U.S. or Mexico, and age category.

**Results.** Choice between U.S.- and Mexican-assembled cars is not associated with age category.

*Style of car chosen is associated with age category.*

Younger buyers are more likely to choose a sporty coupe.

Older buyers are more likely to buy a sedan or wagon.



$\chi^2 = 2.5, ns; \chi^2 = 3.0, ns; \chi^2 = .9, ns$

**Conclusions.**

Production in Mexico is not expected to affect car buyer choices, providing the opportunity to shift assembly South to take advantage of cheaper labor.

**Limitations.** A larger sample would enable examination of more representative age categories, and specifically, a broader middle segment and older oldest segment.

## 7.5 Contingency Analysis Is Demanding

Contingency analysis requires a large and balanced dataset to insure a stable chi square. Even large samples may contain small proportions of particular categories, forcing combinations that aren't ideal. In the **American Cars** example, a broad category was used for the oldest age segment, combining fairly different ages, 33 through 60, and a narrow category was defined for the middle age segment, ages 28 through 32. The sample, though large, was not balanced and contained a large proportion of car buyers ages 30 through 39. This group was split and combined with sparse younger and older age categories to allow expected cell counts greater than five. With smaller samples, we may be left with just two categories for a variable, which may limit hypothesis testing. In the **Recruiting Stars** example, final results could not be used to assess the association between recruiting and poor employee performance after Poor and Average performing employees were combined.

## 7.6 Contingency Analysis Is Quick, Easy, and Readily Understood

Despite the fairly demanding data requirements, contingency analysis is appealing because it is simple, and results are easily understood. For very large samples, sparse cells are not a problem and many categories may be used, increasing the specificity of results and allowing a range of hypothesis tests.

For smaller samples, other alternatives, such as logit analysis (discussed in detail in Chapter 13, exist for analyzing categorical variable associations. These carry fewer data demands and allow incorporation of multiple variables. Multivariate analysis helps us avoid drawing incorrect conclusions in cases where Simpson's Paradox might mislead.

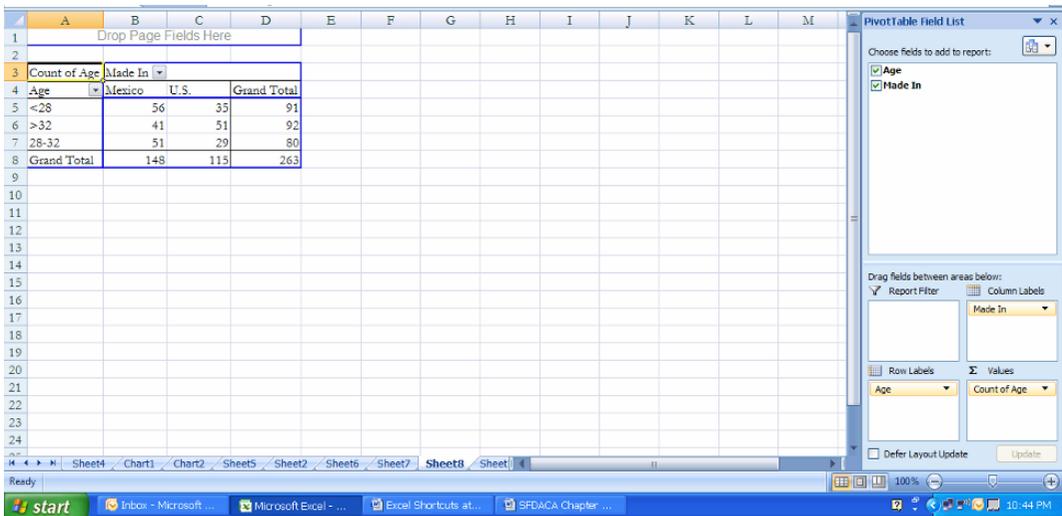
## Excel 7.1 Construct crosstabulations and assess association between categorical variables with PivotTables and PivotCharts

*American Cars.* In order to explore the possible association between choice of U.S.-assembled and Mexican-assembled cars by age, we will begin by making a *Pivot Table* to see the crosstabulation.

Open **Excel 7.1 American Cars.xls**.

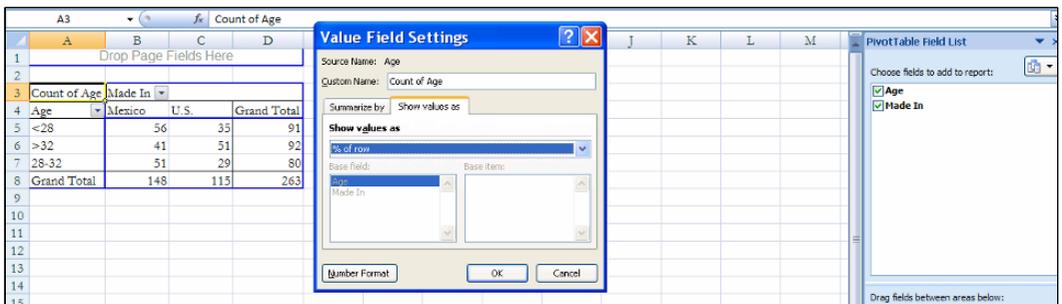
Select filled cells in the *Age* and *Made In* categories, in columns **A** and **B**, then insert a PivotTable.

Drag *Age* to **ROW**, *Made In* to **COLUMN**, and *Age* to **DATA**.



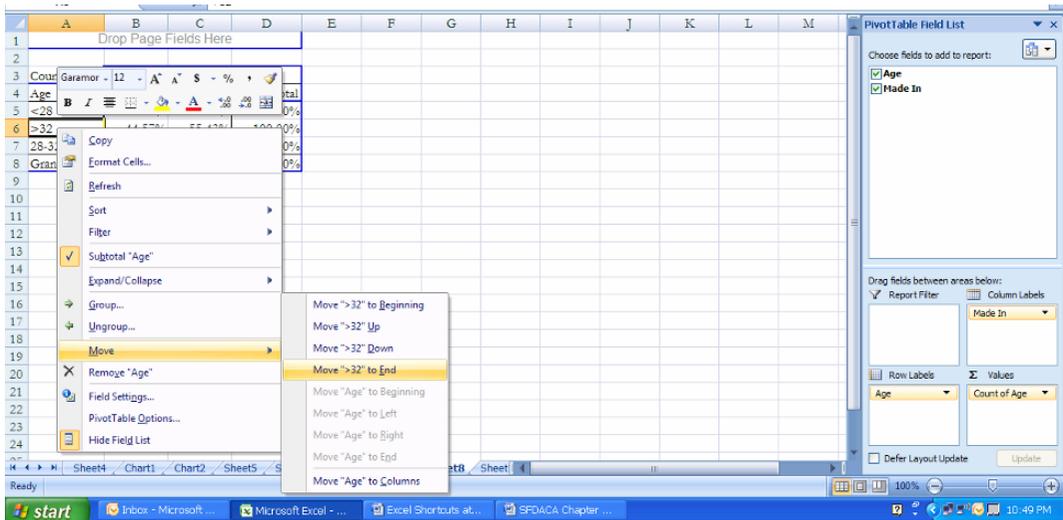
We are interested in the percent of each *age* category that choose cars *Made In* the U.S. and Mexico.

Double click **Count of Age** and **Show values as % Row**, **Ok**:  
 Select cells in the table, **B5:D8**, then reduce decimals.



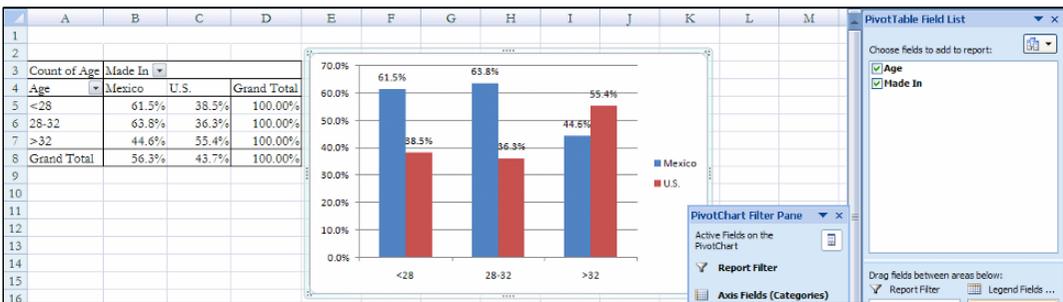
Age	Mexico	U.S.	Grand Total
<28	62%	38%	100%
>32	45%	55%	100%
28-32	64%	36%	100%
Grand Total	56%	44%	100%

To put the age categories in order, select and right click the >32 cell, **Move, Move to End.**

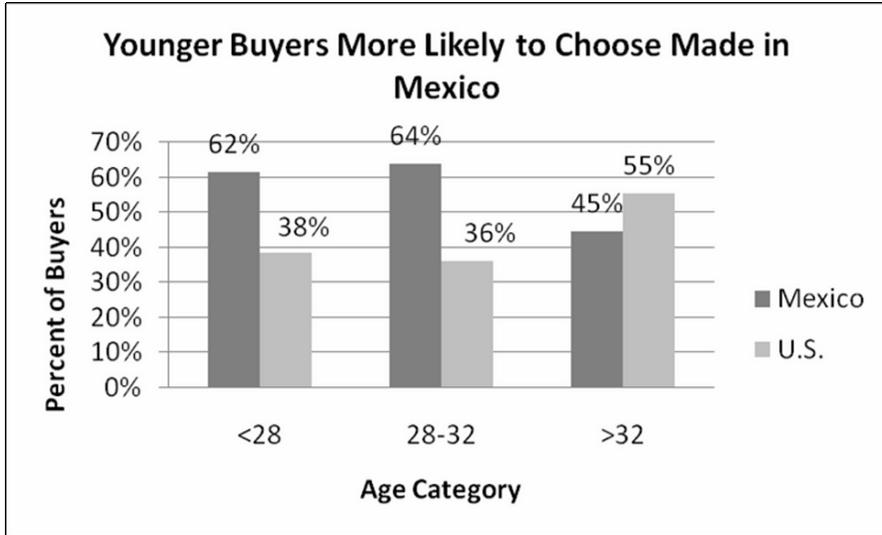


To see the **PivotChart** of *Made In* by *Age*, select the **PivotChart** icon:

**Add Data Labels.**



Add title and axes titles.



We see that fewer of the oldest car buyers, 45% bought cars assembled in Mexico, while a majority, 62 to 64%, of the younger buyers chose cars assembled in Mexico.

Excel 7.2 Use chi square to test association

To find the chi square statistic, change the PivotChart cells back to counts. Double click **Count of Age** in A3 and choose **Options, Normal, OK:**

The screenshot shows an Excel spreadsheet with a PivotTable in cells A3:A8. The PivotTable has 'Count of Age' in cell A3, 'Made In' in B3, and columns for 'Mexico', 'U.S.', and 'Grand Total'. The data rows are: <28 (56 Mexico, 35 U.S., 91 Grand Total), 28-32 (51 Mexico, 29 U.S., 80 Grand Total), >32 (41 Mexico, 51 U.S., 92 Grand Total), and Grand Total (148 Mexico, 115 U.S., 263 Grand Total).

Two dialog boxes are open:

- PivotTable Field:** Source field: Age; Name: Count of Age; Summarize by: Count; Show data as: Normal.
- PivotTable Field List:** Drag items to the PivotTable report; Items: Age, Made In; Add To: Row Area.

For chi square, we will make a table of *expected* cell counts. We will also make a table of cell contributions to chi square.

Select the two empty rows above the PivotTable, plus the PivotTable, **A1:D8**, then use shortcuts to copy, **Cntl+C**.

Paste into **E1:H8** with values and formats, but not formulas, using the shortcut: **Alt VSU**, **Ok**.

Paste a second copy into **I1:L8**, again with values and formats, but not formulas:

Change the table title in **E3** to **Expected**.

Change the table title in **I3** to **Chi square**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3	Count of Age	Made In			expected	Made In			chi-square	Made In							
4	Age	Mexico	U.S.	Grand Total	Age	Mexico	U.S.	Grand Tot	Age	Mexico	U.S.	Grand Total					
5	<28	56	35	91	<28	56	35	91	<28	56	35	91					
6	28-32	51	29	80	28-32	51	29	80	28-32	51	29	80					
7	>32	41	51	92	>32	41	51	92	>32	41	51	92					
8	Grand Total	148	115	263	Grand Tot	148	115	263	Grand Tot	148	115	263					

A cell  $e_{ij}$  in the  $i$ 'th row and  $j$ 'th column of the *expected* table is the product of

- the row proportion,  $\left(\frac{n_{age_i}}{N}\right)$ , the percent <28, 28-32 or >32, **\$D5/\$D\$8**, **\$D6/\$D\$8**, or **\$D7/\$D\$8**,
- the column proportions,  $\left(\frac{n_{MadeIn_j}}{N}\right)$ , or the percent of cars made in Mexico or the U.S., **B\$8/\$D\$8**, and **C\$8/\$D\$8**, and
- the sample size, **\$D\$8**:

$$e_{ij} = \left(\frac{n_{age_i}}{N}\right) \left(\frac{n_{MadeIn_j}}{N}\right) N$$

$$= n_{age_i} n_{MadeIn_j} / N.$$

A dollar sign with **D** locks the column and a dollar sign with **8** locks the row, so that we can grab and drag the formula through the table:

In **F5** enter the formula for the expected count,  $n_{<28} * n_{Mexico} / N$

**=\$D5\*B\$8/D8 f4 [Enter]**.

Select the new cell, grab and drag over through **H** and down through **8**, filling in the expected table:

Count of Age	Made In			expected	Made In			chi-square	Made In		
Age	Mexico	U.S.	Grand Total	Age	Mexico	U.S.	Grand Tot	Age	Mexico	U.S.	Grand Total
<28	56	35	91	<28	51.20913	39.79087	91	<28	56	35	91
28-32	51	29	80	28-32	45.01901	34.98099	80	28-32	51	29	80
>32	41	51	92	>32	51.77186	40.22814	92	>32	41	51	92
Grand Total	148	115	263	Grand Tot	148	115	263	Grand Tot	148	115	263

Find each (row,column) cell’s contribution to chi square, the squared difference between expected  $e_{i,j}$  and actual counts  $n_{i,j}$  in the cell in the  $i$ 'th column and  $j$ 'th row, divided by expecteds:

$$\chi^2_{i,j} = \frac{(e_{i,j} - n_{i,j})^2}{e_{i,j}}$$

In the first cell of the chi square table, **J5**, enter  $=\text{(F5-B5)}^2/\text{F5}$ .

Select **J5**, grab and drag over through **K** and down through row **7**:

Count of Age	Made In			expected	Made In			chi-square	Made In		
Age	Mexico	U.S.	Grand Total	Age	Mexico	U.S.	Grand Tot	Age	Mexico	U.S.	Grand Total
<28	56	35	91	<28	51.20913	39.79087	91	<28	0.448211	0.576828	91
28-32	51	29	80	28-32	45.01901	34.98099	80	28-32	0.794603	1.022619	80
>32	41	51	92	>32	51.77186	40.22814	92	>32	2.241237	2.884375	92
Grand Total	148	115	263	Grand Tot	148	115	263	Grand Tot	148	115	263

In **J8** enter the label *chisquare*, then use the Excel function **SUM(array1,array2)** to add the cell contributions to find the chi square statistic.

In **L8** enter  $=\text{SUM}(J5:K7)$ [Enter]:

Count of Age	Made In			expected	Made In			chi-square	Made In		
Age	Mexico	U.S.	Grand Total	Age	Mexico	U.S.	Grand Tot	Age	Mexico	U.S.	Grand Total
<28	56	35	91	<28	51.20913	39.79087	91	<28	0.448211	0.576828	91
28-32	51	29	80	28-32	45.01901	34.98099	80	28-32	0.794603	1.022619	80
>32	41	51	92	>32	51.77186	40.22814	92	>32	2.241237	2.884375	92
Grand Total	148	115	263	Grand Tot	148	115	263	chi-square	148	115	7.967873

In **K9** type in the label *p-value*.

Use the Excel function **CHIDIST(chisquare,df)** with your **chisquare** in **L8** and degrees of freedom **df** of **2** ( $=\text{(number of rows-1)}*\text{(number of columns -1)}$ ):

In L9 enter =CHIDIST(L8,2) [Enter]:

Count of Age	Made In	expected	Made In	chi-square	Made In
Age	Mexico U.S.	Grand Total	Age	Mexico U.S.	Grand Total
<28	56 35	91	<28	51.20913 39.79087	91
28-32	51 29	80	28-32	45.01901 34.98099	80
>32	41 51	92	>32	51.77186 40.22814	92
Grand Total	148 115	263	Grand Tot	148 115	263
			chi-square	148	115
					7.967873
					p-value
					0.018612

Based on sample evidence, we reject the null hypothesis that country of manufacture and age are independent. We conclude that the choice between cars made in the U.S. and cars made in Mexico depends on age.

### Excel 7.3 Conduct contingency analysis with summary data

Sometimes our data are in summary form. That is, we know the sample size, and we know the percent of the sample in each category.

*Marketing Cereal to Children.* Kooldogg expects that many Saturday morning cartoon viewers would be attracted to their sugared cereals. A heavy advertising budget for sugared cereals is allocated to Saturday morning television. We will use contingency analysis to analyze the association between Saturday morning cartoon viewing and frequent consumption of Kooldogg cereal with sugar added. From a survey of 300 households, we know whether or not children ages 2 through 5 *Watch Saturday Morning Cartoons* on a regular basis (at least twice a month) and whether or not those children *Eat Kooldogg Cereal with Added Sugar* (at least once a week).

### Open Excel 7.3 Kooldogg Kids Ads.xls.

Select the summary data in columns **A, B, and C**, and make a PivotTable, with *Watches Saturday Morning Cartoons* in **ROW**, *Eats Kooldogg Sugary Cereal* in **COLUMN**, and drop *Number of Children* in **DATA**:

	A	B	C	D	E	F	G	H	I	J	K	L	M
3	Sum of Number of children	Column Labels											
4	Row Labels	0	1	Grand Total									
5	0	36	4	40									
6	1	4	256	260									
7	Grand Total	40	260	300									

Copy rows 1 and 2 with the table and paste with formats and values, Alt HVSU, into E1:H7 and I1:I7.

Find the expected cell counts in **E5:F6** under the assumption that Kooldog cereal consumption is independent of Saturday morning TV viewing.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3	Sum of Number of children	Column Labels			expected	Column Labels		chisquare	Column Labels								
4	Row Labels		0	1	Grand Total	Row Label	0	1	Grand Tot	Row Label	0	1	Grand Total				
5	0		36	4	40	0	5.333333	34.66667	40	0	36	4	40				
6	1		4	256	260	1	34.66667	225.3333	260	1	4	256	260				
7	Grand Total		40	260	300	Grand Tot	40	260	300	Grand Tot	40	260	300				

Find cell contributions to chi square in **J5:K6**, with squared differences between expected cell counts in **F5:G6** and actual cell counts in **B5:C6**, divided by expected cell counts in **F5:G6**.

Sum the cell contributions to chi square in **J5:K6** to find chisquare in **L7**.

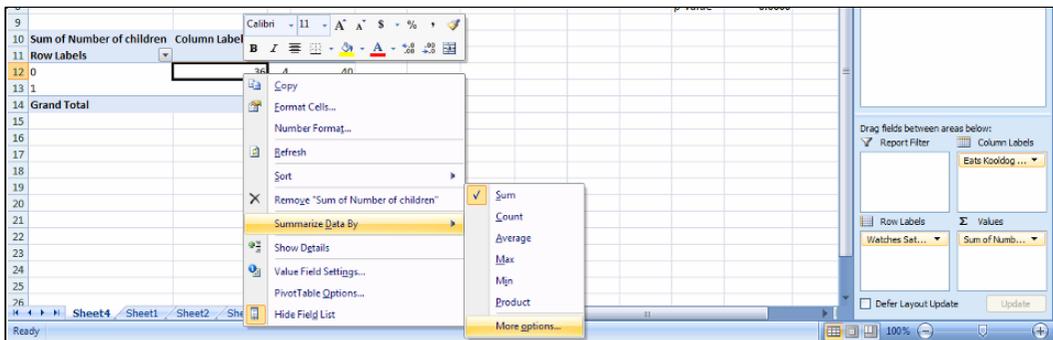
In **L8**, use **CHIDIST()** to Find the *p-value* of chi square in **L7**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3	Sum of Number of children	Column Labels			expected	Column Labels		chisquare	Column Labels								
4	Row Labels		0	1	Grand Total	Row Label	0	1	Grand Tot	Row Label	0	1	Grand Total				
5	0		36	4	40	0	5	35	40	0	176.3	27.1	40				
6	1		4	256	260	1	35	225	260	1	27.1	4.2	260				
7	Grand Total		40	260	300	Grand Tot	40	260	300	chisquare	40	260	234.8				
8													p-value				0.0000

The *p value* is very small (with 53 zeros following the decimal point). Based on sample evidence, we reject the null hypothesis of independence and conclude that eating cereal with added sugar is associated with Saturday morning cartoon viewing.

To see the association, copy rows **1** and **2** with the PivotTable **A1:D7**, and paste below the original in **A8:D12**, this time with formulas, using **Cntl+V**:

Change the cell counts to percents of row: Right click a cell in the copied table, **Summarize Data By, More options, Show Data As: % of row, OK**:



Select **B11** and type in the label **Don't Eat Kooldogs**.

Select **C11** and type in the label **Eat Kooldogs**.

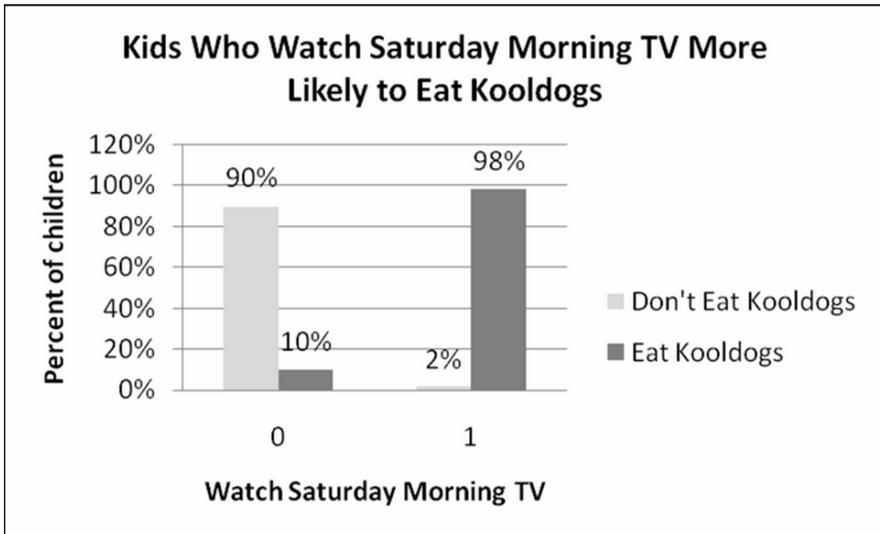
	0	1	Grand Total
Sum of Number of children	4	256	260
Grand Total	40	260	300

Row Labels	Don't Eat Kooldogs	Eat Kooldogs	Grand Total
0	90.00%	10.00%	100.00%
1	1.54%	98.46%	100.00%
Grand Total	13.33%	86.67%	100.00%

Make a **PivotChart** with shortcuts **Alt JTC** to see the association. (**JT** selects the Pivot menu and **C** inserts a **PivotChart**.)

Add data labels, a title and axes titles:



Management would conclude:

*The majority of children surveyed (87%) eat Kooldogg cereal with added sugar and an even greater proportion, 98%, of those who watch Saturday morning cartoons eat our cereal with added sugar. In contrast, only 10% of children who do not watch Saturday mornings eat our cereal with added sugar. Since most children (87%) watch Saturday morning cartoons, our heavy advertising in this time slot seems justified, since evidence suggests that consumption of our sugared cereals is associated with Saturday cartoon viewing.*

## Excel Shortcuts at Your Fingertips

### By Shortcut Key

**Alt** activates the shortcuts menus, linking keyboard letters to Excel menus. Press **Alt**, then release and press letters linked to the menus you want.

The following are examples of shortcuts. Press **Alt**, then

**H 9** to select the Home menu and the reduce decimals function

**H DC** to select the Home menu and the Delete function to delete column(s)

**H IC** to select the Home menu and Insert function and to insert a column to the left of the selected cell or column

**HIE** selects the Home menu and Insert function and inserts cut or copied cells to the left of the selected column or cell

**AY2** to select the Data and Data Analysis menus

**AS** to select the Data and the Sort menus

**NC** to select the Insert function and to insert a column chart

**ND** to select the Insert function and to insert a scatterplot

**NE** to select the Insert function and to insert a pie chart

**NVT** to select the Insert function, the Pivot menu, and to insert a PivotTable

**NX** to select the Insert function and to insert a text box

**WFR** to select the View and Freeze panes menus, and to Freeze rows

**JAB** to select the Layout and Data Labels menus

**JARM** to select the Layout, the Error Bar, and the custom Error Bar menus

**JAT** to select the Layout and Title menus

**JAI** to select the Layout and Axis Labels menus

**JTC** to make a **PivotChart** from a PivotTable

**VSU** to paste with values and formats, but not formulas

**Shift+arrow** selects cells scrolled over

**Cntl+C** to copy

**Cntl+X** cuts selected cells and places them on the clipboard.

**Cntl+down arrow** scrolls through all cells in the same column that contain data and stops at the last filled cell.

**Cntl+R** fills in values of empty cells using a formula from the first cell in a selected array

**Cntl+Shift+down arrow** selects all filled cells in the column.

## By Goal

If you want to

Activate shortcuts menus, press **Alt**, then release.

Add data labels in a column chart: select a column, then **Alt JAB**

Add error bars in a column chart: select a column, then **Alt JARM**

Add a title: **Alt JAT**

Add axis label: **Alt JAI**

Analyze data: **Alt AY2**

Copy cells: select the cells, then **Cntl+C**

Delete a column: **Alt HDC**

Freeze the top row: **Alt WFR**

Insert copied cells: **Alt HIE**

Insert a column: **Alt HIC**

Insert a column chart: **Alt NC**

Insert a pie chart: **Alt NE**

Insert a PivotChart from a Pivot Table: **Alt JTC**

Insert a PivotTable: **Alt NVT**

Insert a row: **Alt HIR**

Insert a scatterplot: **Alt ND**

Insert a text box: **Alt NX**

Move cells or a column: select the cells or column, **Cntl+X**, then select the new location, **Alt HIE**

Move to the end of a column: **Cntl+down arrow**

Paste with values and formats, but not formulas: **Alt VSU**

Reduce decimals: **Alt H9**

Select all of the filled cells in a column: select the first cell in the column, then **Cntl+Shift+down arrow**

Sort data: **Alt AS**

## Assignment 7-1 747s and Jets<sup>2</sup>

Boeing Aircraft Company management believes that demand for particular types of aircraft is associated with particular global region across their three largest markets, North America, Europe, and China. To better plan and set strategy, they have asked you to identify region(s) where demand is uniquely strong for 747s and for regional jets.

**Assignment 7-1 JETS747.xls** contains Boeing's actual and projected deliveries 2005-2024 of each type of aircraft in each of the three regions.

- a. Use contingency analysis to test the hypothesis that *demand* for particular aircraft is associated with *global region*.
- b. If the association is significant, explain the nature of association.
- c. Include a PivotChart and explain what it illustrates.

## Assignment 7-2 Fit Matters

Procter & Gamble management would like to know whether intent to try their new preemie diaper concept is associated with the importance of fit. If Likely Triers value fit more than Unlikely Triers, fit could be emphasized in advertisements.

**Assignment 7-2 Fit Matters.xls** contains data from a concept test of 97 mothers of preemie diapers, including trial *Intention* and *Fit Importance*, measured on a 9-point scale.

You may decide to combine categories.

- Use contingency analysis to test the hypothesis that *intent* to try is associated with the *importance of fit*.
- If the association is significant, explain the nature of association.
- Include a PivotChart and explain what it illustrates.

---

<sup>2</sup> This case is a hypothetical scenario using actual data.

### *Assignment 7-3 Allied Airlines*

Rolls-Royce management has observed the growth in commercial airline alliances. Airline companies which are allied tend to purchase the same aircraft. Management would like to know whether or not alliance is associated with global region.

Data including the number of allied airline companies, *Allied*, and *Global Region* are contained in **Assignment 7-3 Allied Airlines.xls**.

You may decide to combine global regions.

- Use contingency analysis to test the hypothesis of association between *alliance* and *global region*.
- If the association is significant, describe the nature of association.
- Include a PivotChart and explain what it illustrates.

## *CASE 7-1 Hybrids for American Car*

Rising gas prices and environmental concerns have led some customers to switch to hybrid cars. In 2004, sales of hybrids increased by 81%, nearly doubling 2003 hybrid sales. Nonetheless, Polk Research reports that less than one percent (.081%) switched from conventional cars to hybrids in the 12 months of 2005.

American Car (AC) offers two hybrids, AC Sapphire and AC Durado, an SUV and a pickup. AC offers no hybrid automobiles. Major competitors, Ford, Toyota and Honda, offer hybrid automobiles. AC executives believe that with their hybrid SUV and pickup, they will be able to attract loyal AC customers who desire a hybrid. Shawn Green, AC Division Head, is worried that customers who were driving sedans, coupes or wagons may not want a truck or an SUV. They might switch from AC to Ford, Toyota or Honda in order to purchase a hybrid car.

To investigate further, Mr. Green commissioned a survey of car buyers. The new car purchases of a representative random sample of 4,000 buyers were sorted into eight groups, based on the type of car they had owned and *Traded* (Prestige, Sport, Compact SUV, Large, and Full-size SUV) and whether or not they bought *Hybrid* or *Conventional*. These data are in **Case 7-1 Hybrid.xls**. The number of *Buyers* indicates popularity of each *Traded, Hybrid* combination.

Conduct contingency analysis with this data to determine whether *choice of hybrid vehicles* depends on *type of vehicle owned previously*.

Specifically,

- Is there an association between the *type of car owned and Traded* and *choice of a Hybrid* instead of a Conventional car?  
In other words, are owners of particular types of cars more likely than others to trade for a hybrid?
- What is the probability that a new car buyer will choose a *hybrid*?
- Which segments are more likely than others to switch to hybrids, and exactly how likely is hybrid choice among these segments?

Illustrate your results with a PivotChart. Include a bottom-line title.

What are the implications of results for American Car Division?

What is your advice to Mr. Green?

## CASE 7-2 Tony's GREAT Advertising

Kellogg spends a hefty proportion of its advertising budget to expose children to ads for sweetened cereal on Saturday mornings. Kellogg brand ads feature cartoon hero characters similar to the cartoon hero characters that children watch on Saturday morning shows. This following press release is an example:

*Advertising Age*, Dec 6, 2004 v75 i49 p1

**Kellogg pounces on toddlers; Tiger Power to wrest tot monopoly away from General Mills' \$500M Cheerios brand.** (News) *Stephanie Thompson*.

Byline: STEPHANIE THOMPSON

In the first serious challenge to General Mills' \$500 million Cheerios juggernaut, Kellogg is launching a toddler cereal dubbed Tiger Power.

The cereal, to arrive on shelves in January, will be endorsed by none other than Frosted Flakes icon Tony the Tiger and will be "one of our biggest launches next year," according to Kellogg spokeswoman Jenny Enochson. Kellogg will position the cereal-high in calcium, fiber and protein-as "food to grow" for the 2-to-5 set in a mom-targeted roughly \$20 million TV and print campaign that begins in March from Publicis Groupe's Leo Burnett, Chicago.

Cereal category leader Kellogg is banking on Tiger Power's nutritional profile as well as the friendly face of its tiger icon, a new shape and a supposed "great taste with or without milk" to make a big showing in take-along treats for tots.

Kellogg spent \$7.3 million on Frosted Flakes in 2003 and \$7 million on the brand for January through July of this year.

Tony Grate, the brand manager for Frosted Flakes would like to know whether there is an association between Saturday morning cartoon viewing and consumption of his brand.

The Saturday morning TV viewing behaviors, *Saturday Morning Cartoons*, and consumption of Frosted Flakes, *Frosted Flake Eater*, are contained in **Case 7-2 Frosted Flakes.xls**. A random sample of 300 children ages 2 through 5 were sorted into four groups based on whether or not each watches at least three hours of television on Saturday morning at least twice a month and whether or not each consumes Frosted Flakes at least twice times a week. The number of *Children* indicates popularity of each *Saturday Morning Cartoons*, *Frosted Flake Eater* combination.<sup>3</sup>

---

<sup>3</sup> These data are fictitious, though designed to reflect a realistic scenario.

- Is there an association between watching *Saturday morning cartoons* and consumption of Frosted Flakes?
- What is the probability that a *cartoon watcher* consumes Frosted Flakes?
- Which group is more likely to consume Frosted Flakes, and exactly how likely is Frosted Flake consumption among this group?

Illustrate your results with a properly labeled PivotChart. Include a bottom-line title.

What are the implications of results for Tony Grate?

## 8

# Building Multiple Regression Models

Models are used to accomplish *two* complementary goals: *identification of key drivers of performance* and *prediction of performance under alternative scenarios*. The variables selected affect both the explanatory accuracy and power of models, as well as forecasting precision. In this chapter, we focus on variable selection, the first step in the process used to build powerful and accurate multiple regression models.

We use logic to choose variables initially. Some of the variables which logically belong in a model may be insignificant, either because they truly have no impact, or because their influence is part of the joint influence of a correlated set of predictors which together drive performance. *Multicollinear* predictors create the illusion that important variables are insignificant. *Partial F test(s)* are used to decide whether seemingly insignificant variables contribute to variance explained. If an insignificant predictor adds no explanatory power, it is removed from the model. It is either not a performance driver, or it is redundant because other variables reflect the same driving dimension. Using *partial F tests* does not cure multicollinearity, but acknowledges its presence and helps us assess the incremental worth of variables that may be redundant or insignificant.

### 8.1 Multiple Regression Models Identify Drivers and Forecast

Multiple regression models are used to achieve two complementary goals: identification of key *drivers* of performance and prediction of performance under alternative scenarios. This prediction can be either what would have happened had an alternate course of action been taken, or what can be expected to happen under alternative scenarios in the future.

Decision makers want to know, given uncontrollable external influences, which controllable variables make a difference in performance. We also want to know the nature and extent of each of the influences when considered together with the full set of important influences. A multiple regression model will provide this information.

Once key drivers of performance have been identified and our model has been validated, we can use it to compare performance predictions, either of the past or in the future, under alternative scenarios. This *sensitivity analysis* allows managers to compare expected performance levels and to make better decisions.

### 8.2 Use Your Logic to Choose Model Components

The first step in model building happens before we look at data or use software. Using logic, personal experience, and others' experiences, we first decide which of the potential influences ought to be included in a model. From the set of variables with

available data, which could reasonably be expected to influence performance? In most cases, we need a reason for including each independent variable in our model. Independent variables tend to be related to each other in our correlated world, and we unnecessarily complicate models if we include variables which don't logically affect the dependent performance variable. We will explore this complication from correlated predictors, *multicollinearity*, later in the chapter.

*Example 8.1 Sakura Motors Quest for Cleaner Cars.* The new product development group at Sakura Motors is in the midst of designing a new line of cars which will offer reduced greenhouse gas emissions for sale to drivers in global markets where air pollution is a major concern. They expect to develop a car that will emit only 5 tons of greenhouse gases per year.

What car characteristics drive emissions? The management team believes that smaller, lighter cars with smaller, more fuel efficient engines will be cleaner. The U.S. Government publishes data on the fuel economy of car models sold in the U.S. ([fuelconomy.gov](http://fuelconomy.gov)), which includes *manufacturer*, *model*, engine size (*cylinders*), and gas mileage (*MPG*) for each category of car. This data source also includes *emissions* of tons of greenhouse gases per year. A second database, [consumerreports.org](http://consumerreports.org), provides data on acceleration in *seconds* to go from 0 to 60 miles per hour, which reflects car model sluggishness, and two measures of size, *passengers* and *curb weight*. Management believes that responsiveness and size may have to be sacrificed to build a cleaner car.

The multiple linear regression model of *emissions* will include these car characteristics, *miles per gallon (MPG)*, *seconds* to accelerate from 0 to 60, *horsepower*, *liters*, *cylinders*, *passenger* capacity, and weight in *pounds(K)*, each thought to drive *emissions*:

$$\begin{aligned} \text{emissions}_i = & b_0 + b_1 \text{MPG}_i + b_2 \text{seconds}_i + b_3 \text{pounds}(K)_i + b_4 \text{passengers}_i + b_5 \text{horsepower}_i \\ & + b_6 \text{cylinders}_i + b_7 \text{liters}_i \end{aligned}$$

Where  $\text{emissions}_i$  is the expected tons of annual emissions of the  $i$ th car model,

$b_0$  is the intercept indicating expected emissions if *MPG*, *seconds*, *pounds(K)*, *passengers*, *horsepower*, *cylinders* and *liters* were zero,

$b_1, b_2, b_3, b_4, b_5, b_6, b_7$  are the regression coefficient estimates indicating the expected marginal impact on emissions of a unit change in each car characteristic when other characteristics are at average levels, and

$\text{MPG}_i, \text{seconds}_i, \text{horsepower}_i, \text{cylinders}_i, \text{liters}_i, \text{passengers}_i, \text{pounds}(K)_i$  are characteristics of the  $i$ th car model.

When we include more than one independent variable in a linear regression, the coefficient estimates, or parameters estimates, are *marginal*. They estimate the marginal impact of each predictor on performance, given average levels of each of the other predictors.

The new product development team asked the model builder to choose a sample of car models which represents extremes of emissions, worst and best. Thirty-five car models were included in the sample. These included imported and domestic cars, subcompacts, compacts, intermediates, full-size sedans, wagons, SUVs, and pickups. Within this set there are considerable differences in all of the car characteristics, shown in Table 8.1.

<i>Car Characteristic</i>	<i>Minimum</i>	<i>Median</i>	<i>Maximum</i>
<i>Emissions (tons)</i>	5.2	8.7	12.5
<i>MPG</i>	15	22	34
<i>Seconds (0 to 60)</i>	7	9	12
<i>Passengers</i>	4	5	9
<i>Pounds(K)</i>	2.5	4.0	5.9
<i>Horsepower</i>	108	224	300
<i>Cylinders</i>	4	6	8
<i>Liters</i>	1.5	3.3	6.0

Table 8.1 Car characteristics in the Sakura Motors sample

### 8.3 Multicollinear Variables Are Likely When Few Variable Combinations Are Popular In a Sample

Since these data come directly from the set of cars actually available in the market, many characteristic combinations do not exist. For example, there is no car with a 1.5 liter engine that weighs 4,000 pounds. We expect the seven car characteristics to be related to each other. We are knowingly introducing correlated independent variables, also called *multicollinear* independent variables, into our model, because the characteristic combinations which are not represented do not exist.

Multiple linear regression will identify the car characteristics related linearly to emissions. Results from Excel are shown in Table 8.2.

---

**SUMMARY OUTPUT**


---

*Regression Statistics*

Multiple R	
R Square	0.928
Adjusted R Square	0.908
Standard Error	.644
Observations	34

---

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	7	138	19.8	47.7	0.0001
Residual	26	11	.4		
Total	33	149			

---

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>
Intercept	9.2	1.90	4.8	<.0001
<i>seconds</i>	.23	.099	2.3	.03
<i>mpg</i>	-.23	.037	-6.2	<.0001
<i>liters</i>	.41	.29	1.4	.17
<i>cylinders</i>	-.035	.19	-.2	.85
<i>horsepower</i>	-.00052	.0037	-.1	.89
<i>pounds (K)</i>	.54	.30	1.9	.08
<i>passengers</i>	-.086	.12	-.7	.48

---

Table 8.2 Multiple linear regression of emissions with seven car characteristics

*RSquare* is .928, or 93%, indicating that, *together*, variation in the seven car characteristics account for 93% of the variation in emissions. The *standard error* is .644, which indicates that forecasts of emissions would be within 1.29 tons of average actual emissions for a particular car configuration.

#### 8.4 *F* Tests the Joint Significance of the Set of Independent Variables

*Significance F* is = .0001, indicating that it is unlikely that we would observe these data patterns, were none of the seven car characteristics driving emissions. It may be that just one of the seven characteristics drives emissions, or it may be that all seven are significant influences. The *F* test is a general test of the percent of variation explained by the set of predictors together, and, equivalently, a test of the hypothesis that *RSquare* is 0%.

## 8.5 Insignificant Parameter Estimates Signal Multicollinearity

To determine which of the seven car characteristics are significant drivers of emissions, we initially look at the significance of *t tests* of the individual regression parameter estimates. Results suggest that only *seconds* to accelerate 0 to 60 and *MPG* drive differences in emissions. Neither engine size characteristics, *horsepower*, *liters* and *cylinders*, nor car size characteristics, *passengers* or *pounds (K)* appears to influence emissions. Coefficient estimates for *horsepower*, *cylinders* and *passengers* have the “wrong signs.” Larger cars with larger engines are expected to emit more pollutants. These are surprising and nonintuitive results.

When predictors which ought to be significant drivers appear to be insignificant, or when parameter estimates are of the wrong sign, we suspect *multicollinearity*. Multicollinearity, the correlation between predictors, thwarts driver identification. When the independent variables are themselves related, they jointly influence performance. It is difficult to tell which individual variables are more important drivers, since they vary together. Because of their correlation, the standard errors  $s_{b_i}$  of the partial slope coefficient estimates,  $b_i$ , are inflated. We are not very certain of each true influence in the population since their influence is joint. Our confidence intervals of the true partial slopes are large, since these are multiples of the standard errors of the partial slope estimates. Individual predictors seem to be insignificant though they may be truly significant.

## 8.6 Combine or Eliminate Collinear Predictors

We have two remedies for multicollinearity cloudiness:

- We can combine correlated variables, and
- we can eliminate variables that are contributing redundant information.

Correlations between the predictors reveal that *horsepower*, *cylinders* and *liters* are highly correlated with each other ( $r_{horsepower,liters} = .76; r_{cylinders,liters} = .92; r_{cylinders,horsepower} = .77$ ) and with *seconds*, *MPG*, *pounds(K)*, and *passengers*, as shown in Table 8.3.

	<i>MPG</i>	<i>seconds</i>	<i>liters</i>	<i>horsepower</i>	<i>cylinders</i>	<i>pounds (K)</i>	<i>passengers</i>
<i>MPG</i>	1						
<i>seconds</i>	-.05	1					
<i>liters</i>	-.81	-.17	1				
<i>horsepower</i>	-.53	-.36	<b>.76</b>	1			
<i>cylinders</i>	-.74	-.19	<b>.92</b>	<b>.77</b>	1		
<i>pounds (K)</i>	-.77	-.01	.84	.72	.81	1	
<i>passengers</i>	-.53	-.05	.59	.55	.60	.70	1

Table 8.3 Pairwise correlations between predictors

Cars with larger engines have more power. We will eliminate *horsepower* and *cylinders* from the model, expecting that they are redundant measures of engine size. If explanatory power is not substantially reduced, we can designate *liters* as the measure of engine size which reflects *cylinders* and *horsepower*.

*Passenger* capacity is highly correlated with weight (*pounds(K)*):  $r_{\text{passengers,pounds}} = .70$ . Larger, more spacious cars weigh more. We will eliminate *passengers* from the model, expecting that it is a redundant measure of car size. If explanatory power is not sacrificed, *pounds(K)* will reflect car size. We will not eliminate multicollinearity, but we will reduce it by removing correlated predictors. The revised *partial* model becomes:

$$\text{emissions}_i = b_0 + b_1 \text{MPG}_i + b_2 \text{seconds}_i + b_3 \text{liters}_i + b_4 \text{pounds}(K)_i$$

Regression results using this *partial* model are shown in Table 8.4.

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
R Square				0.926	
Adjusted R Square				0.916	
Standard Error				0.617	
Observations				34	
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4	138	34.5	90.8	0.0000
Residual	29	11	.4		
Total	33	149			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	
Intercept	9.0	1.8	5.0	<.0001	
<i>seconds</i>	0.24	.087	2.8	.01	
<i>mpg</i>	-0.23	.034	-6.7	<.0001	
<i>liters</i>	0.36	.20	1.8	.08	
<i>pounds (K)</i>	0.43	.24	1.8	.08	

Table 8.4 Regression of emissions with four car characteristics

The *partial* model *RSquare*, .926, is less than one percentage point lower than the *full* model *RSquare*, .929. With just four of the seven car characteristics, we can account for 93% of the variation in emissions. We have lost little explanatory power and the standard error has dropped from .644 to .617, reducing the margin of error in forecasts by 4% ( $= (.644 - .617) / .644$ ). Model *F* is significant, suggesting that one or more of the four predictors influences emissions. Two of the predictors are significant drivers. All coefficient estimates have correct signs. As we found in the full model, emissions are lower for responsive cars with higher fuel economy.

## 8.7 Partial F Tests the Significance of Changes in Model Power

Can *horsepower*, *cylinders* and *passengers* be eliminated without loss of explanatory and predictive power? Multicollinearity is reduced when we remove variables, increasing the certainty of parameter estimates for variables left in the model. With this small change, we do not need to test the significance of the change in *RSquare*. When *RSquare* does change by more than 1%, we use a *Partial F* test to assess the significance of the decline:

$$F_{k-g, N-1-k} = \frac{(RSquare_{full} - RSquare_{partial}) / g}{(1 - RSquare_{full}) / (N - 1 - k)}$$

Where  $RSquare_{full}$  is *RSquare* from the larger model before variables are removed,  $RSquare_{partial}$  is *RSquare* from the smaller model after variables are removed,

$g$  is the number of predictors removed from the full model

$N$  is the sample size,

$k$  is the number of predictors in the full model, and

$(N-1-k)$  is the residual degrees of freedom (df) from the original model.

We expect a larger change in *RSquare* if we remove a larger number of variables, so the change comparison is per predictor removed,  $g$ .

In the **Sakura Motors** model, *Partial F* to test the significance of incremental explanatory power of *horsepower*, *cylinders* and *passengers* is:

$$F_{3,26} \frac{(.928 - .926)/(3)}{(1 - .928)/(34 - 1 - 7)} = \frac{.0017/3}{.072/26} = \frac{.00058}{.0028} = .21, \text{ Partial } F \text{ Significance} = .89$$

For these degrees of freedom, 3 and 26, an  $F$  value of .21 includes only 11% (=1-.89%) of the  $F$  distribution area and is smaller than the 95% required for significance of .05. *RSquare* did *not* change significantly when the three redundant variables were eliminated. *Horsepower*, *cylinders* and *passengers* do *not* add sufficient explanatory power to the model and will remain out. The partial model now becomes our full model.

Though we can confidently eliminate *horsepower*, *cylinders* and *passengers*, the model still contains two variables which aren't significant. *Pounds(K)* and *liters* may also be redundant, since both are highly correlated with fuel economy ( $r_{\text{pounds,mpg}} = -.77$ ;  $r_{\text{liters,mpg}} = -.81$ ) in Figure 8.5

	MPG	liters	pounds (K)
MPG	1		
liters	-.81	1	
pounds (K)	-.77	.84	1

Table 8.5 Pairwise correlations

We will eliminate these to reduce multicollinearity, observing the drop in explanatory power. Then we can again use a *partial F test* to decide whether they will remain out or return to the model. Regression results are in Table 8.6.

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
R Square		0.886			
Adjusted R Square		0.879			
Standard Error		.740			
Observations		34			
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	132	66.1	121	0.0000
Residual	31	17	.5		
Total	33	149			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	
Intercept	15.0	1.0	14.8	<.0001	
<i>seconds</i>	.16	.096	1.6	.11	
<i>mpg</i>	-.34	.022	-15.4	<.0001	

Table 8.6 Partial model regression

*RSquare* has dropped noticeably, from 92.7% to 88.6%, and the standard error increased by 20%, from .617 to .740. Is this a significant reduction in explanatory power? The *partial F* test will allow us to decide:

$$\text{Partial } F = F_{2,29} = \frac{(.926 - .886)/2}{(1 - .926)/(34 - 1 - 4)} = \frac{.040/2}{.074/29} = \frac{.020}{.0025} = 7.8, \text{ Significance } F_{2,29} = .002$$

The *partial F* of 7.8 is significant at a 99% level of confidence (*Significance*  $F_{2,29} = .002 < .01$ ). We conclude that *pounds(K)* and *liters* do add explanatory power to the model, significantly improving *RSquare*. They also reduce standard error, improving the precision of model forecasts. We cannot remove them. Jointly, with *MPG* and *seconds*, they drive emissions.

Our final multiple linear regression model of emissions is:

$$\text{emissions}_i = 9.0^a + .24^b \text{ seconds}_i - .23^a \text{ MPG}_i + .36^{a*} \text{ liters}_i + .43^{a*} \text{ pounds}(K)_i$$

(1.8)    (.09)                    (.03)                    (.20)                    (.24)

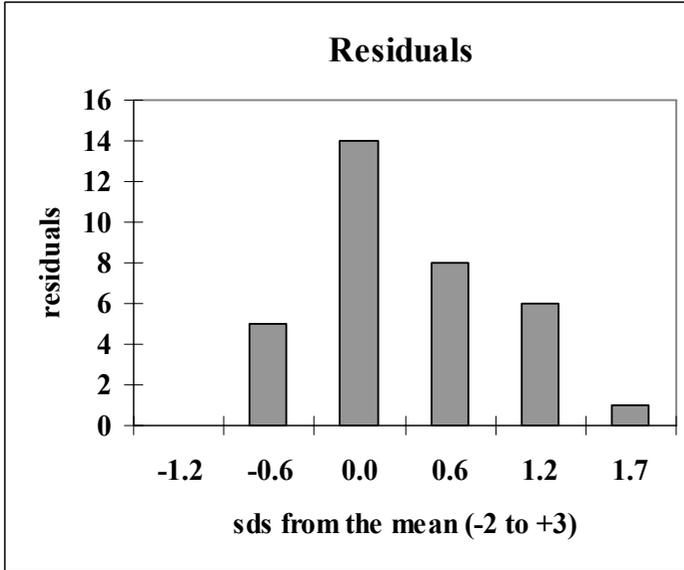
$$RSquare^a = .93$$

<sup>a</sup>Significant at a .0001 level or better.

<sup>b</sup>Significant at a .01 level or better.

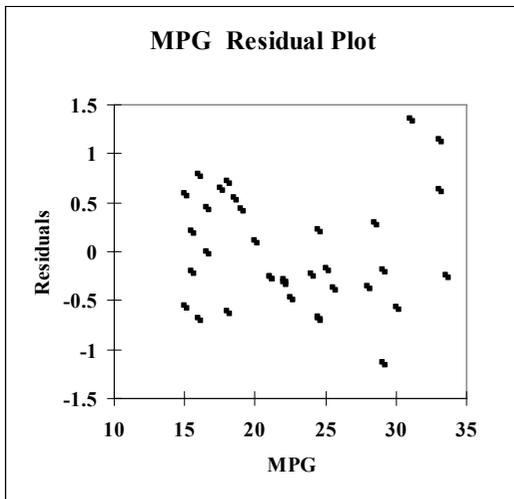
\* Jointly significant at .002 or better.

To determine whether or not our model satisfies the assumptions of linear regression, we look at the distribution of residuals, just as we do with a simple regression model.



In Figure 8.1, the residuals are approximately *Normal*, but do show positive skew, suggesting that rescaling to logarithms or an alternative nonlinear model would improve our fit.

Figure 8.1 Distribution of residuals



From Figure 8.2, we see that we achieve a slightly better fit for less fuel efficient cars, indicating a small degree of heteroskedasticity. Rescaling *MPG* or *emissions* or both in logarithms is likely to reduce this heteroskedasticity.

Figure 8.2 Residuals by MPG

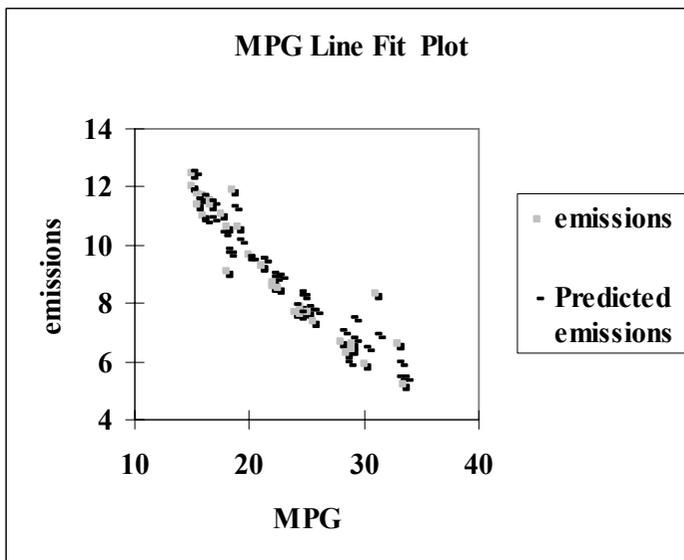
### 8.8 Sensitivity Analysis Quantifies the Marginal Impact Of Drivers

We want to compare the significant drivers to identify those which make the greatest difference. We will forecast emissions at average levels of each of the car characteristics. Then, we will compare forecasts at minimum and maximum levels of each, holding the other three at mean levels. The sensitivity analysis is summarized in Table 8.7, below:

<i>MPG</i>	<i>seconds to accelerate 0 topounds (K)</i>	<i>liters</i>	<i>expected emissions</i>	<i>improvement (reduction) in expected emissions</i>	
<b>15</b>	9	3.5	4.1	10.7	
<b>33.5</b>	9	3.5	4.1	6.5	<b>4.2</b>
22.6	<b>11.9</b>	3.5	4.1	9.7	
22.6	<b>6.7</b>	3.5	4.1	8.4	1.2
22.6	9	<b>6</b>	4.1	9.9	
22.6	9	<b>1.5</b>	4.1	8.3	1.6
22.6	9	3.5	<b>5.9</b>	9.8	
22.6	9	3.5	<b>2.5</b>	8.3	1.5

Table 8.7 Emissions response to car characteristics

**MPG.** Within a representative range of values for each of the car characteristics, fuel economy makes the largest difference in emissions, shown in Figure 8.3. Improving fuel economy by 19 MPG is associated with an expected reduction in emissions of 4.2 tons per year.



This is a large improvement, though not enough alone to meet the 5.0 tons per year goal. Fuel economy improvements will need to be made in conjunction with improvements in one or more of the other car characteristics.

Figure 8.3 Emissions by MPG

Our linear model suggests that improving average fuel economy by 5 MPG, from 25 to 30, would produce an expected average improvement in emissions of about one ton (.80 to 1.50 tons) per year, assuming other car characteristics were at mean levels, which is shown in Figure 8.4

$$\begin{aligned} \Delta MPG [b_{MPG} - 2s_{b_{MPG}}] &\leq \Delta MPG \beta_{MPG} \leq \Delta MPG [b_{MPG} + 2s_{b_{MPG}}] \\ (30 - 25)[- .23 - 2(.034)] &\leq (30 - 25)\beta_{emissions} \leq (30 - 25)[- .23 + 2(.034)] \\ (5)(-.30) &\leq (5)\beta_{emissions} \leq (5)(-.16) \\ -1.50 &\leq (5)\beta_{emissions} \leq -.90 \end{aligned}$$

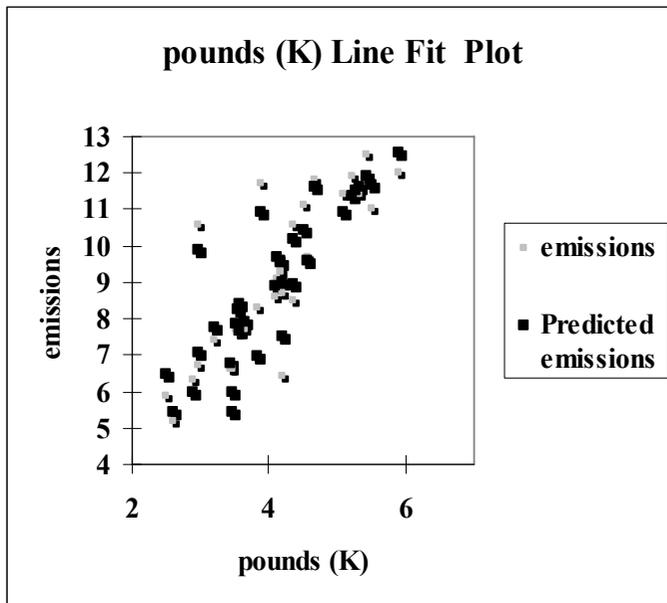
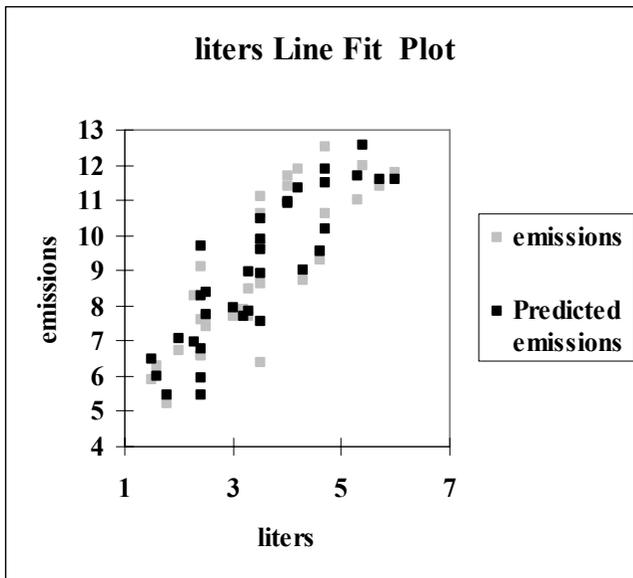


Figure 8.4 Predicted and actual emissions by pounds



**Pounds(K) and Liters.** Reducing car weight by 4,500 pounds or reducing engine size by 3.5 liters improves expected emissions by 1.5 to 1.6 tons per year, which is illustrated in Figure 8.5.

Figure 8.5 Predicted and actual emissions by liters

Even the combination of a lighter car with a smaller engine is not enough to reach the emissions goal of five tons per year. In combination with fuel economy improvements, either car weight or engine size improvements could make the goal attainable.

**Seconds.** Improving car responsiveness could improve expected emissions by more than a ton. Combined with any of the other car characteristics, responsiveness could help Sakura achieve their emissions goal.

Our model provides clear indications for the new product development team. To improve emissions, they will need to design more responsive, lighter-weight cars with smaller engines and superior fuel economy.

The Quantitative Analysis Director summarized model results in the following memo to Sakura Management:

## MEMO

**Re: Light, responsive, fuel efficient cars with smaller engines are cleanest**  
**To: Sakura Product Development Director**  
**From: Benjamin Nowak, Quantitative Analysis Director**  
**Date: June 2007**

Lighter, more responsive, fuel efficient cars with smaller engines are cleanest. Improvements in gas mileage and responsiveness, with reductions in weight or engine size will allow Sakura to achieve the emissions target of five tons per year.

A regression model of emissions was built from a representative sample of 34 diverse car models, considering fuel economy, acceleration, engine size and car size.

**Model results.** Differences in fuel economy, weight, engine size, and acceleration account for 93% of the variation in car emissions. Forecasts from these car characteristics are expected to be no further than 1.2 tons from actual average emissions for a particular car profile.

Fuel economy is the most powerful driver of emissions. Increasing gas mileage by five MPG is expected to reduce annual emissions by .8 to 1.5 tons.

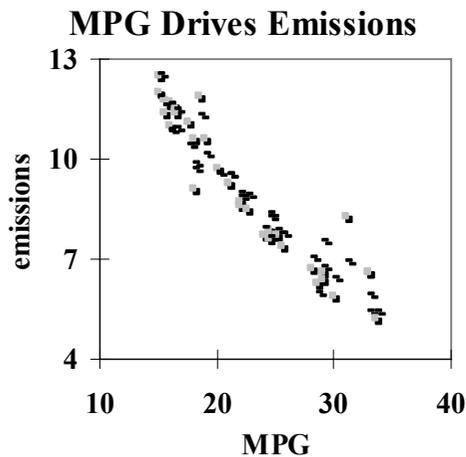
A one ton reduction in weight is expected to improve reduce emissions by as much as 1.8 tons.

Reducing engine size by three liters produces an expected reduction in emissions of as much as 2.3 tons.

Reducing acceleration from 0 to 60 by four seconds would improve emissions by .2 to 1.7 tons.

**Conclusions.** Fortunately, cleaner cars are also more fuel efficient and more responsive. This will allow Sakura to design cleaner models without sacrificing responsiveness. Improvements in fuel economy and responsiveness, with reductions in weight or engine size will enable Sakura to meet the emissions target of five tons per year.

Model results assume existing engine technology. With development of cleaner, more fuel efficient, responsive technologies, even lower emissions could possibly be achieved.



$$\text{emissions}_i = 9.0^a + .24^b \text{ seconds}_i - .23^a \text{ MPG}_i + .36^{a,*} \text{ liters}_i + .43^{a,*} \text{ pounds}(K)$$

*RSquare: .93<sup>a</sup>*

<sup>a</sup>Significant at .01

\*Jointly significant at .02

## 8.9 Model Building Begins With Logic and Considers Multicollinearity

Novice model builders sometimes mistakenly think that the computer can choose those variables which belong in a model. Computers have no experience making decisions and can never replace decision makers' logic. (Have you ever tried holding a conversation with a computer?) The first step in superior model building is to use your head. Use logic and experience to identify independent variables which ought to influence the performance variable which you are interested in explaining and forecasting. Both your height and GDP increased over the past ten years. Given data on your annual height and annual GDP, the computer could churn out a significant parameter estimate relating variation in your height to variation in GDP (or variation in GDP to variation in your height). Decision makers must use their logic and experience to select model variables. Software will quantify and calibrate the influences we know ought to exist.

It is a multicollinear world. Sets of variables together jointly influence performance. Using ratios of collinear predictors reduces multicollinearity. *Partial F tests* help us eliminate redundancies to more accurately explain performance and forecast. *Partial F* allows us to test the significance of reductions in *RSquare* that occur whenever we remove variables.

From the logically sound set of variables, pruned to eliminate redundancies and reduce multicollinearity, we have a solid base for superior model building. To this we will consider adding variables to account for seasonality or cyclicity in time series and the use of indicators to build in influences of segment differences, structural shifts and shocks in Chapter 10. In Chapter 11, we will explore alternative nonlinear models for situations where response is not constant, but where the rate of change model doesn't fit perfectly.

### Excel 8.1 Build and fit a multiple linear regression model

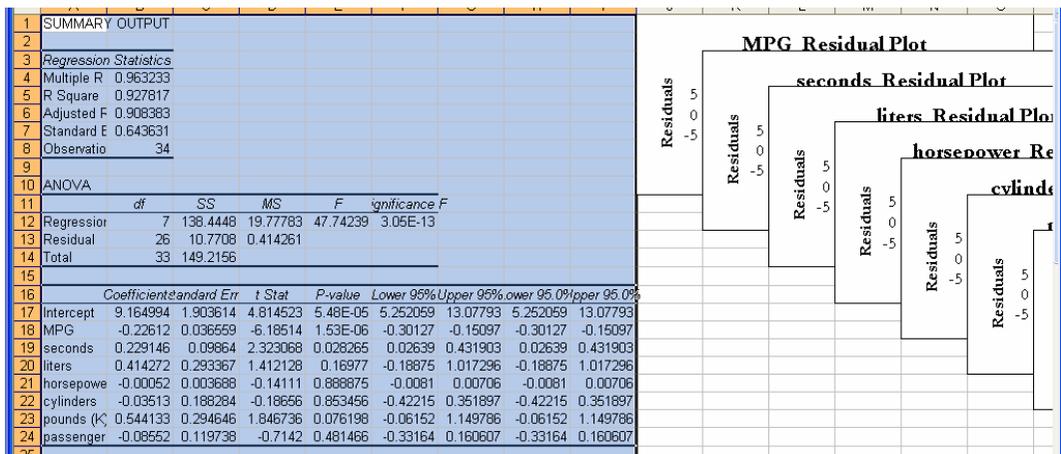
*Sakura Motors Quest for a Clean Car.* We will assist Sakura Motors in their quest for a less polluting car model, using data from [bea.gov](http://bea.gov) and [consumerreports.org](http://consumerreports.org), which together provide information on individual car models.

The dataset, **Excel 8.1 Sakura Motors.xls** contains data on 35 car models, representing U.S., European, and Asian manufacturers and a variety of sizes and styles.

Management is unsure which car characteristics influence *emissions*, but they suspect that fuel economy, *MPG*, acceleration capability, measured as *seconds* to accelerate from 0 to 60 mph, engine size, *cylinders*, *liters*, and *horsepower*, car *passenger* capacity, and weight in *pounds (K)* may be significant influences. Smaller, lighter models with smaller, less powerful engines are expected to be cleanest. We will fit a multiple linear regression model of these influences on *emissions*.

Open the dataset and run regression with the dependent variable *emissions* **C1:C35** in **Input Y Range** and the independent variables, *MPG*, *seconds*, *cylinders*, *liters*, *horsepower*, *passengers*, and *pounds* in **D1:J35** in the **Input X Range**.

Choose **Input: Labels and Residuals: Residuals, Residual Plots and Line Fit Plots, OK:**



**Multicollinearity symptoms.** While the model is significant (*Significance F* <.0001), only two of the car characteristics are significant (*p value* <.05). We are not certain that *pounds(K)*, *liters*, *cylinders*, *passengers*, and *horsepower* are influential, since their *p values* >.05. *Horsepower*, *cylinders* and *passengers* have “incorrect” negative signs. Cars with greater horsepower, more cylinders, and more passenger space ought to be bigger polluters. Together, the lack of significance of seemingly important predictors and the three sign reversals signal multicollinearity.

We will look at the correlations to confirm suspicions that *liters*, *horsepower* and *cylinders* are correlated (and together reflect car power) and that *pounds(K)* and *passengers* are correlated (and together reflect car size). This may allow us to eliminate two of the power variables and one of the size variables to reduce multicollinearity.

Run correlations between the car characteristics in **D1:J35**:

	MPG	seconds	liters	horsepower	cylinders	pounds (K)	passengers
MPG	1						
seconds	-0.04901	1					
liters	-0.81008	-0.17065	1				
horsepower	-0.52917	-0.36347	0.762617	1			
cylinders	-0.74357	-0.18889	0.924148	0.770781	1		
pounds (K)	-0.76895	-0.01287	0.835189	0.71753	0.807687	1	
passenger	-0.52581	-0.04871	0.592526	0.545275	0.602062	0.702565	1

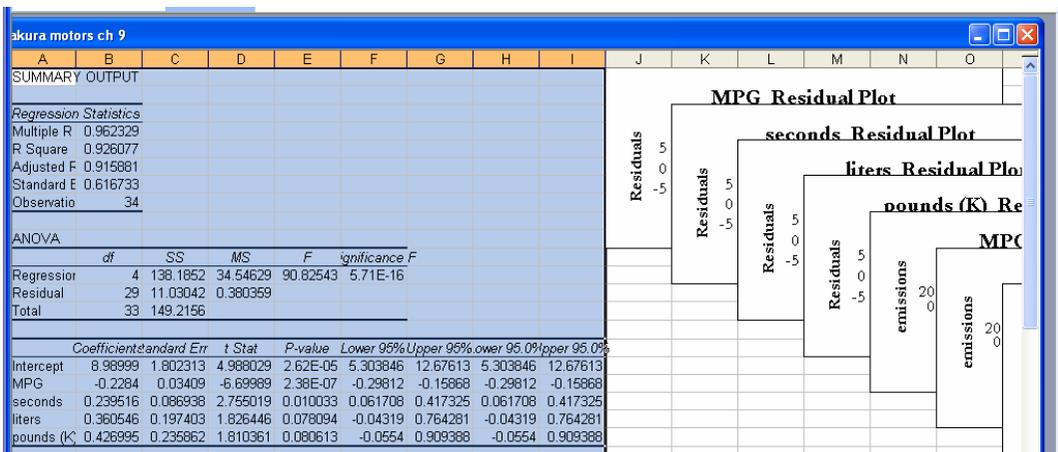
Eliminating two of the three measures of power will reduce multicollinearity. We will also eliminate one of the two measures of size to reduce multicollinearity.

**Use *Partial F* to test significance of contribution to *RSquare*.** We will eliminate characteristics that appear to add little explanatory power. This does not mean that they are not important. More likely, they are closely related to other important characteristics and contribute redundant information.

In the *Sakura Motors* sheet, rearrange the columns so that the variables that we want to keep in the model, *MPG*, *seconds*, *liters*, and *pounds(K)* are adjacent to each other in columns **D** through **G** and follow *emissions* in column **C**.

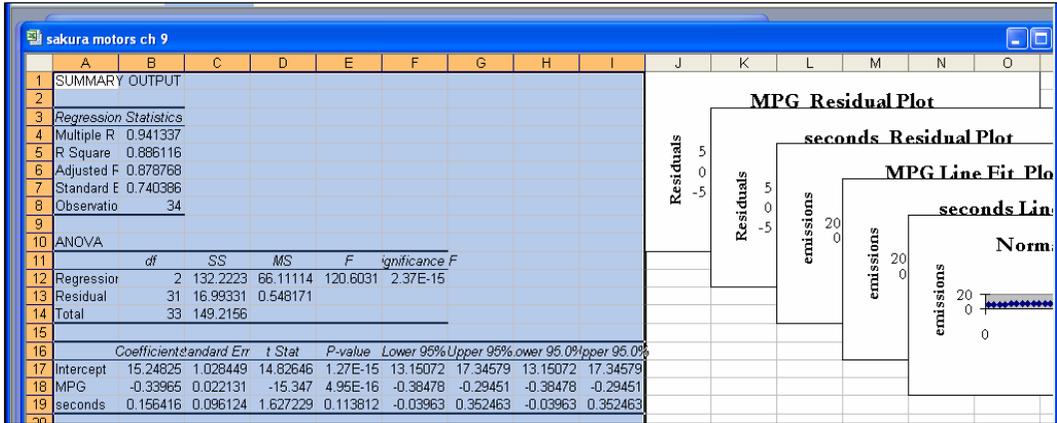
To make the four drivers adjacent, select and cut the *pounds(K)* column **I**, then use shortcuts to paste into column **G**: select **G**, **Alt HIE**.

Run the partial model regression, changing **Input X Range** to **\$D\$1:\$G\$35**.



We will eliminate *liters* and *pounds(K)* to see if they are redundant.

Re-run the regression, changing **Input X Range** to **D1:E35**:

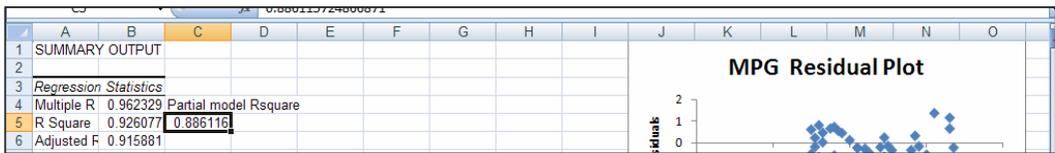


Liters and Pounds(K) contributed unique explanatory power to the model.

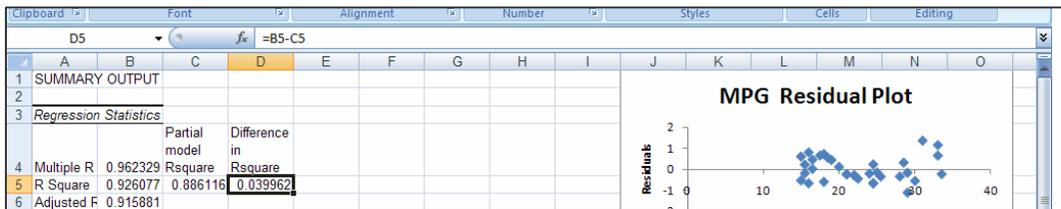
*Partial F* compares reduction in *RSquare*, per variable removed, to unexplained variation, divided by the Residual degrees of freedom in the larger model. Using *Partial F*, we assess the joint significance of the variables removed by focusing on reduction in explanatory power following their removal.

Enter the label *partial model RSquare* in **C4** of the larger model output sheet.

Copy *RSquare* in **B5** from the model (with only *MPG* and *seconds*) and paste it into the larger model output sheet in **C5**.



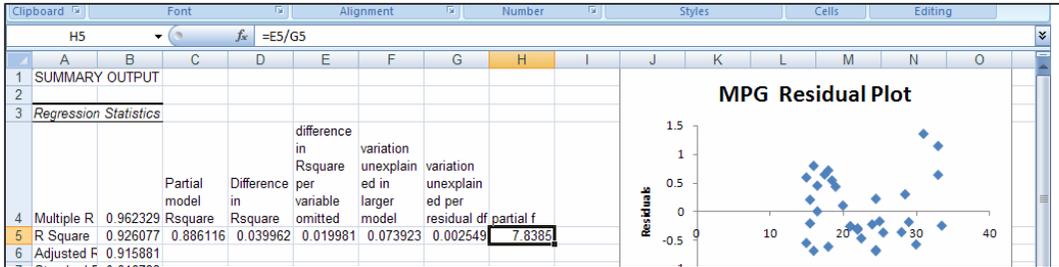
In **D4** enter the label *difference in RSquare*, and in **D5**, find the change in *RSquare* by entering **=B5-C5 [Enter]**.





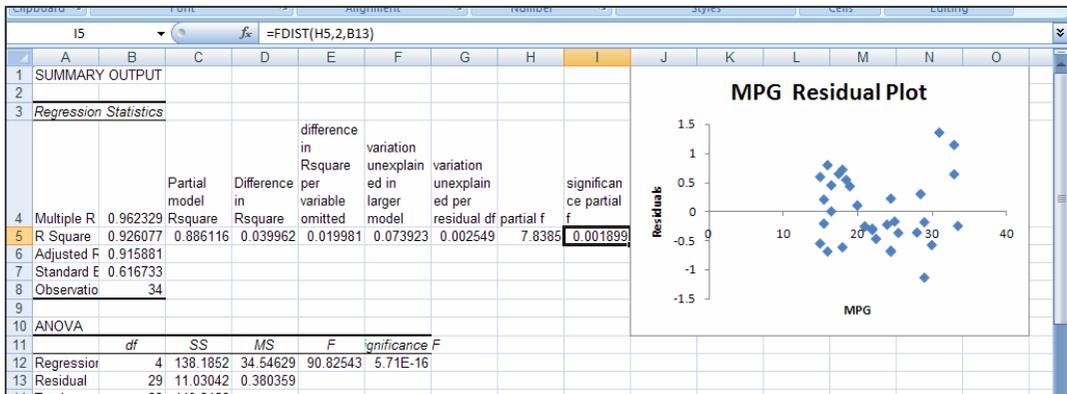
Enter *Partial F* in **H4** and in **H5** enter the formula for *partial F* =E5/G5[Enter], which is

$$Partial F_{2,29} = \frac{(.926 - .886)/2}{(1 - .926)/(34 - 1 - 4)} = \frac{.040/2}{.074/29} = \frac{.020}{.0025} = 7.8,$$



To find the level of significance of this *F* value, with 2 (variables omitted) and 29 (*residual df* in the larger model) degrees of freedom, use the Excel FDIST(F,df) function.

Enter *significance Partial F* in **I4** and in **I5** enter =FDIST(H5,2,B13) [Enter]:

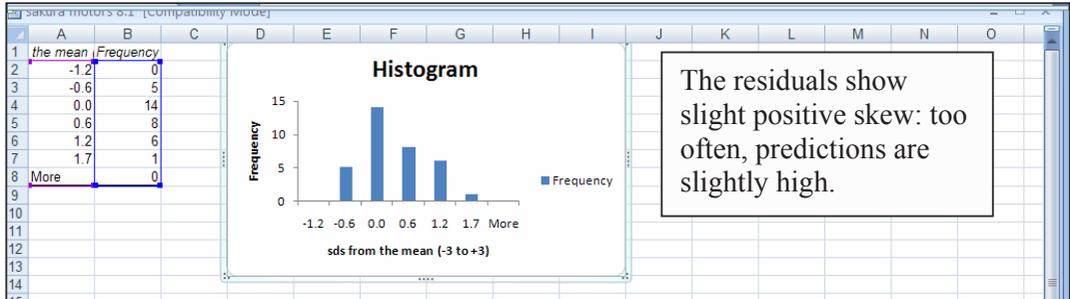


Read the significance level from the cell, .0019 in this case:

This is a very small probability. It is unlikely that we would observe this difference in *RSquare* if *liters* and *pounds(K)* were contributing redundant information. They will remain in the model, since, from the *Partial F test*, we conclude that they are jointly significant at a level of significance less than .0019.

**Look at residuals to check model assumptions.** We want to be sure that the model residuals are free of patterns and *Normally* distributed. Excel gives us the residuals (predicted minus actual) in the regression output sheet.

Make a histogram of the residuals in **D27:F33**.



### Excel 8.2 Use sensitivity analysis to compare the marginal impacts of drivers

For sensitivity analysis, we will need to identify a “low” and a “high” value for each of the four predictors, the minimum and maximum. For each, we will compare predictions given low and high values to find the range of response. To study marginal response to a predictor, we vary only that predictor and set the remaining predictors at their mean values.

To find the minimum, maximum, and mean values for each of the four predictors, use the Excel functions **MAX(array)**, **AVERAGE(array)** and **MIN(array)**.

Enter labels *maximum*, *mean* and *minimum* in **B37:B39**.

- In **D37**, enter **=MAX(D2:D35)[Enter]**.
- In **D38**, enter **=AVERAGE(D2:D35)[Enter]**.
- In **D39**, enter **=MIN(D2:D35)[Enter]**.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (k)	Coefficient	predicted emissions	horsepower	cylinders	passengers	
29	toyota	land cruiser	12.5	15	9.7	4.7	5.435		11.90262	275	8	8	
30	toyota	sequoia	11.5	16.5	9.7	4.7	5.28		11.49384	273	8	8	
31	toyota	sienna	8.5	22.5	8.6	3.3	4.365		8.964516	230	6	8	
32	toyota	tundra	11.4	16.5	8.8	4	5.095		10.9469	245	6	6	
33	vw	toureg	11.9	18.5	11.9	4.2	5.21		11.35381	240	8	5	
34	honda	accord hybrid	6.6	33	6.9	2.4	3.475		5.454643	255	4	5	
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2		7.529488	268	6	4	
36		maximum		33.5									
38		mean		22.6029412									
39		minimum		15									

Select **D37:D39**, grab and drag through **G37:G39**:

Our benchmark, or “typical” car will achieve 22.6 *MPG*, accelerate from 0 to 60 in 9 *seconds* with a 3.5 *liter* engine, and it will weigh 4.1 thousand pounds.

Within the existing range of car designs, a car could achieve the “best” gas mileage of 33.5 *MPG*, or it could have the worst gas mileage of 15 *MPG*. Comparing the difference in expected *Emissions* when all but one driver are at mean levels allows us to isolate the impact of that driver. This will tell us how relatively important each driver is, and which have the greater potential to reduce *Emissions*.

D37													=MAX(D1:D35)	
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	predicted	emissions	coefficients	
29	toyota	land cruiser	12.5	15	9.7	4.7	5.435		11.90262	275	8	8		
30	toyota	sequoia	11.5	16.5	9.7	4.7	5.28		11.49384	273	8	8		
31	toyota	sienna	8.5	22.5	8.6	3.3	4.365		8.964516	230	6	8		
32	toyota	tundra	11.4	16.5	8.8	4	5.095		10.9469	245	6	6		
33	vw	toureg	11.9	18.5	11.9	4.2	5.21		11.35381	240	8	5		
34	honda	accord hybrid	6.6	33	6.9	2.4	3.475		5.454643	255	4	5		
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2		7.529488	268	6	4		
36														
37		maximum		33.5	11.9	6	5.9							
38		mean		22.6029412	9.00294118	3.49117647	4.0675							
39		minimum		15	6.7	1.5	2.485							

In C40 through C42, enter labels for cars with *best*, *typical*, and *worst MPG*.  
 In row 40, enter the maximum *MPG* and sample mean values for *seconds*, *pounds* and *liters*.

In row 41, enter the means for all four predictors.

In row 42, enter the minimum *MPG* and sample mean values for *seconds*, *pounds* and *liters*:

sakura motors o.r. (compatibility mode)													
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	predicted	emissions	coefficients
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4			
36													
37			min	15	6.7	1.5	2.5						
38			average	22.6	9.0	3.5	4.1						
39			max	33.5	11.9	6	5.9						
40			best MPG	33.5	9	3.5	4.1						
41			typical	23	9	3.5	4.1						
42			worst MPG	15	9	3.5	4.1						

In rows 43 through 45, enter labels for *worst*, *typical* and *best acceleration*, maximum, average, and minimum values for *seconds*, and average values for the other three characteristics.

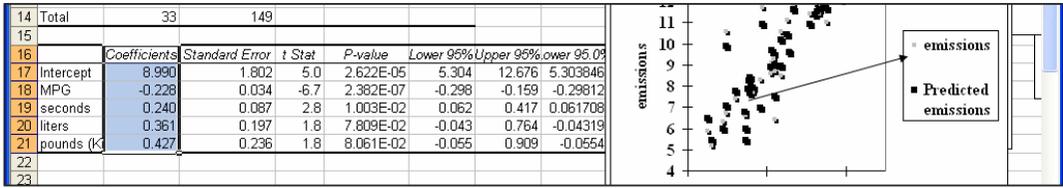
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	predicted	emissions
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4			
36													
37			min	15	6.7	1.5	2.5						
38			average	22.6	9.0	3.5	4.1						
39			max	33.5	11.9	6	5.9						
40			best MPG	33.5	9	3.5	4.1						
41			typical	23	9	3.5	4.1						
42			worst MPG	15	9	3.5	4.1						
43			worst accel	23	11.9	3.5	4.1						
44			typical	23	9	3.5	4.1						
45			best accel	23	6.7	3.5	4.1						

In rows 46 through 48, enter labels for *largest*, *typical* and *smallest engine*, maximum, average, and minimum values for *liters*, and average values for the other characteristics.

In rows 49 through 51, enter labels for *heaviest*, *typical* and *lightest*, maximum, average, and minimum values for *pounds(K)* with average values for the other characteristics:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	predicted	emissions
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4			
36													
37			min	15	6.7	1.5	2.5						
38			average	22.6	9.0	3.5	4.1						
39			max	33.5	11.9	6	5.9						
40			best MPG	33.5	9	3.5	4.1						
41			typical	23	9	3.5	4.1						
42			worst MPG	15	9	3.5	4.1						
43			worst accel	23	11.9	3.5	4.1						
44			typical	23	9	3.5	4.1						
45			best accel	23	6.7	3.5	4.1						
46			largest engine	23	9	6	4.1						
47			typical	23	9	3.5	4.1						
48			smallest engine	23	9	1.5	4.1						
49			heaviest	23	9	3.5	5.9						
50			typical	23	9	3.5	4.1						
51			lightest	23	9	3.5	2.5						

To find *Emissions* predicted by the model for each hypothetical car, copy the coefficients from the regression output sheet **B16:B21**, and paste into column **K**:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients		
2	acura	mdx	9.7	20	8.2	3.5	4.555	265	6	7	8.98999		
3	acura	tl	7.9	24.5	6.7	3.2	3.565	270	6	5	-0.2284		
4	bmw	5series	7.7	24	7.4	3	3.65	215	6	5	0.239516		
5	chevrolet	trailblazer	11.8	15.5	8.3	6	4.66	291	8	5	0.360546		
6	dodge	durango	11.4	15.5	7.6	5.7	5.335	210	8	7	0.426995		
7	ford	crown victoria	9.3	21	8	4.6	4.18	224	8	6			

Add a label *predicted emissions* in L1.

Enter the regression equation formula

$$emi\hat{s}ions_i = b_1 + b_2 * MPG_i + b_3 * seconds_i + b_4 * liters_i + b_5 * pounds(K)_i$$

using the coefficient estimates  $b_1$  through  $b_5$  in column K.

In L2 enter = \$K\$2+\$K\$3\*D2+\$K\$4 \*E2+\$K\$5\*F2+\$K\$6\*G2 [Enter]:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	predicted emissions	differ
2	acura	mdx	9.7	20	8.2	3.5	4.555	265	6	7	8.98999	9.6	
3	acura	tl	7.9	24.5	6.7	3.2	3.565	270	6	5	-0.2284		
4	bmw	5series	7.7	24	7.4	3	3.65	215	6	5	0.239516		
5	chevrolet	trailblazer	11.8	15.5	8.3	6	4.66	291	8	5	0.360546		
6	dodge	durango	11.4	15.5	7.6	5.7	5.335	210	8	7	0.426995		

Drag the lower right corner of the new cell in the new *predicted emissions* column L through row 47 to add predictions for the twelve hypothetical cars, then reduce decimals:

1	Manufacturer	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	emissions	predicted	difference
33	vw	toureg	11.9	18.5	11.9	4.2	5.21	240	8	5		11.4	11.4	
34	honda	accord hybrid	6.6	33	6.9	2.4	3.475	255	4	5		5.5	5.5	
35	lexus	RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4		7.5	7.5	
36												9.0	9.0	
37		max		33.5	11.9	6	5.9					8.9	8.9	
38		average		22.6	9.0	3.5	4.1					9.0	9.0	
39		min		15	6.7	1.5	2.5					8.8	8.8	
40		best MPG		33.5	9	3.5	4.1					6.5	6.5	
41		typical		23	9	3.5	4.1					8.9	8.9	
42		worst MPG		15	9	3.5	4.1					10.7	10.7	
43		worst accel		23	11.9	3.5	4.1					9.6	9.6	
44		typical		23	9	3.5	4.1					8.9	8.9	
45		best accel		23	6.7	3.5	4.1					8.4	8.4	
46		largest engine		23	9	6	4.1					9.8	9.8	
47		typical		23	9	3.5	4.1					8.9	8.9	
48		smallest engine		23	9	1.5	4.1					8.2	8.2	
49		heaviest		23	9	3.5	5.9					9.7	9.7	
50		typical		23	9	3.5	4.1					8.9	8.9	
51		lightest		23	9	3.5	2.5					8.2	8.2	

The difference in expected *emissions* given maximum and minimum *MPG* suggests the potential difference that *MPG* could make.

In **M** enter the label *difference* and in **M42**, enter **=L42-L40** [Enter].

	B	C	D	E	F	G	H	I	J	K	L	M	
1	Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	emissions	predicted	difference
33	toureg	11.9	18.5	11.9	4.2	5.21	240	8	5		11.4		
34	accord hybrid	6.6	33	6.9	2.4	3.475	255	4	5		5.5		
35	RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4		7.5		
36											9.0		
37	max		33.5	11.9	6	5.9					8.9		
38	average		22.6	9.0	3.5	4.1					9.0		
39	min		15	6.7	1.5	2.5					8.8		
40	best MPG		33.5	9	3.5	4.1					6.5		
41	typical		23	9	3.5	4.1					8.9		
42	worst MPG		15	9	3.5	4.1					10.7	4.2	

Improving fuel economy of a typical car, from 15 to 33.5 *MPG*, is expected to reduce *emissions* by  $(10.7-6.5=)$  4.2 tons per year.

To see the potential marginal difference that each of the other characteristics makes,

In **M45**, enter **=L43-L45** [Enter].

In **M48**, enter **=L46-L48** [Enter].

In **M51**, enter **=L49-L51** [Enter]:

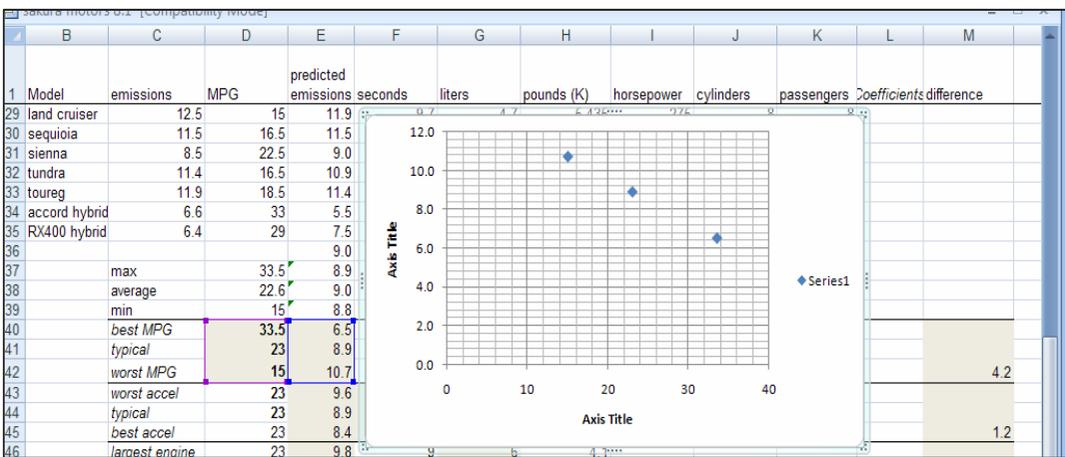
Model	emissions	MPG	seconds	liters	pounds (K)	horsepower	cylinders	passengers	Coefficients	predicted emissions	difference
land cruiser	12.5	15	9.7	4.7	5.435	275	8	8		11.9	
sequoia	11.5	16.5	9.7	4.7	5.28	273	8	8		11.5	
sienna	8.5	22.5	8.6	3.3	4.365	230	6	8		9.0	
tundra	11.4	16.5	8.8	4	5.095	245	6	6		10.9	
toureg	11.9	18.5	11.9	4.2	5.21	240	8	5		11.4	
accord hybrid	6.6	33	6.9	2.4	3.475	255	4	5		5.5	
RX400 hybrid	6.4	29	8.8	3.5	4.2	268	6	4		7.5	
max		33.5	11.9	6	5.9					9.0	
average		22.6	9.0	3.5	4.1					8.9	
min		15	6.7	1.5	2.5					9.0	
best MPG		33.5	9	3.5	4.1					8.8	
typical		23	9	3.5	4.1					6.5	
worst MPG		15	9	3.5	4.1					8.9	
worst accel		23	11.9	3.5	4.1					10.7	4.2
typical		23	9	3.5	4.1					9.6	
best accel		23	6.7	3.5	4.1					8.9	
largest engine		23	9	6	4.1					8.4	1.2
typical		23	9	3.5	4.1					9.8	
smallest engine		23	9	1.5	4.1					8.9	
heaviest		23	9	3.5	5.9					8.2	1.6
typical		23	9	3.5	4.1					9.7	
lightest		23	9	3.5	2.5					8.9	
										8.2	1.5

**Scatterplots of marginal response.** To see the impact of each driver, plot actual and predicted *emissions* of hypotheticals.

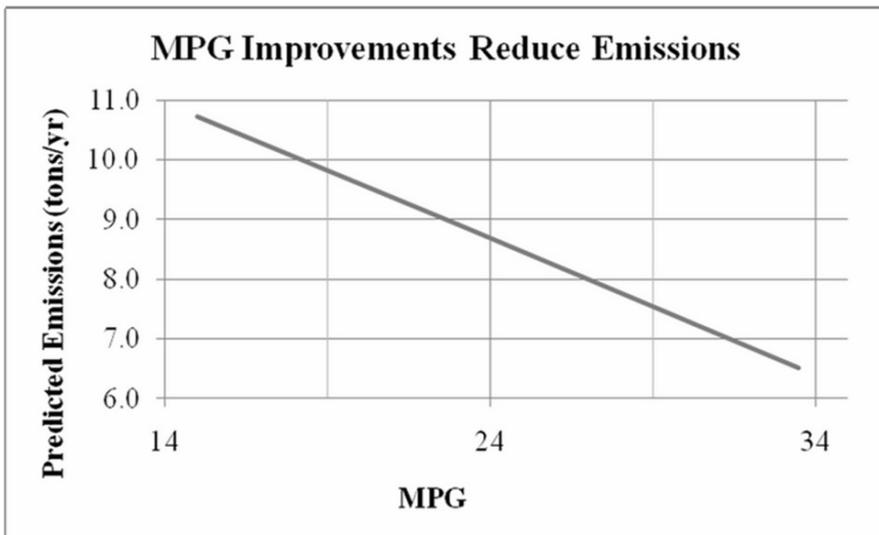
We'll focus on *MPG*.

Rearrange columns so that *MPG*, *Emissions*, and *Predicted emissions* are adjacent: Cut *predicted emissions* in column L and paste into column E.

Select *MPG* and *predicted emissions* columns D and E of the three new rows 40:42 which include the three hypothetical gas mileage levels, then insert a scatterplot, and choose **Layout 10** to see vertical and horizontal reference lines.



Adjust scales for both axes, choose, font and font sizes, and add chart and axes titles:



Improving fuel economy by 19 MPG (from 15 to 34) is expected to reduce emissions 4.2 tons per year (from 10.7 to 6.5).

## Lab Practice 8

### Drivers of Premie Diaper Fit Importance

Procter & Gamble managers were encouraged by concept test results of their Pampers Preemies. Test results revealed that

- superior diaper fit, the benefit which differentiates Pampers Preemies, is an important attribute to premie moms, and
- the most promising target market is unique demographically.

Product manager Deb Henretta wants to know which key demographics are driving the importance of fit. Use the data in **Lab Practice 8 Diaper Fit Drivers.xls** to build a multiple regression model which will provide this information.

Do the set of demographics, *age*, *income*, *family size*, and *number of other kids*, together drive *fit importance*? Y or N

Your evidence will be the significance level of your model *F* test. *Significance F*: \_\_\_\_\_

Based on concept test sample evidence, which particular demographics drive *fit importance*?

	<i>age</i>	<i>income</i>	<i>Family size</i>	<i>Other kids</i>
Significant?	Y or N	Y or N	Y or N	Y or N
Evidence ( <i>p value</i> )				

Which coefficients have the “wrong” sign?

	<i>age</i>	<i>income</i>	<i>Family size</i>	<i>Other kids</i>
Unexpected sign	Y or N	Y or N	Y or N	Y or N

Is it possible that the demographics which seem to be insignificant really matter? Y or N

Find the correlations between each pair of demographic variables and identify those which are highly correlated ( $|r| > .7$ ):

	$ r  > .7?$		$ r  > .7?$
<i>Age, income</i>		<i>Income, family size</i>	
<i>Age, family size</i>		<i>Income, other kids</i>	
<i>Age, other kids</i>		<i>Family size, other kids</i>	

Choose one of the two most strongly correlated demographics to represent the other and re-run your regression.

Is your model explanatory power just as good without the omitted demographic? Y or N

Your evidence is the change in *RSquare*:

Full model *RSquare*: \_\_\_\_ Partial model *RSquare*: \_\_\_\_ Change in *RSquare*: \_\_\_\_

Which demographics drive *fit importance*? (Cross out the variable that you omitted in your partial model.)

	<i>age</i>	<i>income</i>	<i>Family size</i>	<i>Other kids</i>
Significant?	Y or N	Y or N	Y or N	Y or N
Evidence ( <i>p value</i> )				

Which coefficients have the “wrong” sign? (Cross out the variable that you omitted in your partial model.)

	<i>age</i>	<i>income</i>	<i>Family size</i>	<i>Other kids</i>
Unexpected sign	Y or N	Y or N	Y or N	Y or N

Can Procter & Gamble managers safely assume that the demographic variable which was omitted is not a driver of *fit importance* and can be ignored? Y or N

Make a histogram of your residuals. Are the residuals approximately *Normal*? Y or N

Use the partial model coefficients to make *predicted fit importance* with your model regression equation.

Find the *minimum*, *mean* and *maximum* levels for each demographic variable in your model.

Add hypothetical preemie moms to the dataset and find the difference that each demographic makes in driving *fit importance* when other demographics are accounted for. (Cross out the demographic omitted.)

Hypothetical preemie mom	<i>Predicted fit importance</i>	<i>Difference</i>
<i>Oldest</i> with other demos at average levels		
<i>Youngest</i> with other demos at average levels		
<i>Highest income</i> with other demos at averages		
<i>Lowest income</i> with other demos at averages		
<i>Largest family size</i> with other demos at averages		
<i>Smallest family size</i> with other demos at averages		
Most <i>other kids</i> with other demos at averages		
Fewest <i>other kids</i> with other demos at averages		

Differences in \_\_\_\_\_ make the most difference in *fit importance*.

Plot *fit importance* and *predicted fit importance* with the most important demographic driver to illustrate your result. Embed or attach your plot.

## Lab 8 Model Building with Multiple Regression

### Pricing Dell's Navigreat

Dell has experience selling GPS systems built by other firms and plans to introduce a Dell system, the Navigreat. They would like information that will help them set a price.

The Navigreat has

- an innovative, *highly portable* design, *weighing only 5 ounces*, with a *state-of-the-art display*
- a *3.5" screen*, neither large, nor small, relative to competitors.
- innovative technology which guarantees precise *routing time* estimates,

Dell executives believe that these features, *portability*, *weight*, *display quality*, *screen size*, and *routing time* precision, drive the price that customers are willing to pay for a GPS system.

Recent ratings by *Consumer Reports* provide data on the retail *price* of 18 competing brands, as well as

- *portability* (1 to 5 scale), *weight* (ounces), and *display quality* (1 to 5 scale),
- *screen size* (inches)
- *routing time precision* (1 to 5 scale),

These data are in **Lab 8 Dell Navigreat.xls**. Also in the file, in row 21, are the attributes and expected ratings of the Navigreat.

Build a multiple regression model of GPS system *price*, including the characteristics thought by management to be drivers of *price*.

**Regression results.** Is the model *RSquare* significantly greater than 0? Y N

Evidence: *Significance F*= \_\_\_\_\_

Which of the potential drivers have slopes significantly different from 0?

	<i>portability</i>	<i>weight</i>	<i>display</i>	<i>Screen size</i>	<i>Routing time</i>
Slope different from zero	Y or N	Y or N	Y or N	Y or N	Y or N
Evidence ( <i>p-value</i> )					

Which of the drivers have slopes of unexpected sign?

	<i>portability</i>	<i>weight</i>	<i>display</i>	<i>Screen size</i>	<i>Routing time</i>
Slope sign unexpected	Y or N	Y or N	Y or N	Y or N	Y or N

**Confirm suspected multicollinearity.** The GPS system physical design determines its *screen size*, *display quality*, *weight* and *portability*. Run correlations to see if these characteristics are highly correlated.

	Highly correlated ( $r_{x1,x2} > .5$ )
Portability, weight	Y or N
Portability, display	Y or N
Portability, screen size	Y or N
Weight, display	Y or N
Weight, screen size	Y or N
Display, screen size	Y or N

Choose one of the set of correlated characteristics to represent the set, eliminating the other potentially redundant characteristics, and re-run the regression.

Is this partial model *RSquare* significantly greater than 0? Y N

Evidence: *Significance F* = \_\_\_\_\_

Which of the potential drivers in this reduced model have slopes significantly different from 0? (Cross out characteristics that you excluded in this reduced model.)

	portability	weight	display	Screen size	Routing time
Slope different from zero	Y or N	Y or N	Y or N	Y or N	Y or N
Evidence ( <i>p value</i> )					

Which of the drivers have slopes of unexpected sign? (Cross out characteristics that you excluded in this partial model.)

	portability	weight	display	Screen size	Routing time
Slope sign unexpected	Y or N	Y or N	Y or N	Y or N	Y or N

Find *Partial F* to decide whether the partial model's explanatory power is significantly lower than in the full model.

Full model <i>RSquare</i> (1)	Partial model <i>RSquare</i> (2)	Change in <i>RSquare</i> (3) =(1)-(2)	Change per <i>g</i> predictors excluded (4) =(3)/ <i>g</i>	%variation unexplained by full model (5) =1-(1)	%variation unexplained per residual <i>dfs</i> (6) =(5)/(N-1-k)	<i>Partial F</i> (7) =(4)/(6)	<i>p value</i> with <i>g</i> and (N-1-k) <i>dfs</i>

Conclusion:

\_\_\_\_\_ partial model *RSquare* is significantly lower than full model *RSquare*, and potentially redundant variables are jointly significant and cannot be excluded

OR \_\_\_\_\_ partial model *RSquare* is not significantly lower than full model *RSquare*, excluded variables are redundant or unimportant, and can remain excluded.

Determine the improvement in predictive accuracy:

	Full model (1)	Reduced model (2)	Improvement in <i>margin of error</i> (3)=(2)-(1)
<i>Standard error</i>	\$	\$	
<i>Approximate margin of error in 95% predictions</i>	\$	\$	\$

**Assess fit.** Change the Line Fit chart type to scatterplot, adjust axes, and add chart and axis titles. Does the impact of *screen size* on *price* seem to be linear? Y or N

**Assess residuals.** Produce a residual histogram. Are residuals approximately *Normal*? Y or N

**Predict prices.** Copy the *coefficients* and paste into the Navigreat sheet, then use the regression equation to find *expected prices* for each of the GPS systems, including the Navigreat.

Copy the *standard error* and paste into the Navigreat sheet.

Find the *t* value for 95% prediction intervals with your model *residual degrees of freedom*.

Find the *lower and upper 95% prediction intervals* for each model, including the Navigreat.

Will Dell be able to charge a retail price of \$650 for the Navigreat? Y or N

**Sensitivity analysis:** Identify the most important driver of prices by comparing the differences in *expected prices* between four hypothetical GPS systems.

Add these four hypotheticals at the bottom of the file, then extend *expected price*, *lower* and *upper 95% prediction* bounds to include these.

<i>Screen size</i>	<i>Route time rating</i>	<i>Expected price</i>	<i>Difference due to</i>
Largest (5")	Average (4="Good")	\$	Screen size: \$
Smallest (3.4")	Average(4="Good")	\$	
Average (3.8")	Best (5="Excellent:)	\$	Route time rating: \$
Average (3.8")	Worst (2="poor")	\$	

If Dell wants to charge a retail price of \$650 for the Navigreat, what product design modification ought to be made? \_\_\_\_\_

## Assignment 8-1

### *Sakura Motor's Quest for Fuel Efficiency*

The new product development team at Sakura Motors has decided that the new car which they are designing will have superior gas mileage on the highway.

Use the data in **Assignment 8-1 Sakura Motors.xls** to build a model to help the team. Variables in the dataset include:

*MPGH<sub>wy</sub>*

*manufacturer's suggested retail base price*

*engine size (liters)*

*engine cylinders*

*engine horsepower*

*curb weight*

*acceleration in seconds to go from 0 to 60*

*percent of owners satisfied who would buy the model again*

- Use your logic to choose car characteristics which ought to influence highway gas mileage.
- Determine which car characteristics influence *highway gas mileage*. Use *partial F test(s)* to decide whether to remove apparently insignificant variables.
- With sensitivity analysis, find the relative importance of significant influences on *highway fuel economy*
- Find the car characteristic levels which could be expected to achieve **40 miles per gallon** in highway driving.  
(Sakura is not limited to existing designs.)

Write a **one-page single-spaced** memo presenting your model, sensitivity analysis and design recommendations.

- Present your final model in standard format
  - What is the margin of error of model forecasts of MPG?
  - What do the *95% confidence intervals* for the coefficient estimates tell us?
  - Do the coefficient estimates make sense?
- Discuss the relative importance of significant influences, including the expected difference in *fuel economy* that differences in each could be expected to make if other characteristics were held at mean values

- Illustrate your sensitivity analysis with plots of *95% prediction intervals* and *actual fuel economy* for the most important influence and **embed** this in your memo
- Comment on the assumption that MPG Hwy response is constant—are the relationships linear

## 9

# Model Building and Forecasting with Multicollinear Time Series

A regression model from time series data allows us to identify performance drivers and forecast performance given specific predictor values, just as regression models from cross sectional data do. When decision makers want to forecast *future* performance, a time series of past performance is used to identify drivers and fit a model. A time series model can be used to identify drivers whose variation over time is associated with later variation in performance over time.

Three differences in the model building process distinguish cross sectional and time series models:

- the use of lagged predictors,
- addition of trend, seasonality and cyclical variables, and
- the model validation process.

In time series models, the links between drivers and performance are stronger if changes in the drivers precede change in performance. Therefore, lagged predictor variables are often used. Time series models are built using predictor values from past periods to explain and forecast later performance. Figure 9.1 illustrates the differences in model building processes between cross sectional and time series models.

Most business performance variables and most economic indicators are cyclical. Economies cycle through expansion and recession, and performance in most businesses fluctuates following economic fluctuation. Business and economic variables are also often seasonal. We account for cyclical and seasonality by adding cyclical and seasonal predictors.

Before a time series model is used to forecast future performance, it is validated:

- the two most recent observations are excluded to fit the model,
- the model equation is used to forecast performance in those two most recent periods,
- model prediction intervals are compared with actual performance values in those two most recent periods, and if the prediction intervals contain actual performance values, this is evidence that the model has *predictive validity* and can be reliably used to forecast unknown performance in future periods.

## Model Building Process

Cross Sectional
Time Series

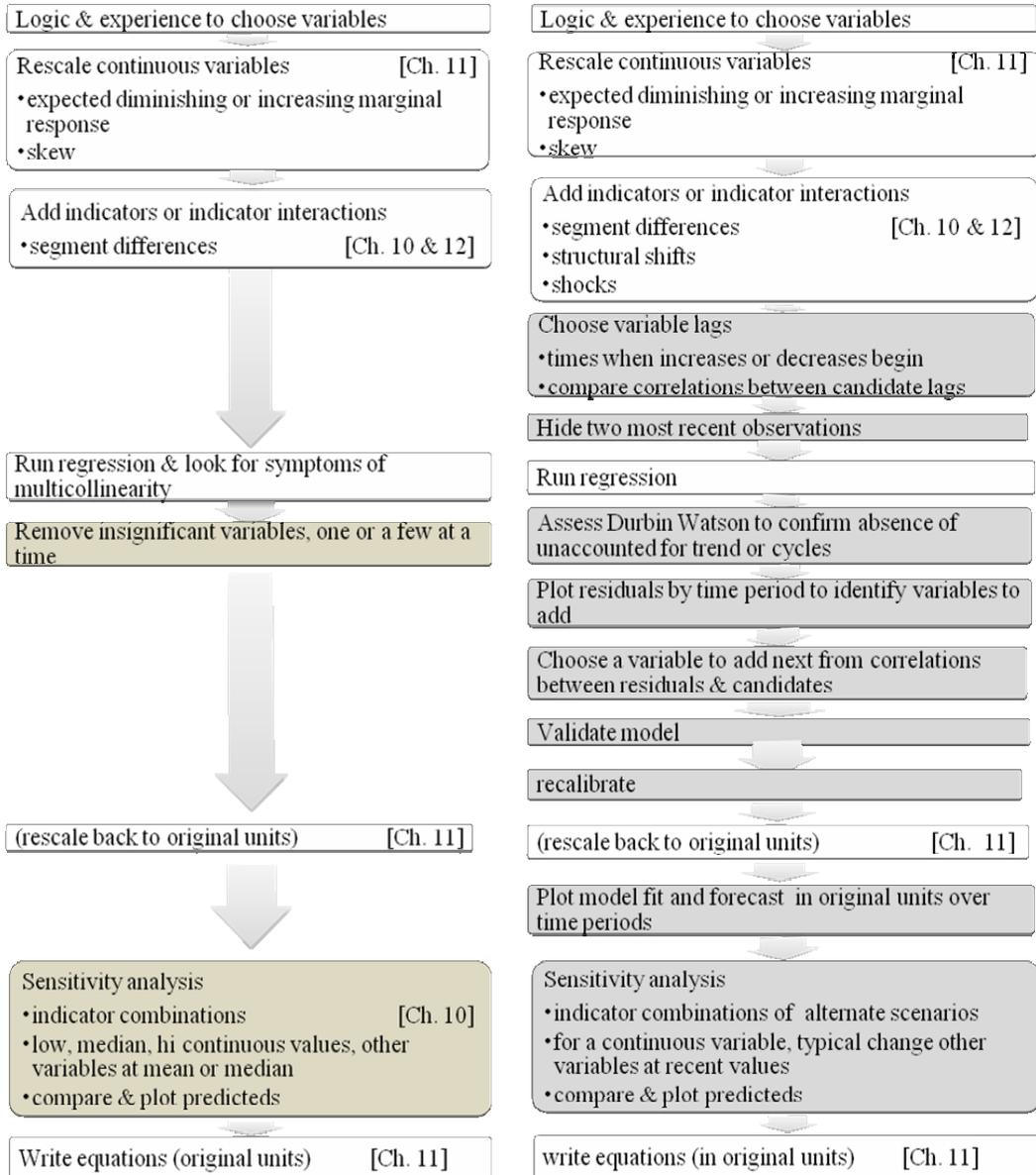


Figure 9.1 Model building processes with cross sectional and time series data

## 9.1 Time Series Models Include Decision Variables, External Forces, Leading Indicators, And Inertia

Most successful forecasting models logically assume that performance in a period,  $Y_t$ , depends upon

- decision variables under the management control,
- external forces, including
  - shocks such as 9/11,
  - market variables,
  - competitive variables,
- Inertia, from past performance
- Leading indicators of the economy, industry or the market
- Seasonality
- Cyclicalities

Ultimately, the multiple regression forecasting models that we build contain these components, which together account for variation in performance. In this chapter, we will introduce trend, inertia and leading indicator components of regression models built from time series.

Performance across time depends on decision variables and the economy. Decision variables, such as spending on advertising, sales effort and research and development tend to move together. In periods of prosperity, spending in all three areas may increase; in periods where performance is sluggish, spending in all three areas may be cut. Firm strategy guides resource allocation to the various firm functions. As a result, it is common for spending and investment variables to be correlated in time-series data.

Many economic indicators also move together across time. In times of economic prosperity, GDP is growing faster, consumer expectations increase and investments increase. Increasing wealth filters down from the economy to consumers and stock holders, where some proportion of gains are channeled back into consumption of investments.

It is common for decision variables, past performance, and leading indicators to be correlated in time-series data. This inherent correlation of performance drivers in time-series data makes logical choice of predictors a critical component of good model building.

It is also often more promising to build models by adding variables, one at a time, looking at residuals for indications of the most promising variables to add next. We will continue to explore multicollinearity in this chapter, including its consequences, diagnosis and alternate remedies.

*Example 9.1 Home Depot Revenues*<sup>1</sup>. Several Home Depot executives were concerned in late 2006 that revenues might slow following a sudden downturn in *New Home Sales*, a leading indicator of the housing market. Traditionally, Home Depot Revenues have grown following growth in *New Home Sales*, since builders and homeowners buy construction materials, flooring, and appliances at Home Depot.

Another group of Home Depot executives was optimistic, pointing to increasing growth in Home Depot Revenues. They believed that Home Depot customers were loyal and became customers for life, returning to purchase home improvement products, flooring and appliances.

## 9.2 Indicators of Economic Prosperity Lead Business Performance

To model the link between changes in a leading indicator and later performance, we could build a *leading indicator* model:

$$\text{revenues}(B)_t = b_0 + b_1 \text{NewHomeSales}(K)_{t-l}$$

where  $l$  denotes the length of lag, or delay from change in new home sales to change in revenues.

## 9.3 Inertia from Loyal Customers Drives Performance

Past performance is often a good predictor of future performance. Performance exhibits *inertia*, as prior patterns tend to be repeated. One likely source of repeat sales is the base of repeat customers who return regularly. When inertia is present, *past period*, or *lagged performance* <sub>$t-1$</sub>  may be a good predictor of current *performance*:

$$\hat{\text{performance}}_t = b_0 + b_1 \text{performance}_{t-l},$$

where  $l$  is the length of lag. If performance is cyclical, performance several periods ago may be a better indicator of current performance than last period's performance.

Amanda, a recent business school graduate with modeling expertise, was asked to build a model of Home Depot Revenues, which would both explain fluctuations and forecast revenues in the next four quarters.

---

<sup>1</sup> This example is a hypothetical scenario based on actual data

Home Depot executives wanted to know

- how much inertia, or repeated buying from customer loyalty existed in Home Depot sales, and
- how strongly the growth in past *New Home Sales* influenced revenues.

After being briefed by the executives, Amanda created a model reflecting their logic. She included as possible drivers in her model:

- an inertia component to capture repeated purchases, *Home Depot revenues*(\$B)<sub>q-l</sub>
- *new home sales*(K)<sub>q-l</sub>

#### 9.4 Compare Scatterplots across Time to Choose Length of Lags For Drivers of Delayed Response: Visual Inspection

Amanda plotted the revenues and each of the suspected drivers. She suspected that four quarter lags were the best choices for inertia and past new home sales growth, since seasonality was expected in both, though six or eight quarter lags were also possibilities. Her scatterplots are shown in Figure 9.2.

Amanda added trend lines for reference. The trend is the average linear growth over the series. She noted quarters in which Home Depot revenues were growing faster than average. These are boxed in black. Quarters when growth was below average she boxed in gray. Faster than average growth and below average growth in the leading indicator was boxed similarly. Then, comparing periods of unusually high or low growth in the leading indicator with those in Home Depot revenues, she identified the response delay.

New home sales slowed in the second quarter of 1999, leading the Home Depot slowdown in the fourth quarter of 2000 by six quarters. New home sales began growing faster in the second quarter of 2003, a year before Home Depot revenues began growing faster. There seemed to be a four to six quarter delay between housing market changes and Home Depot revenue changes.

Both Home Depot revenues and new home sales were highly seasonal, reflecting weather related influences on construction and school year influences on home buying. Including new home sales in the model would remove some of the seasonality in Home Depot revenues.

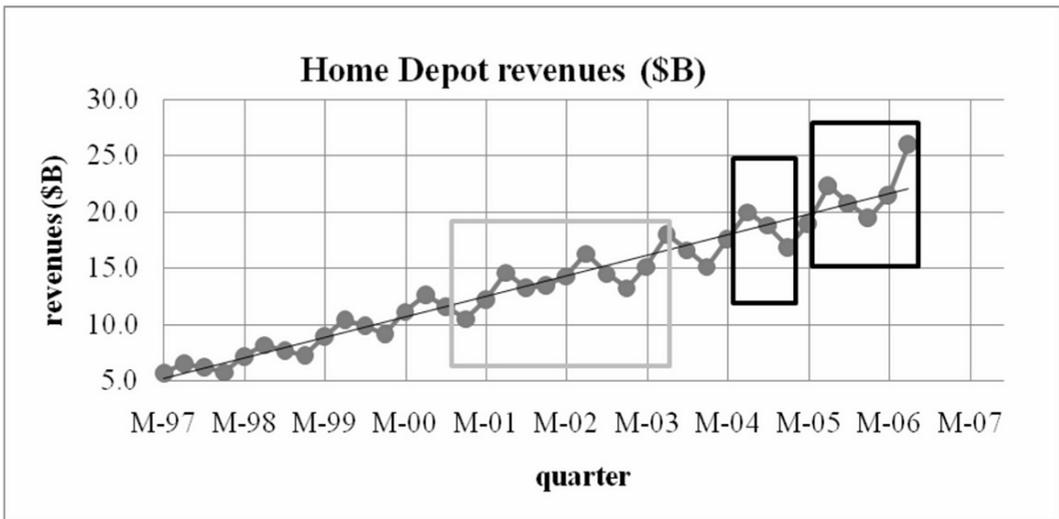
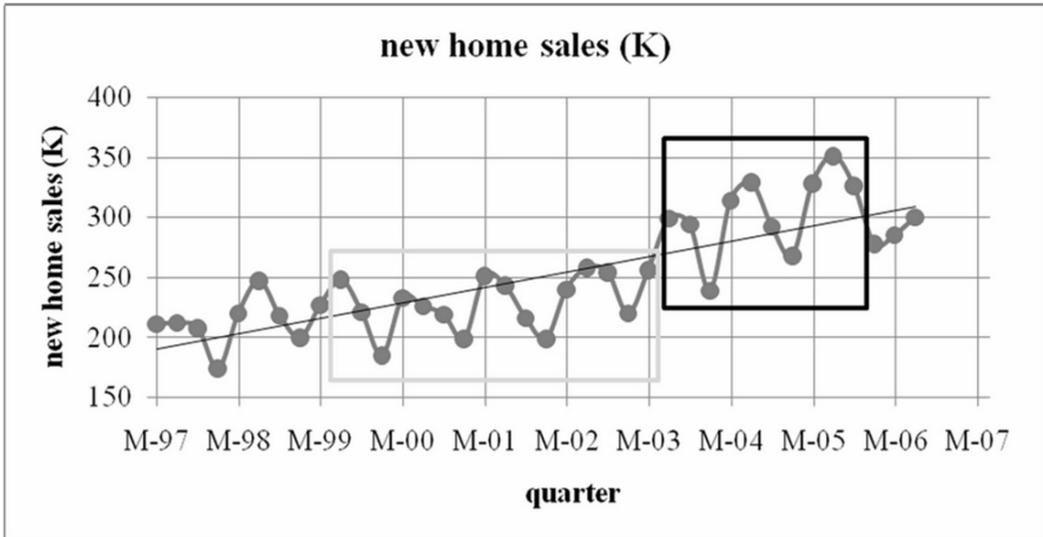


Figure 9.2 Home Depot revenues and new home sales '97 – '07

## 9.5 Hide the Two Most Recent Datapoints to Validate a Time Series Model

Before Amanda fit the multiple regression to quantify the impact of drivers, she excluded the two most recent observations. These *hold out* observations would allow her to compare forecasts for the two most recent periods with actual orders to *validate* her model. If the 95% prediction intervals from the model contained the actual revenues for both quarters, she would be able to conclude that her model is valid. She could then use the model to forecast with confidence.

## 9.6 Correlations Guide Choice of Lags

To reinforce the visual inspection of cycles in revenues and the leading indicator, Amanda looked at correlations between Home Depot revenues and candidate lags for new home sales. She would begin by choosing the leading indicator lag with the highest correlation. Correlations are shown in Table 9.1 and exclude the two most recent quarters.

	<i>new home sales</i> (K) <sub>q-6</sub>	<i>new home sales</i> (K) <sub>q-4</sub>
<i>Home Depot revenues</i> (\$B) <sub>q</sub>	0.626	0.890

Table 9.1 Correlations with Home Depot revenues

The largest correlation is with new home sales lagged four quarters. Amanda ran a regression with *new home sales* (K)<sub>q-4</sub>, then looked at correlations between the residuals and each of the Home Depot revenue lags to choose one to add to the model. Her regression with *new home sales* is shown in Table 9.2.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.890					
R Square	0.791					
Adjusted R Square	0.786					
Standard Error	2.414					
Observations	38					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	796.6	796.6	136.7	0.0000	
Residual	36	209.9	5.8			
Total	37	1006.4				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-11.2	2.2	-5.2	0.0000	-15.5	-6.8
<i>New home sales(K)</i> <sub>q-4</sub>	0.104	0.009	11.7	0.0000	0.086	0.122

Table 9.2 Regression with lagged new home sales

Quarterly variation in past year new home sales accounts for 79% of the quarterly variation in Home Depot Revenues ( $R\ Square=.79$ ). The model is significant ( $Significance\ F=.0000$ ), and the slope estimate is positive (.104). The standard error is \$2.4 (\$B); Amanda could expect forecasts to be within approximately \$4.8B (=2 x \$2.4B) in 95% of quarters.

## 9.7 The Durbin Watson Statistic Identifies Autocorrelation

The Durbin Watson ( $DW$ ) statistic allows us to confirm that trend and cycles in the data have been accounted for. If  $DW$  indicates *autocorrelation*, the correlation of residuals with over time, a trend or cycle has been ignored.

The Durbin Watson statistic compares the sum of squared differences between residuals separated by one time period with the sum of squared residuals:

$$DW = \frac{\sum_2^N (e_q - e_{q-1})^2}{\sum_1^N e_q^2}.$$

If we have accounted for all of the trend and cycles in the data,  $DW$  will exceed two. The leading indicator model has  $DW=1.06$ . This does not exceed two.

In cases where  $DW$  does not exceed two, we refer to a table to determine whether or not unaccounted for cycles, *autocorrelated* residuals, exist. Critical values depend on the number of drivers in a model (including the intercept) and the sample size,  $N$ .  $DW$  critical values can be found online at [stanford.edu/~clint/bench/dwcrit.htm](http://stanford.edu/~clint/bench/dwcrit.htm), found by googling “Durbin Watson critical values.” (In this online table, sample size is indexed by  $T$  and the number of independent variables, plus intercept, is indexed by  $K$ .)

For a model with one independent variable and intercept with a sample size of 38, the critical values at 95% confidence are those in Table 9.3.

$T$	$K$	$dL$	$dU$
38	2	1.43	1.53

With one drivers, plus intercept and a sample size of 38,  $DW$  table values are  $dL=1.43$  and  $dU=1.53$ . Since the model  $DW$  is 1.06, which is below  $dL$ . The residuals are autocorrelated. The data contain trend or cycles not accounted for by the model.

Table 9.3 *Durbin Watson Test* critical values

Examining the residuals is likely to provide clues to identify which variables can be added to account for the trend or cycles.

## 9.8 Assess Residuals to Identify Unaccounted For Trend or Cycles

Model residuals should show neither trend nor cyclical. If we have omitted an important driver, the residuals will not be pattern-free. The residuals will provide clues to help identify which variable to add to the model next.

Amanda plotted the residuals across quarters in Figure 9.3, and observed positive trend and some remaining seasonality. Home Depot revenues were growing faster than new home sales, and the additional unaccounted for growth appears in the residuals. Adding an inertia component, past Home Depot revenues, will account for trend and remaining seasonality. Adding past revenues will also allow Amanda to quantify the loyalty factor that managers believed they had achieved.

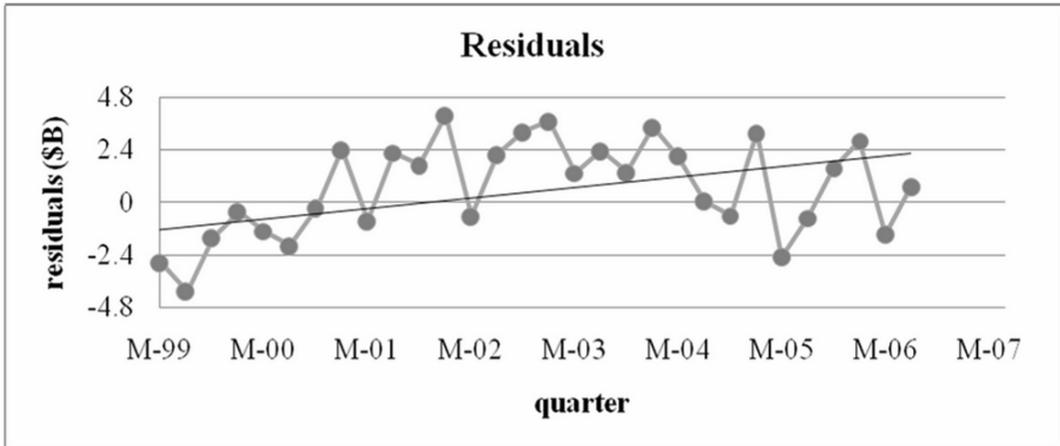


Figure 9.3 Residuals are not pattern-free

Since revenues were seasonal, Amanda considered a four-, six-, and eight-quarter lag for past revenues. The correlations between the candidate past revenue lags and residuals are shown in Table 9.4.

	<i>Home Depot revenues (\$B)<sub>q-4</sub></i>	<i>Home Depot revenues (\$B)<sub>q-6</sub></i>	<i>Home Depot revenues (\$B)<sub>q-8</sub></i>
<i>Residuals</i>	0.482	0.582	0.396

Table 9.4 Correlations with residuals

Past revenues are correlated with residuals, and the highest correlation is with a six quarter lag. Amanda added *Home Depot revenues* with a six quarter lag. Regression results are in Table 9.5.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.971					
R Square	0.943					
Adjusted R Square	0.940					
Standard Error	1.235					
Observations	36					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	836.67	418.34	274.2	0.0000	
Residual	33	50.35	1.53			
Total	35	887.02				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-7.17	1.18	-6.1	0.0000	-9.57	-4.77
<i>New home sales (K) <sub>q-4</sub></i>	0.0607	0.006	9.7	0.0000	0.0480	0.0734
<i>Home Depot revenues (\$B) <sub>q-6</sub></i>	0.582	0.062	9.4	0.0000	0.456	0.708

Table 9.5 Regression with new home sales and past Home Depot revenues

Adding past new home sales added fifteen percent to *RSquare* and reduced the standard error considerably. Forecasts can now be expected to fall within \$2.4B (=2 x 1.2) actual revenues. Both drivers are significant and have the expected positive signs.

There is evidence of loyalty and repeat buying. For each dollar of past revenue, Home Depot management can expect \$.46 to \$.71 revenue six quarters later.

*DW* for this model is 1.85. Since this is less than 2.00, we compare with table values for sample size of 36 and three independent variables (including the intercept):

*Durbin Watson critical values*

$T \quad K \quad dL \quad dU$   
 36. 3. 1.36 1.59

The model  $DW$  exceeds  $dU$ , allowing the conclusion that the residuals are now free of unaccounted for trend or cycles.

### 9.9 Forecast the Recent, Hidden Points to Assess Predictive Validity

With a significant model, logically correct coefficient signs, and residuals free of autocorrelation, Amanda could proceed to assess the predictive validity of her model by comparing actual *Home Depot revenues (\$B)* in the two most recent quarters with the model's 95% prediction intervals. (Recall that those two most recent years were hidden and not used in the regression to fit the model.)

<i>quarter</i>	<i>Home Depot revenues (\$B)</i>	
	<i>lower 95% prediction</i>	<i>upper 95% prediction</i>
S-06	21.2	26.2
D-06	20.2	25.2

The model prediction intervals in Table 9.6 do contain actual revenues in both of the most recent quarters, confirming validity. The model can reliably used to forecast.

Table 9.6 Model Predictions Include Actual Values

### 9.10 Add the Most Recent Datapoints to Recalibrate

With evidence of predictive validity, Amanda used the model to forecast revenues in the next four quarters. Before making the forecast, she added the two most recent observations that were hidden to validate. The recalibrated model became:

$$\hat{revenues}(\$B)_q = -7.07^a + .0611^a \text{NewHomeSales}(K)_{q-4} + .561^a \text{revenues}(\$B)_{q-6}$$

*RSquare: .95*  
<sup>a</sup>*Significant at .01.*

Model forecasts are shown in Figure 9.4 and Table 9.7.

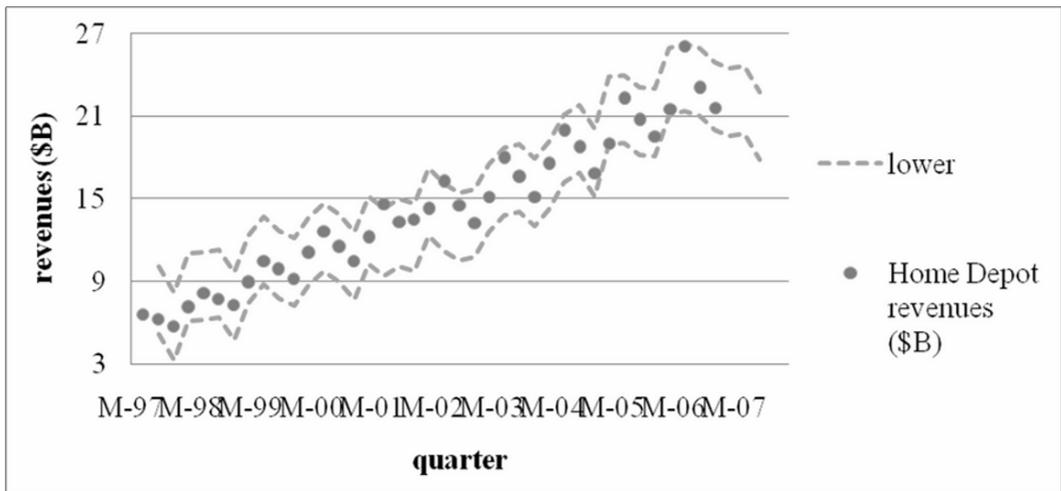


Figure 9.4 Downturn in revenues forecast for 2007

<i>quarter</i>	<i>95% lower prediction</i>	<i>95% upper prediction</i>	<i>Prior year quarterly Revenues</i>	<i>Forecast growth from past year</i>
M-07	19.5	24.5	21.5	2.5%
J-07	19.7	24.7	26.0	-14.7%
S-07	17.9	22.8	23.1	-12.0%

Table 9.7 Quarterly revenue forecast

Revenues in the next quarter are expected to match revenues from the same quarter last year. In the second and third quarters of 2007, revenues are expected to decline substantially. Annual quarterly growth (from each quarter to the same quarter the next year) averaged 12% over the past five years, which suggests that Home Depot revenues will take an unusual turn downwards in 2007, following new home sales.

Amanda summarized her model results for Management:

## MEMO

**Re: Revenue Decline Forecast Following New Home Sales Downturn**  
**To: Home Depot Management**  
**From: Amanda Chanel**  
**Date: June 2007**

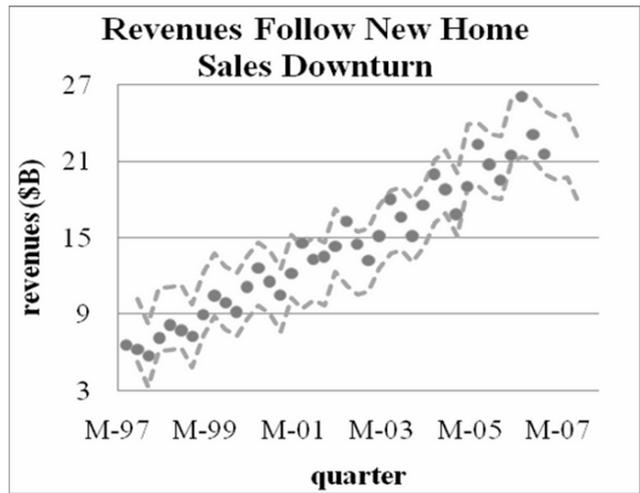
Past growth in revenues and new home sales drive revenue growth.

A regression model of quarterly revenues was built from past quarterly revenues and new home sales. The model accounts for 95% of the variation in revenues and produces valid forecasts within \$2.4 billion of actual revenues.

**Model results.** Results suggest that quarterly revenues are driven housing market movement and inertia from repeat sales to loyal customers.

Following a billion dollar increase in revenues in a quarter, an increase of \$400 to \$700 million in quarterly revenues is expected six quarters later, indicating customer loyalty and repeat sales.

Following a decline of one thousand new homes sold, revenues are expected to decline \$50 to \$70 million in the same quarter the following year.



$$\hat{revenues}(\$B)_q = -7.07^a + .0611^a \text{NewHomeSales}(K)_{q-4} + .561^a \text{revenues}(\$B)_{q-6}$$

*RSquare: .95<sup>a</sup>*

*<sup>a</sup>Significant at .01*

**Three Quarter Forecast.** Home Depot Revenues will decline over the next three quarters, following shrinking new home sales in 2006.

quarter	past year new home sales growth	forecast revenue (\$B)		forecast growth % from same quarter last year
		low	high	
Mar-07	2.5%	19.5	24.5	2.5%
Jun-07	5.3%	19.7	24.7	-14.7%
Sep-07	-16.3%	17.9	22.8	-12.0%

**Conclusions.** Home Depot Revenues contain a stable component of repeat sales to a loyal customer base. Revenues follow housing market indicators, and this vulnerability to declines in the housing market suggests diversification into businesses not closely tied to housing.

## 9.11 Inertia and Leading Indicator Components Are Powerful Drivers and Often Multicollinear

Like cross sectional models, time series models allow us to identify performance drivers and forecast performance. However, time series models differ from cross sectional models, and the model building process with time series contains additional steps.

- Often lagged predictors are used to make driver identification more certain.
- Lagged predictors tend to move together across time and are often highly correlated. Consequently, to minimize multicollinearity issues, model building begins with one predictor, and then others are added, considering their joint influence and incremental model improvement.
- Forecasting accuracy of time series models is tested, or validated, before they are used for prediction of future performance.

Predictors in time series models tend to be highly correlated, since most move with economic variables and most exhibit predictable growth (*trend*). Model building with time series begins with the strongest among logical predictors, and additional predictors are added which improve the model.

Time series typically contain trend, business cycles, and seasonality that are captured with these components. Unaccounted for trend, cycles, or seasonality are detected through inspection of the residual plot and the Durbin Watson statistic. Inertia and one or more leading indicators are often added to multiple regression models built from time series data. Logically, future performance ought to depend upon past performance and economic prosperity—*inertia*. Leading indicators are often stable and predictable performance drivers.

Useful forecasting models must be valid. Holding out the two most recent performance observations allows us to test a model's forecasting capability. With successful prediction of the most recent performance, we can use a recalibrated forecasting model with confidence to forecast what performance will be in future periods.

## Excel 9.1 Build and fit a multiple regression model with multicollinear time series

*Home Depot Revenues.* We will build a model of Home Depot quarterly revenues which includes past revenues growth and past new home sales. The data are in **Excel 9.1 Home Depot.xls**.

**Select length of lag.** To decide how many months *new home sales<sub>q</sub>* will be lagged, make scatterplots of *Home Depot revenues* and *new home sales* over quarters.

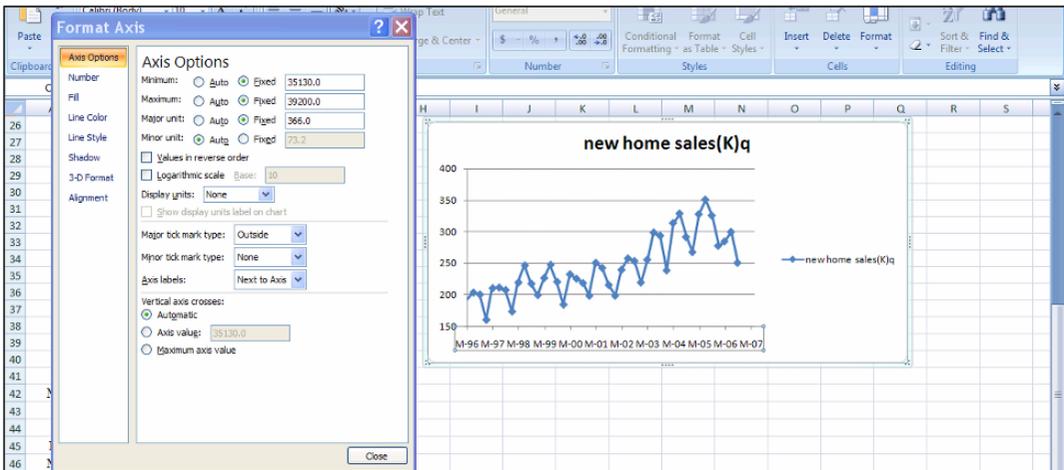
Select the horizontal *quarter* axis, right click and **Format Axis**.

Time periods are measured in days.

To set the axis beginning at March 1996, enter 35130 for **Minimum**.

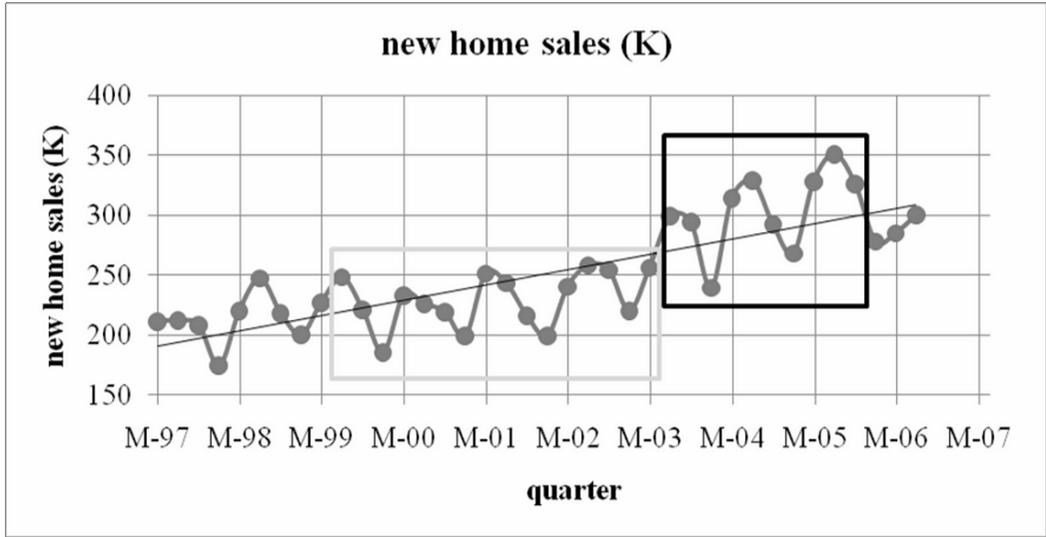
To make the axis end at March 2007, enter 39200 for **Maximum**.

Set **major units** at 366, the number of days in a year:



Use shortcuts to add a trendline: **Alt JAN:**

Visually inspect the *new home sales* scatterplot and record quarters that seem to be growing faster than average:



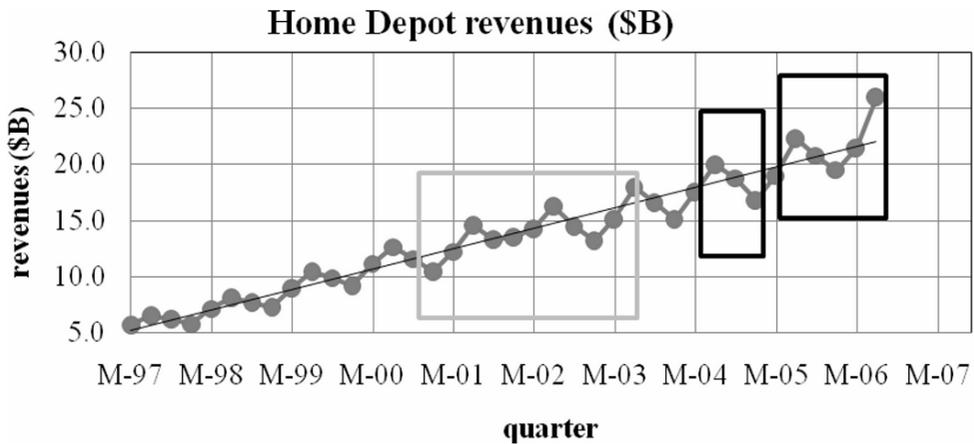
*New home sales begin*

- Slowing in the second quarter of 1999
- growing faster in the second quarter of 2003.

**Format axes** in your *Home Depot revenues* scatterplot so that the graph shows the time period March 1997 through March 2007.

Add a trendline.

Visually inspect the series to identify quarters of greater than average growth:



*Home Depot revenues* begin

- slowing in the fourth quarter of 2000, six quarters after the *new home sales* slowdown
- growing faster in the second quarter of 2004, four quarters after the *new home sales* acceleration.

Four or six quarter lags seem like the best choices for *new home sales*.

To find the correlation between revenues and new home sales lagged four quarters, select and copy *quarters*, *Home Depot revenues* and *new home sales* in columns **A**, **B**, and **C**, and paste into **D**, **E** and **F**.

Change the labels in **D**, **E** and **F** to *quarter from March 1997*, *Home Depot revenues from March 1997*. and *new home sales q-4*.

Delete *quarters* and *Home Depot revenues* for March 1996 through December 1996 in the new columns: select **D2:E5**, **Alt HDD**, **shift cells up**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
		Home Depot revenues (\$B) q	new home sales(K) q	Quarter from M 97	Home Depot revenues (\$B) q from M 97	new home sales(K) q-4													
1	Quarter	(\$B) q	q	97	97	q-4													
2	M-96	4.4	192	M-97	5.7	192													
3	J-96	5.3	204	J-97	6.6	204													
4	S-96	4.9	201	S-97	6.2	201													
5	D-96	5.0	161	D-97	5.7	161													

Find the correlation between *new home sales q-4* and *Home Depot revenues q*, using **E1:F39**.

(There are two more quarters, in rows **41** and **41**, which we are hiding, in order to validate the model later.)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
		Home Depot revenues (\$B) q from	new home sales(K) q-4																
1	Home Depot revenues (\$B) q from																		
2	M 97		1																
3	new home sales(K) q-4	0.8896586	1																
4																			

To find the correlation between revenues and new home sales lagged six quarters, select *quarters*, *Home Depot revenues*, and *new home sales* in columns **A**, **B**, and **C**, copy and paste into **G**, **H**, and **I**.

Change labels in **G**, **H**, and **I** to *quarter from S-97*, *Home Depot revenues (\$B) from S-97*, and *new home sales q-6*.

Delete *quarters* and *Home Depot revenues* from March 1996 through June 1997: select **G2:H7, Alt HDD, shift cells up**.

Find the correlation between *new home sales q-6* and *Home Depot revenues q*, using **H1:I37**.

(There are two more quarters of data which we have hidden for later validation.)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		Home Depot revenues (\$B) q from S 97	new home sales (K) q-6															
2	Home Depot revenues (\$B) q from S 97		1															
3	new home sales (K) q-6	0.625879559																

The correlation with *new home sales* lagged four quarters is higher, so we will use a four-quarter lag.

Run a regression of *Home Depot revenues q* with *new home sales q-4*, using **E1:E39** for the dependent variable and **F1:F39** for the independent variable:



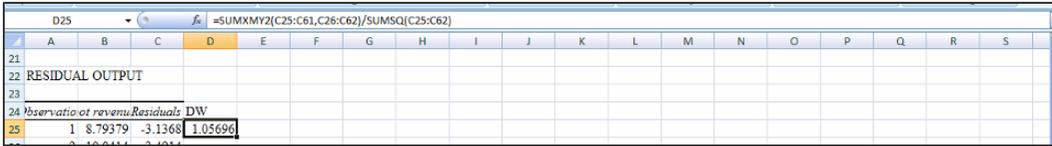
The model is significant and the coefficient sign is positive, as expected. RSquare is relatively low for time series regression, .79, and the standard error is relatively large, 2.41. The margin of error in forecasts would be \$4.8B.

**Assess autocorrelation of the residuals.** If *new home sales* are growing at the same rate as revenues, we will have accounted for trend in the data. *New home sales* are also highly seasonal, like revenues, and cycle with the economy, like revenues do. It is possible that we have accounted for all of the trend, seasonality and cyclicity in the data. In this case, there will be no significant autocorrelation in the residuals.

The Durbin Watson statistic will allow us to assess autocorrelation in the residuals. Next to the residuals in the regression page, find the Durbin Watson statistic using the two Excel functions, **sumxmy2(array1,array2)** and **sumsq(array)**. **Sumxmy2** sums the squared differences between adjacent residuals. For **array1**, enter all but the last residual, and for **array2**, enter all but the first residual. **Sumsq** sums the squared residuals. Enter all of the residuals in this array.

In **D25**, enter `=sumxmy2(c25:c60,c26:c61)/sumsq(c25:c61)` [Enter].

Add the label *DW* in **D24**:



*DW* is less than two, so we consult the online tables. Google “Durbin Watson critical values” to find the Stanford University site: [stanford.edu/~clint/bench/dw05a.htm](http://stanford.edu/~clint/bench/dw05a.htm).

For our sample size, 38, and two independent variables (including the intercept), the critical values are:

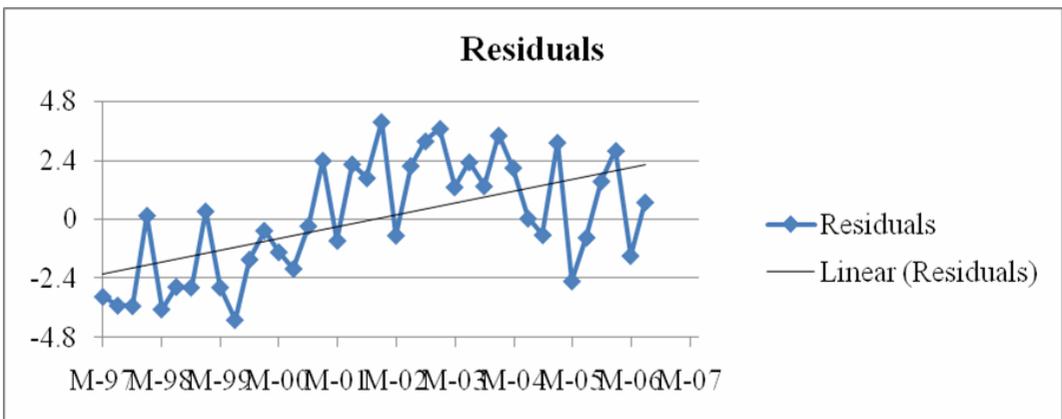
<i>T</i>	<i>K</i>	<i>dL</i>	<i>dU</i>
38.	2.	1.42702	1.53475

*DW* for the model is less than the lower critical value. We conclude that the residuals contain unaccounted for trend or cycles. The next step is to make a scatterplot of the residuals to identify trend, cycles, or seasonality that we can account for by adding one or more variables to the model.

Copy the residuals from the regression page and paste next to the quarters in column **D**, then make a scatterplot over quarters:

**Format Axis** so that quarters range from March 1997 (35490 days) to March 2007 (39200) with major unit of one year (366 days).

Use shortcuts to add a trendline, **Alt JAN**:



It is apparent that *Home Depot revenues* are growing faster than *new home sales*, since a positive trend is left in the residuals. There is also some evidence of seasonality not yet accounted for. Adding an inertia component, *Home Depot revenues* lagged by four, six, or eight quarters, will remove trend and seasonality from the residuals.

**Choose lag for inertia component.** To decide on the number of quarters to lag past revenues, use shortcuts to make the four-quarter lag.

Select column **B**, *Home Depot revenues q*, and copy **Cntl+C**.

Insert in column **F**, following *residuals*: select **F**, **Alt HIE**.

Delete the four cells corresponding to 1996: select **F2:F6**, **Alt HDD**, **Shift cells up**.

Label this new column **F** *Home Depot revenues q-4*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Quarter	Home Depot revenues (\$B) q	new home sales(K) q	Quarter from M	Residuals	Home Depot revenues (\$B) q	Home Depot revenues (\$B) q from M	new home sales(K) q-4	quarter from S	Home Depot revenues (\$B) q from S	new home sales (K) q-6							
2	M-96	4.4	192	M-97	-3.1368	4.4	5.7	192	S-97	6.2	192							
3	J-96	5.3	204	J-97	-3.4914	5.3	6.6	204	D-97	5.7	204							

Find the correlation between the *residuals* and *Home Depot revenues q-4*, using **E1:F39**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1			Home Depot revenues																
2		Residuals	Residuals (\$B) q-4																
3		Home Depot revenues (\$B) q-4	0.481708	1															

To compare with the six quarter lag, add a new column for the six-quarter lag, *Home Depot revenues q-6* in **L**.

Copy **B** *Home Depot revenues q*, select **L** and insert the copied column **Alt HIE**.

Delete the cells corresponding to March 1996 through June 1997: select **L2:L7**, **Alt HDD**, **shift cells up**.

Add a second column *residuals* in **M**.

Select *residuals* in **E** and copy: **Cntl+C**. Select **M** and insert: **Alt HIE**.

Delete the cells corresponding to March and June 1997: select **M2:M3**, **Alt HDD**, **shift cells up**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Quarter	Home Depot revenues (\$B) q	new home sales(K) q	Quarter from M	Residuals	Home Depot revenues (\$B) q-4	Home Depot revenues (\$B) q from M	new home sales(K) q-4	quarter from S	Home Depot revenues (\$B) q from S	new home sales (K) q-6	Home Depot revenues (\$B) q-6	residuals					
2	M-96	4.4	192	M-97	-3.1368	4.4	5.7	192	S-97	6.2	192	4.4	-3.5					
3	J-96	5.3	204	J-97	-3.4914	5.3	6.6	204	D-97	5.7	204	5.3	0.2					
4	S-96	4.9	201	S-97	-3.5125	4.9	6.2	201	M-98	7.1	201	4.9	-3.6					

Use **L1:M37** to find the correlation:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1			Home Depot revenues q-6																
2		residuals	1																
3		Home Depot revenues q-6	0.582358	1															
4																			

The six-quarter lag of revenues is more highly correlated with residuals, so we will add this to the model.

To use the four-quarter lag of *new home sales* with the six-quarter lag of *Home Depot revenues*, the regression data will use quarters beginning in September 1997.

Copy *new home sales q-4* in **H**, then use shortcuts to insert in column **K**: Select **K**, **Alt HIE**.

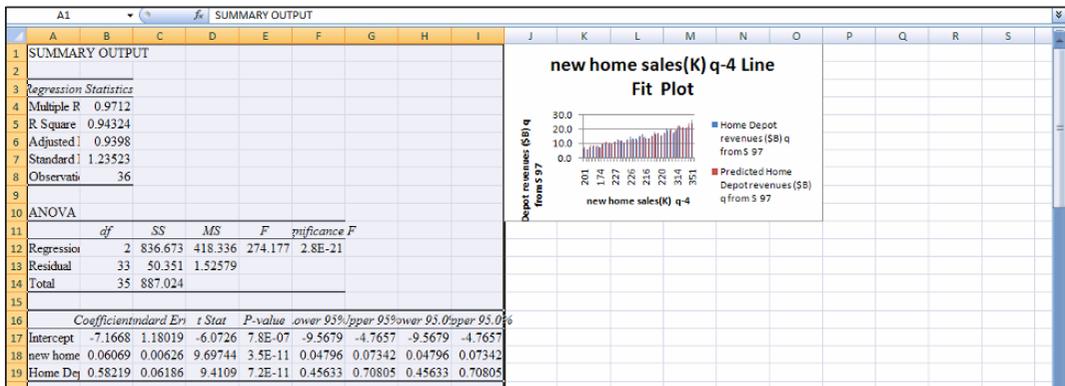
Select the cells corresponding to March and June 1997, **K2:K3**, and use shortcuts to delete: **Alt HDD**, **Shift cells up**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Quarter	(\$B) q	new home sales(K) q	Quarter from M	Residuals (\$B) q-4	Home Depot revenues (\$B) q from M	Home Depot revenues (\$B) q from M	new home sales(K) q-4	quarter from S	Home Depot revenues (\$B) q from S	new home sales(K) q-4	new home sales(K) q-6	Home Depot revenues q-6	residuals		Home Depot revenues (\$B) q	new home sales(K) q-4	hdr q-6
2	M-96	4.4	192	M-97	-3.1368	4.4	5.7	192	S-97	6.2	201	192	4.4	-3.5		6.2	201	4.4
3	J-96	5.3	204	J-97	-3.4914	5.3	6.6	204	D-97	5.7	161	204	5.3	0.2		5.7	161	5.3
4	S-96	4.9	201	S-97	-3.5125	4.9	6.2	201	M-98	7.1	211	201	4.9	-3.6		7.1	211	4.9
5	M-96	5.0	161	M-97	0.16027	5.0	5.7	161	L-97	8.1	217	161	5.0	3.7		8.1	217	5.0

Use shortcuts to move *Home Depot revenues q-6* to column **L**: select **M**, **Ctrl+X**, select **L**, **Alt HIE**.

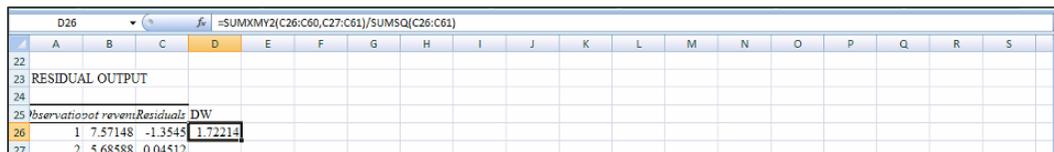
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Quarter	(\$B) q	new home sales(K) q	Quarter from M	Residuals (\$B) q-4	Home Depot revenues (\$B) q from M	Home Depot revenues (\$B) q from M	new home sales(K) q-4	quarter from S	Home Depot revenues (\$B) q from S	new home sales(K) q-4	Home Depot revenues q-6	new home sales(K) q-6	residuals		Home Depot revenues (\$B) q	new home sales(K) q-4	hdr q-6
2	M-96	4.4	192	M-97	-3.1368	4.4	5.7	192	S-97	6.2	201	4.4	192	-3.5		6.2	201	4.4

Run the two-variable regression, Using *Home Depot revenues* in **J1:J37** as the **Input Y range** and *new home sales q-4* and *Home Depot revenues q-6* in **K1:L37** as the **Input X range**:



The two-variable model is significant, and both coefficients have the expected positive sign. *RSquare* as increased to .94, and the standard error is now much smaller: 1.24. The forecast margin of error is now approximately \$2.48B.

**Assess autocorrelation.** To see whether trend and seasonality have been accounted for, find *DW*:



*DW* is less than two, so we consult the online table, finding critical values for sample size 36 and three independent variables (including the intercept):

T	K	dL	dU
36.	3.	1.35365	1.58716

The residuals are now free of significant autocorrelation.

With a significant model, correct signs, an acceptable *RSquare* and standard error, and residuals free of autocorrelation, we are ready to validate the model to see whether it produces accurate forecasts.

**Test the model’s forecasting validity.** To test model validity, copy the regression coefficients in **B16:B19**, and paste into **O** of the original worksheet.

Use the regression equation to make *predicted Home Depot revenues (\$B)* in **P**:

$$\widehat{\text{Home Depot revenues}} (\$B)_q = b_0 + b_1 \text{ new home sales}_{q-4} + b_2 \text{ Home Depot revenues}_{q-6}$$

In **P2** enter =O2 f4 +O3 f4 \*K2+O4 f4 \*L2 [Enter].

Select the new cell, grab and drag through row **42**, filling in the column:

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
	Home Depot revenues (\$B) q from 97	new home sales(K) q-4	quarter from 97	Home Depot revenues (\$B) q from 97	new home sales(K) q-4	Home Depot revenues sales(K) q-6	new home sales(K) q-6	residuals	coefficient	revenues									
1																			
2		5.7	192	S-97	6.2	201	4.4	192	-3.5	-7.1668	7.57148								
3		6.6	204	D-97	5.7	161	5.3	204	0.2	0.06069	5.685877								
4		6.2	201	M-98	7.1	211	4.9	201	-3.6	0.58219	8.504411								
5		5.7	161	J-98	8.1	212	5.0	161	-2.7		8.586643								
6		7.1	211	S-98	7.7	208	5.7	211	-2.8		8.750249								

Make the 95% lower and upper prediction intervals.

First copy the regression standard error from **B7** and paste into **Q2**.

Find the appropriate *t* value for 33 residual degrees of freedom by entering in **R2 = TINV(.05, 33)** [Enter].

Make the 95% lower and 95% upper Home Depot revenues (\$B) in **S** and **T** by subtracting and adding the margin of error, which is *t* in **R2** times the standard error in **Q2**:

In **S2**, enter =P2-Q2 f4 \*R2 f4.

In **T2**, enter = P2+Q2 f4 \*R2 f4.

	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
	Home Depot revenues (\$B) q from 97	new home sales(K) q-4	Home Depot revenues sales(K) q-6	new home sales(K) q-6	residuals	coefficient	revenues	standard error	t	lower 95% prediction	upper 95% prediction							
1																		
2		6.2	201	4.4	192	-3.5	-7.1668	7.57148	1.23523	2.03452	5.1	10.1						
3		5.7	161	5.3	204	0.2	0.06069	5.685877			3.2	8.2						
4		7.1	211	4.9	201	-3.6	0.58219	8.504411			6.0	11.0						

Select the new cells **S2:T2**, grab and drag through row **42**, filling in the prediction interval columns:

quarter from S 97	Home Depot revenues (\$B) q from S 97	new home sales(K) q-4	Home Depot revenues q-6	new home sales (K) q-6	residuals	coefficient	predicted Home Depot revenues	standard error	t	lower 95% prediction	upper 95% prediction
32	M-05	19.0	314	16.6	294	-2.5	21.55317			19.0	24.1
33	J-05	22.3	329	15.1	239	-0.7	21.60597			19.1	24.1
34	S-05	20.7	292	17.6	314	1.6	20.77223			18.3	23.3
35	D-05	19.5	268	20.0	329	2.8	20.71873			18.2	23.2
36	M-06	21.5	328	18.8	292	-1.5	23.66852			21.2	26.2
37	J-06	26.0	351	16.8	268	0.7	23.92331			21.4	26.4
38	S-06	23.1	326	19.0	328		23.66416			21.2	26.2
39	D-06	21.5	278	22.3	351		22.69087			20.2	25.2
40	M-07		285	20.7	326		22.2069			19.7	24.7
41	J-07		300	19.5	278		22.38662			19.9	24.9
42			251	21.5	285		20.56086			18.0	23.1
43				26.0	300						
44				23.1	251						

Comparing the prediction intervals in rows **38** and **39** for the two most recent quarters, September and December 2006, we find that the prediction intervals in **S** and **T** do contain actual revenues in **J**. The model is valid and produces accurate forecasts.

**Recalibrate to forecast.** Recalibrate the model by rerunning the regression with rows **1** through **39**, this time including the two most recent quarters.

df	SS	MS	F	significance F
12	2.966627	483.313	327.669	2.2E-23
13	35	51.6251	1.475	
14	37	1018.25		

	Coefficient	Standard Error	t Stat	P-value	lower 95%	upper 95%	lower 95.0%	upper 95.0%
Intercept	-7.0688	1.12691	-6.2727	3.4E-07	-9.3565	-4.781	-9.3565	-4.781
new home	0.06112	0.00597	10.2383	4.6E-12	0.049	0.07324	0.049	0.07324
Home Dep	0.56118	0.05612	9.99978	8.5E-12	0.44725	0.67511	0.44725	0.67511

With the two most recent quarters included, *RSquare* is slightly higher, and now .95, and the standard error is slightly lower, and now 1.21. The forecast margin of error becomes \$2.4B.

The final model equation is:

$$\hat{Home\ Depot\ rev\ (\$B)_q} = -7.07 + .061\ new\ home\ sales_{q-4} + .561\ Home\ Depot\ rev\ (\$B)_{q-6}$$

Copy and paste the recalibrated *coefficients* over the validation coefficients in **O** which will update the *predicted Home Depot revenues* in **P**.

Copy and paste the recalibrated *standard error* over the validation standard error in **Q**, and update *t* to reflect 35 residual dfs, which will update the prediction columns.

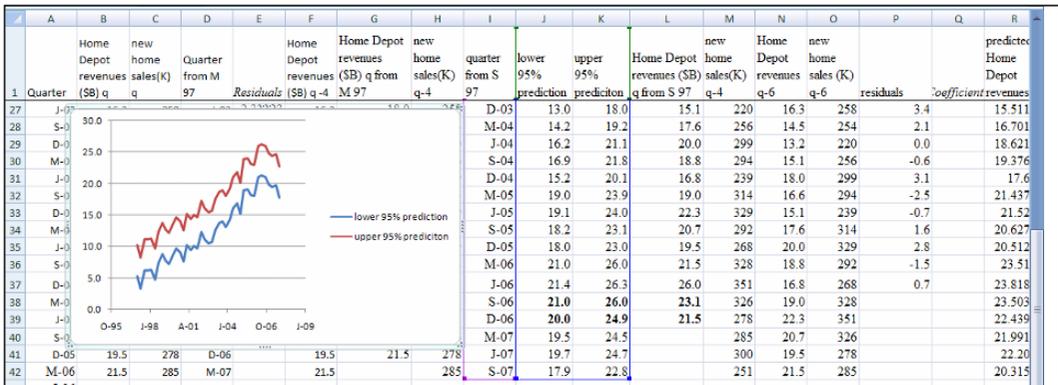
quarter from S	Home Depot revenues (\$B) q from S 97	new home sales(K) q-4	Home Depot revenues q-6	new home sales (K) q-6	residuals	coefficient	revenues	predicted Home Depot revenues	standard error	t	lower 95% prediction	upper 95% prediction
1 97												
2 S-97	6.2	201	4.4	192	-3.5	-7.0688	7.664253	1.2145	2.03011		5.2	10.1
3 D-97	5.7	161	5.3	204	0.2	0.06112	5.741901				3.3	8.2
4 M-98	7.1	211	4.9	201	-3.6	0.56118	8.589715				6.1	11.1
5 J-98	8.1	212	5.0	161	-2.7		8.671598				6.2	11.1

**Illustrate the fit and forecast.** To see the model fit and forecast, plot *Home Depot revenues (\$B)* and *95% predicted lower and upper values* by *quarter*.

Rearrange columns.

Select and cut the prediction interval columns **S** and **T**, **Cntl+X**, then insert into columns **J** and **K**: select **J**, **Alt HIE**.

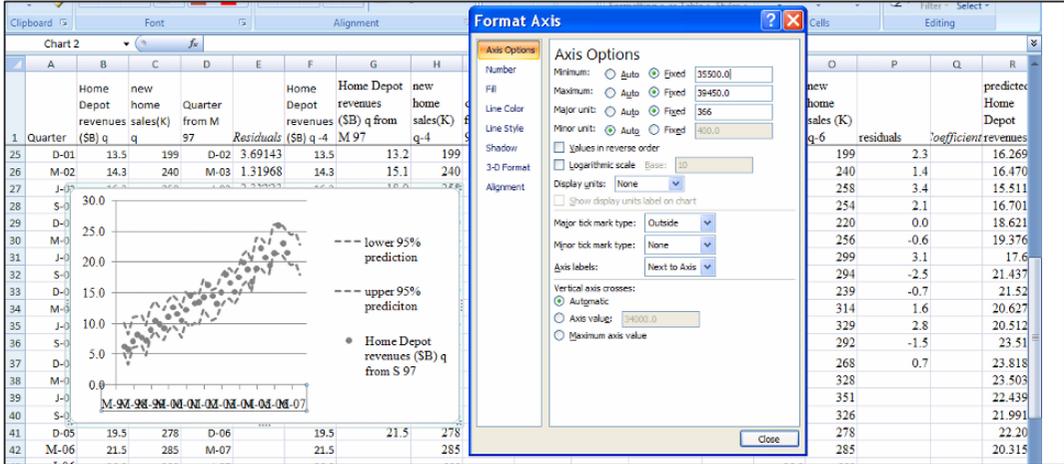
Make a scatterplot. Select **I1:K44**, **Alt ND**.



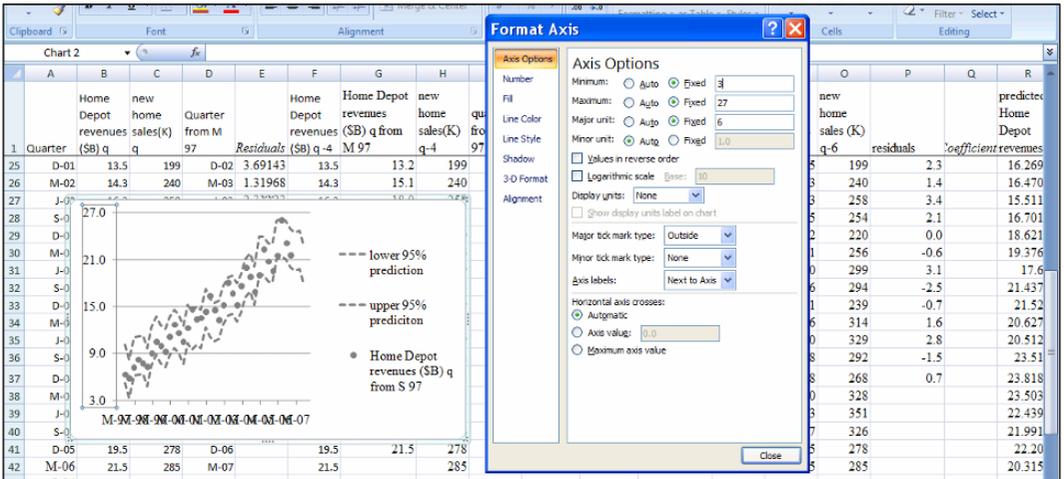


Rescale the horizontal axis to show March 1997 through March 2007. Select the quarters, right click and **Format Axis**.

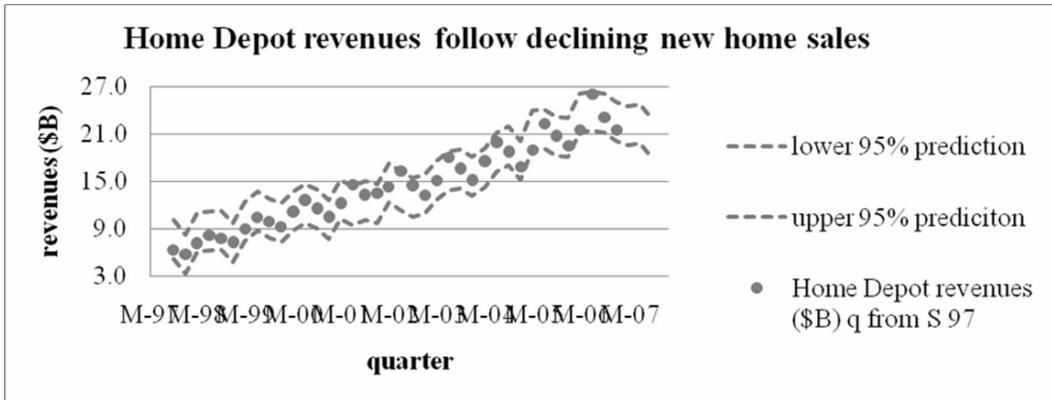
Set the **Minimum** to 35400, the **Maximum** to 39400, and **Major unit** to 366, then **Close**:



Reset the vertical axis to **Minimum 3**, **Maximum 27**, and **Major unit 6**:



Choose **Chart Layout 1** from the **Design** menu and type in chart and axes titles:



**Assess the impact of drivers.** We will use the regression equation to look at the impact of each of the drivers on model forecasts.

**Impact of past year new home sales.** To see the impact of the leading indicator, add *growth in past year new home sales* in column U.

In U3, enter  $= (M3 - M2) / M2$  [Enter].

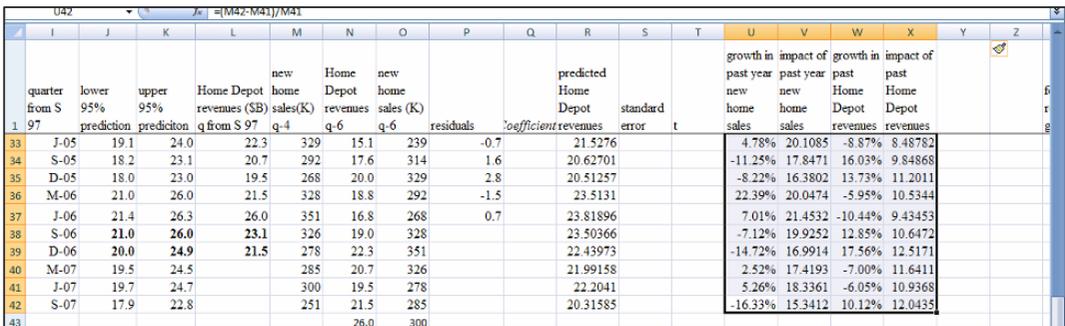
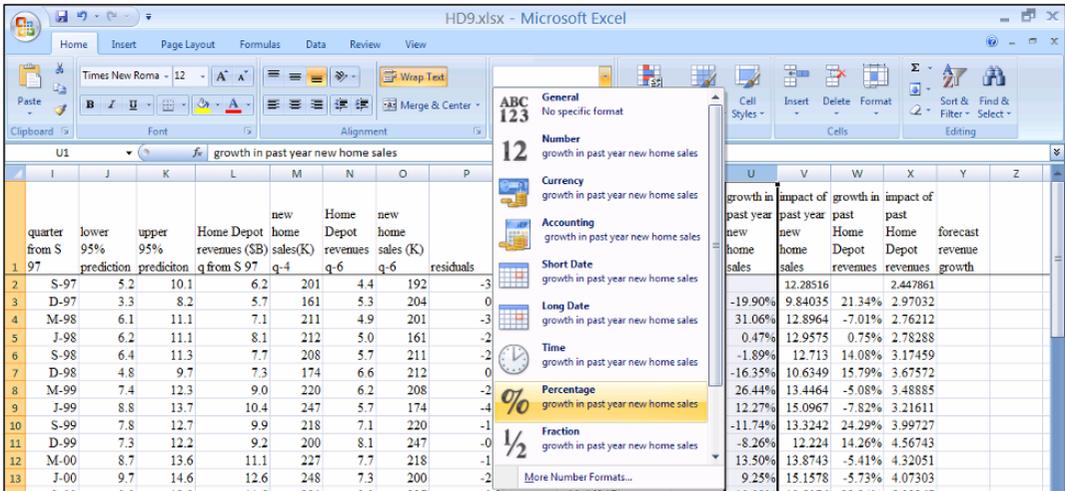
Add the *impact of past year new home sales* in column V and in V2 enter  $= Q3 f4 * M2$  [Enter].

**Impact of past Home Depot revenues.** To see the impact of inertia, *growth in past Home Depot revenues* in column **W**, and in **W3**, enter  $= (W3 - W2) / W2$  [Enter].

Add *impact of past Home Depot revenues* in column **X** and in **X2** enter  $= Q4\ f4 * N2$  [Enter].

Select **U2:X2**, grab, and drag through row **42**, filling in the cells.

Change growth rates in **U** and **W** to percents:



Growth in *new home sales* in column **U** has been noticeably lower in the last four quarters, relative to the same quarters the year before. The impact of declining *new home sales* leads declining revenues.

Growth in past revenues, in column **W** has improved, relative to the same quarters the year before. This has dampened the impact of declining new home sales on revenues.

**Forecast revenue growth.** Find *forecast revenue growth* in column Y.

In Y3 enter  $=\text{(L3-L2)}/\text{L2}$  [Enter].

In Y40 enter  $=\text{(R40-L39)}/\text{L39}$  [Enter].

In Y41 enter  $=\text{(R41-R40)}/\text{F40}$  [Enter].

Grab this new cell and drag through row 42.

Change *forecast revenue growth* rates to percents:

	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	quarter from S 97	lower 95% prediction	upper 95% prediction	Home Depot revenues q from S 97	new Home Depot sales(K) q-4	Home Depot revenues q-6	new Home Depot sales (K) q-6	residuals	coefficient	predicted Home Depot revenues	standard error	t	growth in new home sales past year	impact of new home sales past year	growth in Home Depot revenues past	impact of Home Depot revenues past	forecast revenue growth	
35	D-05	18.0	23.0	19.5	268	20.0	329	2.8	20.51257				-8.22%	16.3802	13.73%	11.2011	-6.05%	
36	M-06	21.0	26.0	21.5	328	18.8	292	-1.5	23.5131				22.39%	20.0474	-5.95%	10.5344	10.12%	
37	J-06	21.4	26.3	26.0	351	16.8	268	0.7	23.81896				7.01%	21.4532	-10.44%	9.43453	21.27%	
38	S-06	21.0	26.0	23.1	326	19.0	328		23.50366				-7.12%	19.9252	12.85%	10.6472	-11.30%	
39	D-06	20.0	24.9	21.5	278	22.3	351		22.43973				-14.72%	16.9914	17.56%	12.5171	-6.68%	
40	M-07	19.5	24.5		285	20.7	326		21.99158				2.52%	17.4193	-7.00%	11.6411	2.09%	
41	J-07	19.7	24.7		300	19.5	278		22.2041				5.26%	18.3361	-6.05%	10.9368	0.97%	
42	S-07	17.9	22.8		251	21.5	285		20.31585				-16.33%	15.3412	10.12%	12.0435	-8.50%	
43						26.0	300											

Forecast revenue growth in 2007 is noticeably lower than in the same quarters in 2006, following declining growth in *new home sales*, but dampened by growing Home Depot customer loyalty.

## Chapter 9 Lab: HP Revenue Forecast

Mark Hurd, Hewlett Packard's new CEO would like to promise shareholders that worldwide revenues will reach \$100 billion by 2008. You have been hired to confirm that this seems likely. He is concerned by Chinese competitors who are gaining ground as China industrializes.

Data are in **Lab 9 HP forecast.xls**, and contain annual *HP revenues* in billion dollars, *GDP* in trillion dollars, *Dell revenues* in billion dollars, and *Chinese per capita GDP* in thousand dollars for the twenty years 1985 through 2008. *Chinese per capita GDP* for 2007 and 2008 are World Bank estimates.

**Make scatterplots to see GDP Leading HP Revenues.** To see how *GDP* leads *HP revenues*, make scatterplots of each by year, and add trendlines to both. Add an 'X' to cells for years in which you see slowed growth:

Year	93	94	95	96	97	98	99	00	01	02	03
GDP slowed											
HP slowed											

Following slowing of *GDP*, *HP* sometimes slows \_\_\_ 2 years later \_\_\_ 3 years later.

Copy *Year* and *HP revenues* into new columns then:

- delete the 8 cells for years 1985 through 1988 and
- delete *HP revenues* in 2004 and 2005 to hide them for later validation.

**Add lagged indicators.** Add in years 1989-2009 seven new columns:

- *GDP t-2* and *GDP t-3*
- *Dell t-2*, *Dell t-3* and *Dell t-4*
- *Chinese per capita GDP t-2*, and *Chinese per capita GDP t-3*

Find the correlations between *HP revenue* and each of the seven lagged variables, then choose the lag with the highest correlation to run a simple leading indicator regression using years 1989-2003.

**Assess autocorrelation.** Look up the Durbin Watson critical values in <http://www.stanford.edu/~clint/bench/dw05a.htm> *dL*: \_\_\_\_ *dU*: \_\_\_\_

Find the model Durbin Watson value using the residuals: \_\_\_\_

Conclude: The model \_\_\_\_ has unaccounted for trend or cycles,  
 \_\_\_\_ may have unaccounted for trend or cycles, or  
 \_\_\_\_ is free of unaccounted for trend or cycles.

Copy the residuals into the HP sheet and find correlations with the three Dell lags and two Chinese lags.

Choose the lagged variable with the highest correlation with residuals to add to your regression.

Compare *RSquares* and *standard errors*:

	<i>RSquare</i>	<i>Standard error</i>
Model with <i>GDP</i>		(\$B)
Model with <i>GDP</i> & additional variable		(\$B)

Look up the Durbin Watson critical values: *dL*: \_\_\_\_ *dU*: \_\_\_\_

Find the model Durbin Watson value: \_\_\_\_

Conclude: The model \_\_\_\_ has unaccounted for trend or cycles,  
 \_\_\_\_ may have unaccounted for trend or cycles.  
 \_\_\_\_ is free of unaccounted for trend or cycles.

Copy the residuals into the HP sheet and find correlations with the two lags for the variable not yet in the model.

Choose the lagged variable with the highest correlation with residuals to add to your regression.

Compare *RSquares* and *standard errors*:

	<i>RSquare</i>	<i>Standard error</i>
Model with <i>GDP</i> & additional variable		(B\$)
Model with <i>GDP</i> , <i>Dell</i> & <i>Chinese per capita GDP</i>		(B\$)

What does the coefficient sign for the lagged *Chinese per capita GDP* variable tell us?

---

Look up the Durbin Watson critical values: *dL*: \_\_\_\_ *dU*: \_\_\_\_

Find the model Durbin Watson value: \_\_\_\_

Conclude: The model \_\_\_\_ has unaccounted for trend or cycles,  
 \_\_\_\_ may have unaccounted for trend or cycles.  
 \_\_\_\_ is free of unaccounted for trend or cycles.

**Validate your model.** Copy the *coefficients* and *standard error* into the HP sheet and use the regression equation to make *Predicted HP revenues* and *lower* and *upper 95% prediction intervals*.

Do prediction intervals contain the hidden *HP revenues* for 2004 and 2005? Y or N

**Recalibrate** by running the regression again with years through 2005.

Can Chairman Hurd claim that HP revenues will reach \$100 billion by 2008? Y or N

## *CASE 9-1 Dell: Overcoming Roadblocks to Growth\**

Data are in **Case 9-1 Dell Revenue Forecast.xls** and contain *Dell Revenues (B\$)*, *U.S. GDP (T\$)*, *Hewlett Packard Revenues (\$B)*, and *China GDP per capita (K\$)* for years 1985 through 2008. (*China GDP per capita* in 2007 and 2008 is an estimate.) *Inertia from past Dell revenues is highly correlated with past Hewlett Packard Revenues. To reduce potential multicollinearity problems, the ratio of Hewlett Packard Revenues to Dell Revenues is also included. If you choose to use the ratio, you should not include Hewlett Packard Revenues or past Dell Revenues.* The ratio may reflect benefits to Dell from Hewlett Packard's marketing efforts, since Hewlett Packard is the larger firm. *You are not limited to the variables in the dataset. The case may give you ideas for other variables that could be useful drivers. You should, however, be able to build a valid forecasting model with the variables provided.*

### **Proposed Steps to Build a Forecasting Model**

1. Plot *Dell Revenues* by year for years 1985 through 2004 and *U.S. GDP* by year for years 1985 through 2006.

Identify the length of delays between changes in *GDP* growth and changes in *Dell revenue* growth, considering two, three or four years.

2. Create columns to use in your model which begin in 1989, adding *GDP* from two, three or four years past.

Check correlations to confirm your choice of lag, and then run a simple regression of *Dell Revenue* with lagged *GDP*.

3. Check model significance, and the coefficient sign, then assess autocorrelation with the Durbin Watson statistic.
4. If your model is not significant, choose a different *GDP* lag, re-run and re-assess; if your model is significant, choose a second driver to add from six candidate drivers: *Hewlett Packard Revenues* and the *ratio of Hewlett Packard to Dell revenues* with two, three and four year lags.

Copy the residuals onto your Dell sheet, and then use correlations with the residuals to choose a lagged competitive driver to add to your model: past *Hewlett Packard Revenues* or past *Hewlett Packard to Dell revenues*.

5. Check model significance, compare *RSquare* and the standard error with your one-variable model, check *p values* and coefficient signs of the two drivers, and then assess autocorrelation.

---

\*Harvard Business School case HKU575

6. If either marginal slope (i.e. coefficient) is not significant, choose a different driver to add to *GDP*, re-run and re-assess; if both slopes are significant, choose a two-, three-, or four-year *Chinese GDP* lag to add.
7. Check model significance, compare *RSquare* and the standard error with your two-variable model, check *p values* and coefficient signs of the three drivers, then assess autocorrelation.
8. If one of the marginal slopes (coefficients) is not significant, choose a different lag; if all slopes are significant, use the regression equation to validate your model.
9. If your model is not valid, try a different lag and re-assess.
10. Once validated, recalibrate your model and make a scatterplot showing your fit and forecast.

(This plot should contain the lower and upper 95% prediction intervals and actual Dell Revenues over years 1989 through 2009.)

**Deliverables.** Present your final model in a one-page, single-spaced memo to Dell executives. (You built a forecasting model from historical time series, using a twenty years of data from the Bureau of Economic Analysis, annual reports, and the International Monetary Fund.) Embed your scatterplot and include your regression equation in standard format.

- Explain how each of the drivers in your model affects *revenues*, including the range of average impact and the length of delay.
- Include *95% prediction intervals* for 2008 and 2009.

Attach your final model regression sheets (i) before recalibration with your Durbin Watson analysis and (ii) after recalibration.

## *CASE 9-2 Mattel Revenues Following the Recalls*

Despite recent press reports that recalls of toys manufactured in China will curb revenues, Mattel management is claiming that revenue growth will double in 2007 and 2008, reaching \$6 billion by 2008.

Mattel management is counting on the growing number of preschool and elementary children to fuel revenues. More children ought to translate to more toy sales.

Management is aware that toys are luxuries and sales are likely to be linked to past growth in GDP.

Mattel managers are also aware that when children choose Hasbro toys, products of their strongest competitor, Mattel has traditionally lost sales.

Build a *valid* Leading Indicator model of Mattel revenues to forecast revenues in 2007 and 2008 from data in **Case 9-2 Mattel.xls**. The dataset contains *Mattel Revenues* (B\$) in billion dollars, *U.S. GDP (\$T)* in trillion dollars, *4-year old population (MM)* in millions, *7-year-old population (MM)* in millions, and *Hasbro revenues (\$B)* in billion dollars for years 1985 through 2006, with population estimates through 2008. Use years 1989 through 2004 to build your model.

First, choose *GDP* from two or three years prior and include this in a regression with *4-* and *7-year olds*.

Next, choose *Hasbro revenues* from two or three years prior.

Write a one-page memo to present your results to management. Include in your memo

- percent of variation in *Mattel revenues* explained with variation in past *GDP*, *4-* and *7-year old* populations, and past *Hasbro revenues*
- margin of error for your forecasts
- the range in *revenue* increase which Mattel can expect following each increase of **\$1T (one trillion dollars)** in *GDP*.  
(Be sure to specify units and when the increase can be expected.)
- the change in *revenue* which Mattel could expect if an additional **1MM (one million)** babies were born four years ago,
- the change in *revenue* expected if an additional **1MM (one million)** babies were born seven years ago

- the expected **revenue** change if **Hasbro revenues** increase by **\$1B (one billion)**, on average.  
(Be sure to specify units and time of the expected change.)
- whether or not your model free of unaccounted for trend and cycles?  
(Use a footnote to refer to the statistic that you are using to draw your conclusion.)
- the range in *revenues* forecast in 2007 and 2008, with 95% confidence
- Likelihood that Mattel will meet its claim to achieve \$6 billion by 2008
- annual *revenue* growth percent average in the past five years, 2002 through 2006 and expected annual growth percent in the next two years
- model validity

Embed a scatterplot of your fit and forecast, including your regression equation, RSquare and significance levels.

### *CASE 9-3 Starbucks in China*

Despite recent press that their revenue growth is stagnating, Starbucks management is claiming that revenues will grow by 20% annually, reaching \$13 billion by 2009.

Starbucks management is counting on the growing coffee consumption in China to fuel revenues. In China, Starbucks coffee is considered a luxury. More and more Chinese will be able to afford the treat, as per capita GDP continues to grow. Two recent articles explain:

*A Tall Espresso Con Panna costs \$1.63, while a small coffee of the day is \$1.50. And a Mocha Frappuccino Grande sets you back a substantial 3.63 at the crowded Starbucks stores of Beijing, Shanghai, and Tianjin. Wait a second – isn't the mainland better known for leaves steeped in water, as demonstrated by the phrase "all the tea in China?" There's no shortage of tea in the country that invented it, but the fact is that java beans are a new sensation for the relatively well-off urban Chinese, who now earn on average \$1,312 per year, up 9.6% this year. [Rural Chinese won't likely be drinking Seattle's finest anytime soon, however; rural incomes, still less than a third of their urban counterparts, this year grew 6.2% to \$407.]*

*In the seven years since H&Q Asia – the former controlling shareholder of Beijing Mei Da Coffee – opened the first Starbucks shop in Beijing in 1999, the Seattle phenomenon has grown to 190 stores in 19 cities in mainland China. "It's not just a drink in China. It's a destination. It's a place to be seen and a place to show how modern one is," adds Technomic Asia's Kedl. And with China's economy growing in double digits, there are likely to be lots more young urban and modern Chinese ready to sip java in a sleek new Starbucks. (Business Week Online, October 26, 2006)*

*Starbucks Corp. executives have forecast that about 20 percent of its international growth will occur in China this year, which has the potential for more than 200 million customers. There already are more than 500 Starbucks Coffee outlets in China, about 300 of which have opened in the past two years, and Martin Coles, president of Starbucks' international division, told a telephone conference of financial analysts that the chain would add 200 more there by 2008. Chairman Howard Schultz, emphasizing Starbucks' current presence in Beijing and 17 provinces, said he anticipates the brand will continue to do well in Hong Kong and gain strength in Taiwan. "We are dreaming very big in China," he said. (Nation's Restaurant News, May 21, 2007)*

Starbucks managers also believe that their loyal customers will continue to return to purchase their favorite coffees, in spite of growing competition.

Build a *valid* Leading Indicator model of Starbucks revenues to forecast revenues in 2007 through 2009 from data in **Case 9-3 Starbucks Revenue.xls**. The dataset contains *Starbucks Revenues* (B\$) in billion dollars, and *China GDP per capita* (\$T) in trillion dollars for years 1988 through 2006, with estimates of *China GDP per capita* through 2008.

Use years 1991 through 2004 to build your model.

First, choose Chinese per capita GDP from two or three years prior.

Next, choose Starbucks revenues from two or three years prior. (Prior revenues reflect inertia in consumer behavior, or the tendency for Starbucks customers to remain loyal, rather than switch to other coffee sources.)

Write a one-page memo presenting your results to management. Be sure to include in your memo:

- percent of variation in Starbucks *revenues* which can be explained with variation in past *Chinese per capita GDP* and past *Starbucks revenues*
- the margin of error for your forecasts
- Following each increase of **\$1K (one thousand dollars)** in *Chinese per capita GDP*, the expected change in *revenues*.  
(Be sure to specify units and the expected time of the change)
- evidence of Starbucks customer loyalty and the extent of this loyalty. . .the range of increase in *Starbucks revenues* expected, following each *revenue* increase of **\$1B (one billion dollars)**
- whether or not your model is free of unaccounted for trend and cycles  
(Use a footnote to include the statistic that you are using to draw your conclusion.)
- the *range* in revenues forecast in 2007, 2008, and 2009 with 95% confidence
- Likelihood that Starbucks' will match its claim to achieve *revenues* of **\$13 billion by 2009**
- Average annual *revenue* growth percent in the past five years, 2002 through 2006 and expected annual growth percent the next three years
- model validity

Embed a scatterplot of your fit and forecast with your regression equation, *RSquare* and *significance* levels.

# 10

## Indicator Variables

In this chapter, we use 0-1 indicator or “dummy” variables to incorporate shocks, structural shifts or segment differences into models. In cross-sectional data, indicators allow us to compare response across groups or segments. In time-series data, indicators allow us to modify responses to account for external shocks or structural shifts. Indicators also offer one option to account for seasonality or cyclicalities in time series.

Model variable selection begins with the choice of potential drivers from logic and experience. Redundant multicollinear variables are then removed. Indicators are added to account for segment differences, shocks, shifts or seasonality, and, if autocorrelation remains, trend, inertia, a leading indicator or an indicator variable may be added to remedy the autocorrelation. These later steps in the variable selection process are considered in this chapter.

This chapter also introduces the use of indicators to analyze data from conjoint analysis experiments. Conjoint analysis is used to quantify customer preferences for better design of new products and services.

### 10.1 Indicators Modify the Intercept to Account for Segment Differences

To compare two segments, we add a 0-1 indicator. One segment becomes the baseline, and the indicator represents the amount of difference from the base segment to the second segment. Indicators are like switches that turn on or off adjustments in a model intercept.

*Example 10.1 Hybrid Fuel Economy.* In a model of the impact of car characteristics on fuel economy:

$$\begin{aligned} \hat{MPG} &= b_0 + b_1 \text{Hybrid} + b_2 \text{Emissions} + b_3 \text{Horsepower} \\ &= 48 + 8.8 \text{Hybrid} - 2.3 \text{Emissions} - .025 \text{Horsepower} \end{aligned}$$

The coefficient estimate of 8.8 for the *hybrid* indicator modifies the intercept. For conventional cars, the *hybrid* indicator is 0, making the intercept for conventional cars 48:

$$\begin{aligned} \hat{MPG} &= 48 + 8.8(0) - 2.3 \text{Emissions} - .025 \text{Horsepower} \\ &= 48 - 2.3 \text{Emissions} - .025 \text{Horsepower} \end{aligned}$$

For hybrids in the sample, the *hybrid* indicator is 1, which adjusts the intercept for hybrids to 56.8 by adding 8.8 to the baseline 48:

$$\begin{aligned} \hat{MPG} &= 48 + 8.8(1) - 2.3 \text{Emissions} - .025 \text{Horsepower} \\ &= 56.8 - 2.3 \text{Emissions} - .025 \text{Horsepower} \end{aligned}$$

The adjustment is switched on when  $hybrid=1$ , but remains switched off if  $hybrid=0$ . The parameter estimate for the indicator tells us that on average, hybrid gas mileage is 8.8 MPG higher than conventional gas mileage.

*Example 10.2 Yankees v Marlins Salaries*<sup>2</sup>. The Yankees General Manager has discovered that the hot rookie whom the Yankees are hoping to sign is also considering an offer from the Marlins. The General Manager would like to know whether there is a difference in salaries between the two teams. He believes that, in addition to a possible difference between the two teams, *Runs* by players ought to affect salaries.

We will build a model of baseball salaries, including *Runs* and an indicator for Team. This variable, *Yankees*, will be equal to 1 if a player is on the Yankee Team, and equal to 0 if the player is a Marlin. The Marlins are our baseline team. Our data are shown in Table 10.1, and regression results are shown in Table 10.2.

<i>Player</i>	<i>Team</i>	<i>Yankees</i>	<i>Position</i>	<i>Runs</i>	<i>Salary(M\$)</i>
Castillo	Marlin	0	Second	72	5.2
Delgado	Marlin	0	First	81	4.0
Pierre	Marlin	0	Outfield	96	3.7
Gonzalez	Marlin	0	Shortstop	45	3.4
Easley	Marlin	0	Second	37	0.8
Cabrera	Marlin	0	Outfield	106	0.4
Aguila	Marlin	0	Outfield	11	0.3
Treanor	Marlin	0	Catcher	10	0.3
Rodriguez	Yankee	1	Shortstop	111	21.7
Jeter	Yankee	1	Shortstop	110	19.6
Sheffield	Yankee	1	Outfield	94	13.0
Williams	Yankee	1	Outfield	48	12.4
Posada	Yankee	1	Catcher	60	11.0
Matsui	Yankee	1	Outfield	97	8.0
tino martinez	Yankee	1	First	41	2.8
womack	Yankee	1	Second	46	2.0
Sierra	Yankee	1	Outfield	13	1.5
Giambi	Yankee	1	Baseman	66	1.3
Flaherty	Yankee	1	Catcher	8	0.8
Crosby	Yankee	1	Outfield	10	0.3
andy phillips	Yankee	1	Second	7	0.3

Table 10.1 Baseball team salaries

<sup>2</sup> This example is a hypothetical scenario based on actual data

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.756					
R Square	0.572					
Adjusted R Square	0.545					
Standard Error	4.204					
Observations	35					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	754	377	21.34	.0000	
Residual	32	66	18			
Total	34	1320				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3.895	1.555	-2.5	0.02	-7.062	-0.728
<i>Yankee</i>	6.306	1.429	4.4	0.0001	3.396	9.217
<i>runs</i>	0.104	0.020	5.1	0.0000	0.062	0.145

Table 10.2 Multiple regression of baseball salaries

From the regression output, our model is:

$$\hat{Salary}(M\$) = -3.90^a + 6.31^b Yankee + .104^b Runs$$

(1.56) (1.43) (0.020)

*RSquare: .57<sup>b</sup>*  
<sup>a</sup>*Significant at .02*  
<sup>b</sup>*Significant at .0001*

The coefficient estimate for the Yankee indicator is 6.31. The intercept for Yankees is 6.31 greater than the intercept for Marlins. The rookie can expect to earn \$6.31 million more if he signs with the Yankees.

His expected salary, with 40 runs last season, is:

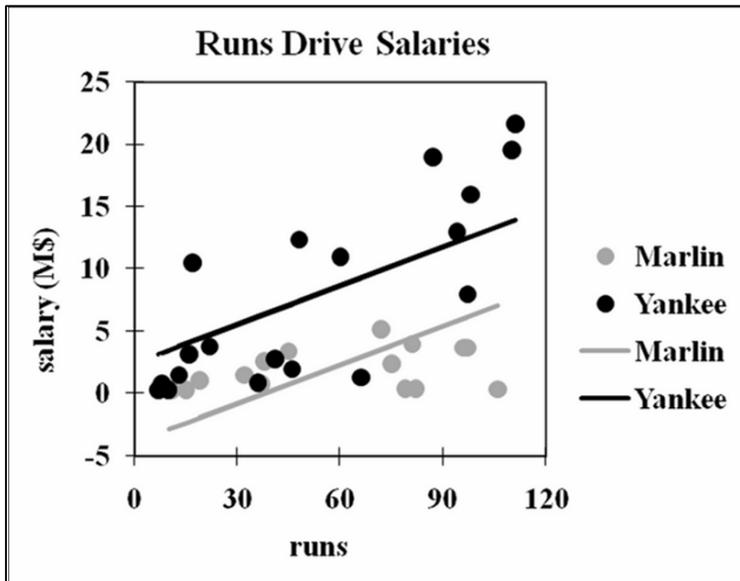
- As a Marlin, setting the *Yankee* indicator to zero:

$$\hat{Salary}(M\$) = -3.90 + .104(40) = -3.90 + 4.16 = .26(M\$) = \$260,000$$

- As a Yankee, setting the *Yankee* indicator to one:

$$\hat{Salary}(M\$) = -3.90 + 6.31 + .104(40) = 2.41 + 4.16 = 6.57(M\$) = \$6,570,000$$

The *Yankee* indicator modifies the intercept of the regression line, increasing it by 6.31.



In Figure 10.1, the intercept represents the baseline Marlins segment; the indicator adjusts the intercept to reflect the difference between Yankees and Marlins.

Figure 10.1 Yankees expect to earn \$6.31 million more

It does not matter which team is the designated baseline. We will get identical results either way.

## 10.2 Indicators Estimate the Value of Product Attributes

New product development managers sometimes use *conjoint analysis* to identify potential customers' most preferred new product design and to estimate the relative importance of product attributes. The conjoint analysis concept assumes that customers' preferences for a product are the sum of the values of each of the product's attributes, and that customers *trade off* features. A customer will give up a desired feature if another, more desired feature is offered.

*Example 10.3 New PDA Design.* As an example, consider preferences for PDAs. Management believes that customers choose PDAs based on desired size, design, keypad, and price. For a new PDA design, they are considering

- three sizes: bigger than shirt-pocket, shirt-pocket, and ultra thin shirt-pocket
- three designs: single unit, clamshell, and slider
- three keypads: standard, touch screen, and QWERTY
- three prices: \$150, \$250 and \$350

Management believes that price is a quality signal, and that customers suspect the quality of less expensive phones.

The least desirable, baseline configuration is expected to be:

bigger than shirt-pocket, single unit, with standard keypad at the lowest price.

To find the *part worth utilities*, or the value of each cell phone feature, indicators are used to represent features that differ from the baseline. The conjoint analysis regression model is:

$$\begin{aligned}
 PDA \text{ preference}_{ij} = & b_0 + b_{1i} \text{shirt-pocket size}_j + b_{2i} \text{ultra thin shirt-size}_j \\
 & + b_{3i} \text{clam shell}_j + b_{4i} \text{slider}_j \\
 & + b_{5i} \text{touch screen}_j + b_{6i} \text{QWERTY}_j \\
 & + b_{7i} \$250_j + b_{8i} \$350_j
 \end{aligned}$$

for the  $i$ 'th customer and the  $j$ 'th PDA configuration.

$b_0$  is the intercept, which reflects preference for the baseline configuration,  $b_{1i}$ ,  $b_{2i}$ ,  $b_{3i}$ ,  $b_{4i}$ ,  $b_{5i}$ ,  $b_{6i}$ ,  $b_{7i}$ , and  $b_{8i}$  are estimates of the *part worth utilities* of features to the  $i$ 'th customer.

The conjoint analysis process assumes that it is easier for customers to rank or rate products or brands, rather than estimating the value of each feature. For price preferences, this may be particularly true. It will be easier to customers to rate hypothetical PDA designs than it would be for customers to estimate the value of a \$250 PDA, relative to a \$150 PDA.

The four PDA attributes could be combined in 81 ( $=3^4$ ) unique ways. 81 hypothetical PDAs would be too many for customers to accurately evaluate. From the 81, a set of nine are carefully chosen so that the chance of each feature is equally likely (33%), and uncorrelated with other features. Slide designs, for example, are equally likely to be paired with each of the three sizes, each of the three keypads, and each of the three prices. This will minimize multicollinearity among the indicators used in the regression of the conjoint model. Such a subset of hypothetical combinations is an *orthogonal array* and is shown in Table 10.3.

<i>Size</i>	<i>Shape</i>	<i>Keypad</i>	<i>Price</i>
Bigger than shirt-pocket	Single unit	Standard	\$150
Bigger than shirt pocket	Clamshell	Touch screen	\$250
Bigger than shirt pocket	Slider	QWERTY	\$350
Shirt-pocket	Single unit	Touch screen	\$350
Shirt-pocket	Clamshell	QWERTY	\$150
Shirt-pocket	Slider	Standard	\$250
Ultra thin shirt-pocket	Single unit	QWERTY	\$250
Ultra thin shirt-pocket	Clamshell	Standard	\$350
Ultra thin shirt-pocket	Slider	Touch screen	\$150

Table 10.3 Nine hypothetical PDA designs in an orthogonal array

Three customers rated the nine hypothetical PDAs after viewing concept descriptions with sketches. The configurations judged extremely attractive were rated 9 and those judged extremely unattractive were rated 1. The regression with eight indicators is shown in Table 10.4.

<b><i>Regression Statistics</i></b>						
Multiple R	0.864					
R Square	0.747					
Adjusted R Square	0.634					
Standard Error	1.644					
Observations	27					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	8	143.3	17.9	6.6	0.0004	
Residual	18	48.7	2.7			
Total	26	192.0				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1.00	0.95	1.1	0.3061	-0.99	2.99
<i>shirt pocket</i>	0.78	0.78	1.0	0.3290	-0.85	2.41
<i>ultra thin shirt pocket</i>	1.89	0.78	2.4	0.0254	0.26	3.52
<i>clamshell</i>	-1.56	0.78	-2.0	0.0600	-3.18	0.07
<i>slider</i>	-1.44	0.78	-1.9	0.0788	-3.07	0.18
<i>touch screen</i>	4.22	0.78	5.4	0.0000	2.59	5.85
<i>QWERTY</i>	3.78	0.78	4.9	0.0001	2.15	5.41
<i>\$250</i>	1.67	0.78	2.2	0.0454	0.04	3.30
<i>\$350</i>	1.67	0.78	2.2	0.0454	0.04	3.30

Table 10.4 Regression of PDA preferences

PDA size, keypad, and price features influence preferences, while design does not. The preferred PDA is *ultra thin* and fits in a *shirt pocket*, with a *touch screen* or *QWERTY keypad*, and is priced at \$250 or \$350.

The *coefficients* estimate the part worth utilities of the PDA features. Expected preference for the ideal design the sum of the part worth utilities for feature included. We will assume an ultra thin PDA that fits in a shirt pocket, with the simplest single unit design, with a touch screen, at the highest price. Design does not affect preferences, so the least expensive option would be used, and the two higher prices are equivalent to customers, so the higher, more profitable price would be charged:

$$\begin{aligned}
 PDA\ preference_j &= 1.00 + 0.78\ shirtpocket_j + 1.89\ ultra\ thin\ shirtpocket_j \\
 &\quad - 1.56\ clamshell_j - 1.44\ slider_j \\
 &\quad + 4.22\ touch\ screen + 3.78\ QWERTY_j \\
 &\quad + 1.67\$250_j + 1.67\ \$350_j \\
 &= 1.00 + 0.78 (0) + 1.89 (1) \\
 &\quad - 1.56 (0) - 1.44 (0) \\
 &\quad + 4.22 (1) + 3.78 (0) \\
 &\quad + 1.67 (0) + 1.67 (1) \\
 &= 8.78
 \end{aligned}$$

The part worth utilities from coefficient estimates are shown in Figure 10.2 and Table 10.5.

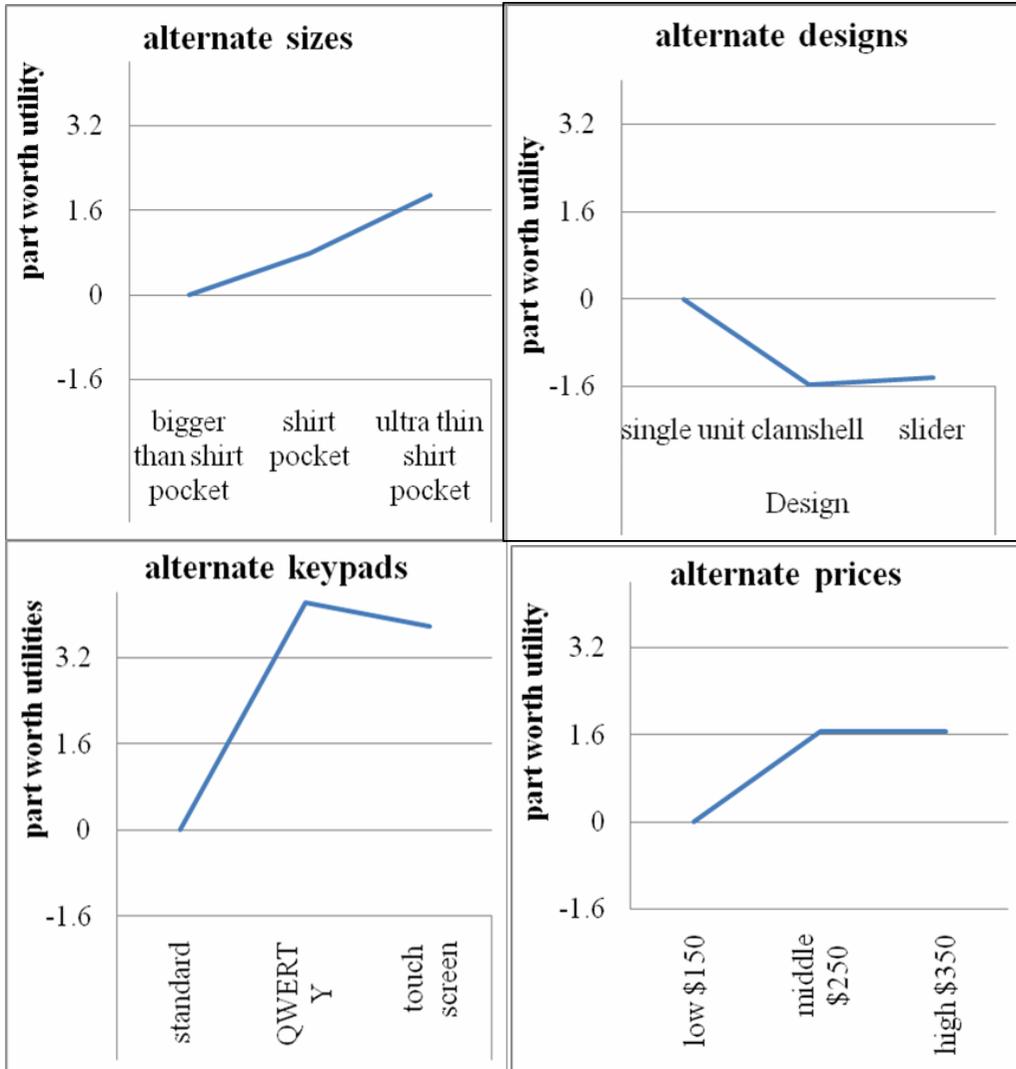


Figure 10.2 PDA part worth utilities

Preferred *ultraslim shirt pocket size* adds 1.89 (=1.89-0) to the preference ranking, a *touch screen* adds 4.22 (=4.22-0), and a price of \$250 adds 1.67 (=4.11-0). The preferred design makes no significant difference, 1.56 (=1.56-0).

The range in part worth utilities for each attribute is an indication of that attribute's importance. Preference depends most on the keypad configuration, which is more than twice as important as size or price.

<i>Attribute</i>	<i>Part worth utility of least preferred</i>	<i>Part worth utility of most preferred</i>	<i>Part worth utility range</i>	<i>Attribute importance</i>
<i>Size</i>	0	1.89	1.89	1.89/9.34= .20
<i>Shape</i>	-1.56	0	1.56	1.56/9.34= .17
<i>keypad</i>	0	4.22	4.22	4.22/9.34= .18
<i>price</i>	0	1.67	1.67	1.67/9.34= .45
<i>Sum of part worth utility ranges:</i>			9.34	

Table 10.5 Relative importance of PDA attributes

Conjoint analysis been used to improve the designs of a wide range of products and services, including:

- seating, food service, scheduling and prices of airline flights
- offer of outpatient services and prices for a hospital
- container design, fragrance and design of a aerosol rug cleaner,
- digital camera pixels, features and prices

Conjoint analysis is versatile and the attributes studied can include characteristics that are difficult to describe, such as fragrance or taste. It is difficult for customers to tell us how important color, package design, or brand name are in shaping preferences, and conjoint analysis often provides believable, valid estimates.

### 10.3 Indicators Quantify Seasonality in Time Series

*Example 10.4 Tyson's Farm Worker Forecast*<sup>3</sup>. Tyson's Management would like to forecast quarterly self employed workers in agriculture. They believe that these self employed workers, family farmers, are leaving the farm to find more profitable work elsewhere. Tyson's meet labor demand left unsatisfied by hiring agricultural workers. They have asked Mark, their master model builder, to build a model to forecast quarterly self employed agriculture workers. In months where the number of workers is expected to be down from the prior year, they will hire additional workers. If these gaps are large enough, they will implement a lobbying campaign to lesson restrictions on illegal immigrant workers who would work for lower wages.

**Choice of the first predictor.** Since Mark was working with a time series, he first chose a logically appealing leading indicator of *self employed agricultural workers: unpaid family workers* in agriculture. Self employed farmers often relied on unpaid family members. If unpaid family workers were leaving agriculture to work in paid jobs elsewhere, this might drive self-employed workers to leave agriculture the following year.

<sup>3</sup> This example is a hypothetical scenario based on actual data

Both segments of workers probably fluctuated with other economic indicators, so Mark began with a single predictor to minimize multicollinearity issues.

**Choice of lag.** In order to forecast *self employed ag workers* from *unpaid family ag workers*, Mark needed to lag the leading indicator. To confirm that twelve months was the appropriate lag for wage and salary workers, he plotted *self employed ag workers* and *unpaid family ag workers*, using 33 months of data from the Bureau of Labor, June, 2004 through April 2007, shown in Figure 10.3.

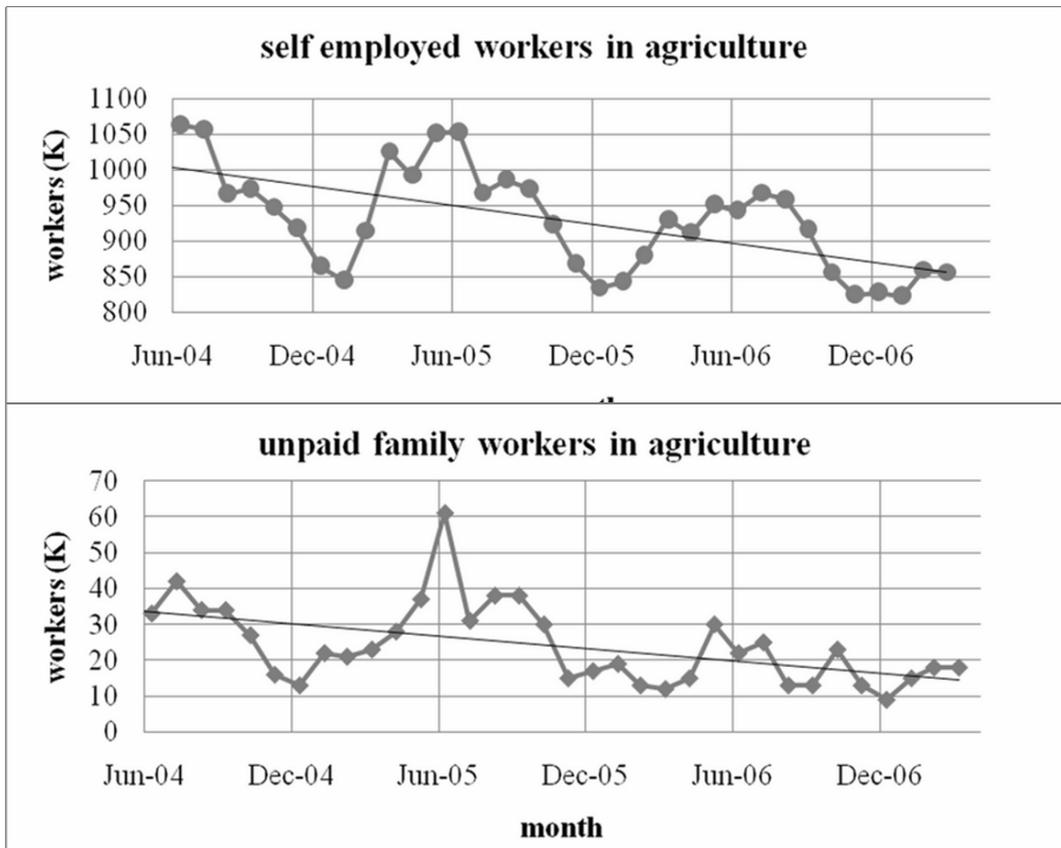


Figure 10.3 Self Employed and Unpaid Workers in Agriculture, June 04 through April 07

Both the number of self-employed workers and unpaid family workers were decreasing. The scatterplots confirm that agricultural labor follows an annual cycle that corresponds to planting and harvesting cycles. Since twelve months is the traditional growing cycle in agriculture, Mark chose a twelve month lag for the regression model, which is below. He hid the two most recent observations, March and April 2007, to later validate the model, since he wanted to be sure that his model could be relied upon to produce solid forecasts.

## SUMMARY OUTPUT

<i>Regression Statistics</i>						
Multiple R	0.560					
R Square	0.313					
Adjusted R Square	0.291					
Standard Error	61.9					
Observations	33					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	54160	54160	14.1	0.0007	
Residual	31	118807	3832			
Total	32	172967				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	837.2	28.4	29.5	0.0000	779.3	895.1
<i>unpaid family workers q-12</i>	3.63	0.97	3.8	0.0007	1.66	5.61
<i>DW:</i>	0.652					

Table 10.6 Regression of self-employed workers in agriculture

The model, shown in Table 10.6, is significant (*Significance F* = .0007), the *RSquare* is low for time series data, .31, and the standard error, 61.9K workers, is large. The coefficient estimate is positive as expected: *self-employed workers* leave agriculture following the exit of *unpaid family workers*.

**Assessment of autocorrelation.** Since time series often contain trend, cycles, and seasonality, those must be accounted for. If these systematic variations in the data are present, but unaccounted for, they will be present in the model residuals. The Durbin Watson statistic will identify presence of unaccounted for trend, cycles, or seasonality in the residuals. Mark found that the residuals are autocorrelated ( $DW=.65 < dL_{33,2}=1.38$ ). Trend, cycles or seasonality are present in the data and have not been accounted for. Mark plotted the residuals in Figure 10.4 to identify potential trend, cycle or seasonality variables.

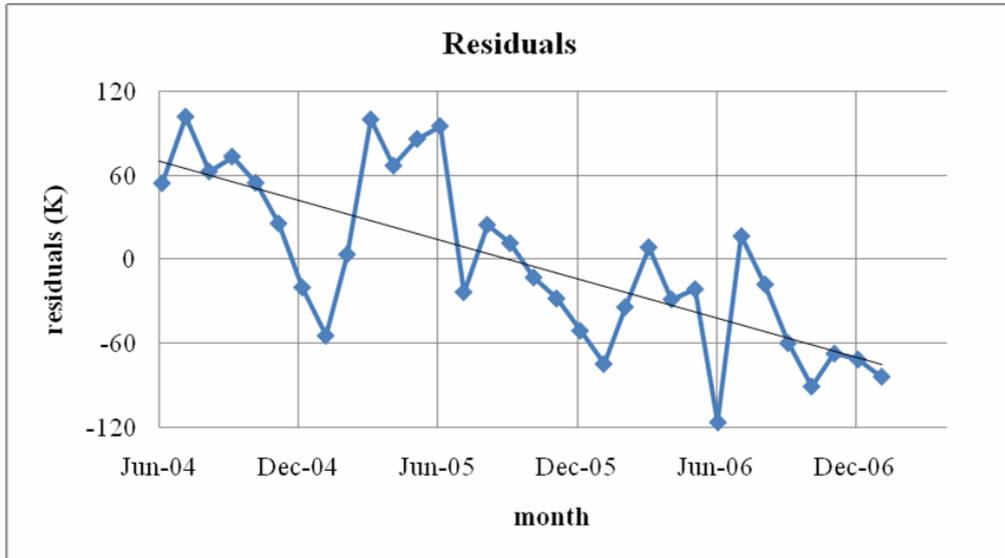


Figure 10.4 Residuals are not pattern free

There is an obvious trend in the residuals, suggesting that the decline in self employed workers was more severe than the decline in unpaid family workers. Mark would add a trend component, *month since May 2004* (equal to one in the first month of the series).

There is also obvious seasonality in the residuals. In Winter months, residuals tend to be negative. Mark would add a Winter indicator to the model. To decide which months to include in the Winter season, Mark made a PivotChart of residuals by month, which is below.

The residuals, shown in Figure 10.5, were lower in Winter months November through March, indicating that the number of self-employed workers were lower in these months. The *Winter* indicator variable would be equal to one in months November, December, January, February, and March, and it would be equal to zero in other months.

The expanded regression model, with a trend component, *month since May 2004*, and a seasonality indicator, *Winter*, is shown in Table 10.7.

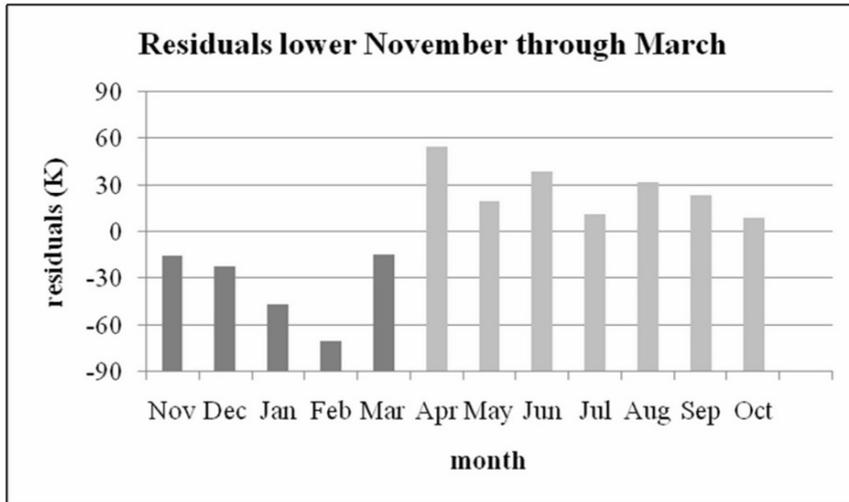


Figure 10.5 Residuals are lower in Winter months November through March

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.919
R Square	0.845
Adjusted R Square	0.828
Standard Error	30.5
Observations	33

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	146072	48691	52.5	0.0000
Residual	29	26895	927		
Total	32	172967			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	993.8	22.7	43.8	0.0000	947.4	1040.2
<i>unpaid family m-12</i>	1.46	0.66	2.2	0.0339	0.12	2.81
<i>month since 5/04</i>	-3.71	0.59	-6.3	0.0000	-4.91	-2.52
<i>Winter</i>	-81.2	15.0	-5.4	0.0000	-111.9	-50.4

DW : 1.83

Table 10.7 Regression with Trend and Seasonality Indicator

*RSquare* is now much higher, .85, and the standard error is now much smaller. Forecasts can be expected to fall within 61K (=2\*30.5K) workers.

The coefficient signs are as Mark expected. The number of self employed workers follows the number of unpaid family workers a year later, though the decline in self employed workers is more severe. About 3.7K more self employed workers leave agriculture each month.

There is significant seasonality in the self-employed labor market. About 81K fewer self-employed work in agriculture in the Winter months. These workers apparently farm during warmer months and work at other jobs outside agriculture in Winter months.

The residuals are now free of autocorrelation. *DW* is 1.83, which exceeds  $dU_{33,4}=1.65$  for this sample of 33 months and a model with four variables, including intercept.

**Model validity.** To assess the model's validity, Mark compared the two most recent, hidden observations with the 95% mean prediction intervals, shown in Table 10.8.

<i>month</i>	<i>95% lower prediction</i>	<i>self-employed workers (K)</i>	<i>95% upper prediction</i>
Mar-07	743	859	868
Apr-07	819	856	944

Table 10.8 Model Validation

The model correctly predicts the number of self-employed workers in the two most recent months. With this evidence of model validity, Mark recalibrated the model by adding these two most recent months, which had been hidden to build the model and validate. The model becomes:

$$\hat{\text{Self employed workers}}(K)_q = 989^a - 75.4^a \text{Winter}_q + 1.54^a \text{unpaid family workers}(K)_{q-12} - 3.64^a q$$

*RSquare*: .84  
<sup>a</sup>significant at .01.

In months April through October, setting the *Winter* indicator to 0, the expected number of *self employed workers* in agriculture is:

$$\begin{aligned} \hat{\text{Self employed workers}}(K)_q &= (989 - 75.4(0)) + 1.54 \text{unpaid family workers}(K)_{q-12} - 3.64 q \\ &= 989 + 1.54 \text{unpaid family workers}(K)_{q-12} - 3.64 q \end{aligned}$$

In months November through March, the *Winter* indicator is 1, and the expected number of *self-employed workers* is:

$$\begin{aligned} \text{Self employed workers}(K)_q &= (989 - 75.4(1)) + 1.54 \text{ unpaid family workers}(K)_{q-12} - 3.64 q \\ &= 914 + 1.54 \text{ unpaid family workers}(K)_{q-12} - 3.64 q \end{aligned}$$

The *Winter* indicator shifts the regression intercept and line down by 75(K) workers, as Figure 10.6 illustrates.

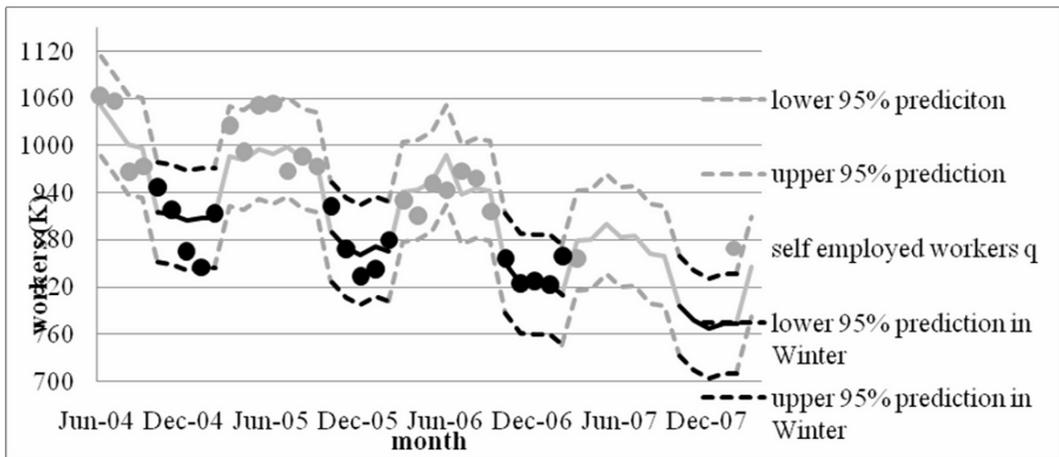


Figure 10.6 Self employed workers are leaving agriculture

Mark would report to Management:

**MEMO**

**Re: Declining Supply of Self Employed Agriculture Workers**

**To:** Tyson Directors of Planning and Legal Affairs

**From:** Mark Weisselburg, Director, Econometric Forecasting and Analysis

**Date:** April 2007

Analysis of workers in agriculture from June 2004 to April 2007 reveals that self employed workers are leaving agriculture.

**Econometric Model.** Using 35 months of data on self employed and unpaid family workers in agriculture from the Bureau of Labor. A model of self employed workers was built from 33 months and correctly forecast the two most recent months.

**Model Results.** Trend, seasonality, and variation in past year unpaid family workers in agriculture account for 84% of the variation in monthly self employed workers. The model forecast margin of error is less than 62 thousand workers.

Following a decline of 1,000 unpaid family workers, the number of self employed workers is expected to decline by as many as 3,000 the following year.

A negative trend in self employed workers is forecast: each month 3,000 to 5,000 self-employed are expected to exit.

A larger number, 50,000 to 100,000, leave during Winter months, but return in warmer weather.

Forecasts are:



*April through October:*

$$\text{Self employed } (K)_q = 989^a - 3.6^a q + 1.5^a \text{ unpaid family } (K)_{q-12}$$

*November through March*

$$\text{Self employed } (K)_q = 924^a - 3.6^a q + 1.5^a \text{ unpaid family } (K)_{q-12}$$

*RSquare: .84<sup>a</sup>*

*<sup>a</sup>Significant at .01.*

Month	M-7	J-7	J-7	A-7	S-7	O-7	N-7	D-7	J-8	F-8	M-8	A-8
<i>lower</i>	820	840	820	820	800	800	730	710	700	710	710	780
<i>upper</i>	940	960	950	950	930	920	860	840	830	840	840	910

**Conclusions.** The number of self employed agriculture workers is expected to continue a stable decline, providing an opportunity for Tyson to assume a greater level of leadership in farming by pressing for legislation to facilitate a greater supply of immigrant labor.

**Other factors.** The pool of wage and salary workers, a potentially driving influence was not considered here.

## 10.4 Indicators Add Structural Shifts in Time Series

Economic and business performance adapts to shocks, such as 911, and structural shifts, such as changes in national leadership. Indicators allow us to incorporate shocks or structural shifts, turning on and off economic or political environments within a time series.

*Example 10.5 Leadership Changes Influence US Imports by India.* US imports by India are growing each year with India's rapidly growing economy. The growing wealth creates growing demand, some of which is satisfied with US products.

The level of international trade between India and the U.S. probably depends upon the political leadership in place. In the past twenty years, political leadership in India has shifted back and forth between the Congress and BJP Parties. The structure of trade practices is influenced by leadership. To represent party leadership, we will include an indicator: *Congress*, representing one of the two dominant parties in India. The baseline leadership until mid-1991 was under the *BJP* Party. When BJP was in power, the indicator *Congress* will equal zero, and when leadership shifts to *Congress*, the indicator becomes one.

We will build a model of India's Imports from the U.S. during the past twenty-one years, 1985 through 2005, which incorporates the 0-1 indicator of the structural shifts due to the party leadership and the effects of a leading indicator, *Indian per capita GDP*. Data is from Ward's Communications and the International Monetary Fund.

**Expected response patterns.** India's imports of U.S. products are growing with India's increasing wealth. Following good modeling practice, we will exclude the two most recent years to later validate our model. Regression results are in Table 10.9.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.930					
R Square	0.865					
Adjusted R Square	0.848					
Standard Error	0.381					
Observations	19					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	14.86	7.43	51.2	0.0000	
Residual	16	2.32	0.15			
Total	18	17.19				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3.12	0.61	-5.1	0.0001	-4.42	-1.82
Congress Party	0.58	0.21	2.7	0.0161	0.12	1.03
Past year GDP per capita (\$K)	15.67	1.56	10.1	0.0000	12.37	18.98
<i>DW: 1.81</i>						

Table 10.9 Indian imports are driven by party and per capita wealth

The model is significant, coefficient signs are positive, as expected, and the residuals are free from significant autocorrelation ( $DW=1.81 > dU_{19,3}=1.54$ ).

The model produces valid forecasts:

year	<i>lower 95% prediction</i>	<i>Indian imports (\$B)</i>	<i>upper 95% prediction</i>
2004	5.24	6.11	6.85
2005	6.40	7.96	8.02

Following recalibration, the model becomes:

$$\text{Indian imports } (\$B)_q = -3.56^a + .67^a \text{Congress}_q + 16.8^a \text{GDP per capita } (K\$)_{q-1}$$

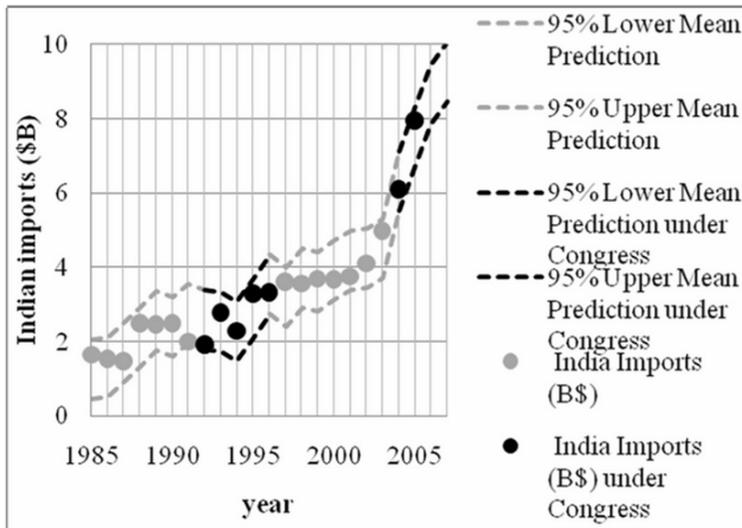
*RSquare: .95*

*<sup>a</sup>significant at .01.*

Relative to the baseline leadership under the BJP Party, Indian imports of US products have been higher under Congress leadership.

$$\begin{aligned}
 \text{Indian imports } (\$B)_q &= -3.56^a + .67^a \text{Congress}_q + 16.8^a \text{GDP per capita}(K\$)_{q-1} \\
 &= -3.56 + .67(0) + 16.8 \text{GDP per capita}(K\$)_{q-1} \\
 &= -3.56 + 16.8 \text{GDP per capita}(K\$)_{q-1} && \text{under BJP leadership,} \\
 &= -3.56^a + .67^a \text{Congress}_q + 16.8^a \text{GDP per capita}(K\$)_{q-1} \\
 &= -3.56 + .67(1) + 16.8 \text{GDP per capita}(K\$)_{q-1} \\
 &= -2.89 + 16.8 \text{GDP per capita}(K\$)_{q-1} && \text{under Congress leadership,}
 \end{aligned}$$

Forecasts are shown in Figure 10.6, with Congress leadership shown in black and BJP leadership shown in gray. Leadership under the Congress Party produces an expected increase of \$.67B each year.



Management gauging export potential to the Indian market would conclude:

*“Indian imports of U.S. products are growing at increasing rates, as India’s economic productivity rate of population growth increases.”*

Figure 10.6 Forecast of India’s imports under alternate party leadership

*Variation in India’s population growth rate and political leadership in India account for 95% of the variation in India’s imports of U.S. products.*

*Imports increase during years when the Congress Party is in power in India. Annually, Indian imports of U.S. products are expected to be about \$.67B more under the current leadership of the Congress Party than if BJP to replace that current leadership.”*

## 10.5 Indicators Allow Comparison of Segments and Scenarios And Quantify Structural Shifts

Indicators allow us to adjust the intercept in linear models to allow for differences in average levels of diverse segments or scenarios. Incorporating indicators in time series models allows us to gauge the impact of structural shifts and to estimate response levels that would have manifested had shocks not occurred. Similarly, if a shock is expected to recur, we can set its indicator to one in future periods to forecast the expected change should the shock occur again.

Indicators are used to analyze conjoint analysis data, and estimate the part worth utilities, or the value of each product feature. The part worth utility estimates enable new product development managers to identify most preferred product designs and the most important attributes driving preferences.

## Excel 10.1 Use indicators to find part worth utilities and attribute importances from conjoint analysis data

Three customers from the target market rated nine hypothetical PDA designs, shown in **Table 10.3**, using a scale from 1 (=least preferred) to 9 (=most preferred). This data is in **Excel 10.1 PDA conjoint.xls**.

Use indicators to estimate the part worth utilities of *size*, *shape*, *keypad* and *price* attribute options for PDAs.

**Baseline hypothetical.** The baseline PDA is

*bigger than shirt pocket*, with *single unit* design, *standard* keypad, at a retail price of *\$150*.

The first hypothetical PDA design in **Table 10.3**, and in rows **2**, **11**, and **20** of the file, corresponds to the baseline.

**Add indicators for differences from baseline.** Add four indicators, two for each PDA attribute, in **G** through **N**: *shirt pocket*, *ultra slim shirt pocket*, *clamshell*, *slider*, *QWERTY*, *touch screen*, *\$250*, and *\$350*. Enter a zero or a one in each of these columns for each of the nine hypotheticals.

The baseline hypothetical, for example, will have zeros in all eight columns, since it is not *shirt pocket* or *ultra slim shirt pocket size*, it does not feature a *clamshell* or *slider design*, it does not have a *QWERTY* or *touch screen keypad*, and it is not priced at *\$250* or *\$350*:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	customer	hypothetical	size	design	key pad	price	rating	shirt pocket	ultra thin shirt pocket	clamshell	slider	touch screen	QWERTY	\$250	\$350				
2	1	1	bigger than shirt pocket	single unit	standard	\$150	1	0	0	0	0	0	0	0	0				
3	1	2	bigger than shirt pocket	clamshell	touch screen	\$250	5	0	0	1	0	1	0	1	0				
4	1	3	bigger than shirt pocket	slider	QWERTY	\$350	5	0	0	0	1	0	1	0	1				
5	1	4	shirt pocket	single unit	touch screen	\$350	7	1	0	0	0	1	0	0	1				
6	1	5	shirt pocket	clamshell	QWERTY	\$150	3	1	0	1	0	0	0	1	0				
7	1	6	shirt pocket	slider	standard	\$250	3	1	0	0	1	0	0	1	0				
8	1	7	ultra thin shirt pocket	single unit	QWERTY	\$250	8	0	1	0	0	0	1	1	0				
9	1	8	ultra thin shirt pocket	clamshell	standard	\$350	5	0	1	1	0	0	0	0	1				
10	1	9	ultra thin shirt pocket	slider	touch screen	\$150	9	0	1	0	1	1	0	0	0				

Use shortcuts to copy and paste the indicator values for the nine hypotheticals into rows **11** through **28**:

Select **H2:O10**, **Cntl+C**, then  
 Select **H11 [Enter]**, **Cntl+C**, select **H20 [Enter]**.

Run a regression of *rating*, with **Input Y Range G1:G28** and the eight indicators, with **Input X Range H1:O28**, labels:

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.864								
R Square	0.747								
Adjusted R Sq	0.634								
Standard Error	1.644								
Observations	27								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	8	143.3	17.9	6.6	0.0004				
Residual	18	48.7	2.7						
Total	26	192.0							
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	1.00	0.95	1.1	0.3061	-0.99	2.99	-0.9945	2.99448	
shirt pocket	0.78	0.78	1.0	0.3290	-0.85	2.41	-0.8507	2.40626	
ultra thin shirt p	1.89	0.78	2.4	0.0254	0.26	3.52	0.26041	3.51737	
clamshell	-1.56	0.78	-2.0	0.0600	-3.18	0.07	-3.184	0.07293	
slider	-1.44	0.78	-1.9	0.0788	-3.07	0.18	-3.0729	0.18404	
touch screen	4.22	0.78	5.4	0.0000	2.59	5.85	2.59374	5.85071	
QWERTY	3.78	0.78	4.9	0.0001	2.15	5.41	2.14929	5.40626	
250	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515	
350	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515	

The regression is significant, and RSquare is .75, suggesting that the feature differences among the PDA hypotheticals account for 75% of the variation in preferences. The standard error is 1.6 on the 9-point rating scale, making the margin of error in model predictions about 3.2 on the 9-point scale.

**Part worth utilities.** The *coefficients* are estimates of the part worth utilities, or the value of each feature. Size, price, and keypad options drive preferences, while design options do not. The most preferred PDAs would be ultrathin shirt pocket size, with a touch screen or QWERTY keypad, at a price of \$250 or \$350.

To find the *expected rating of the ideal design*, add the coefficients corresponding to these features. For an ultrathin shirt pocket size, single unit, with touchscreen at \$350

in **J25**, enter **=SUM(B16,B18,B21,B24)** [Enter]:

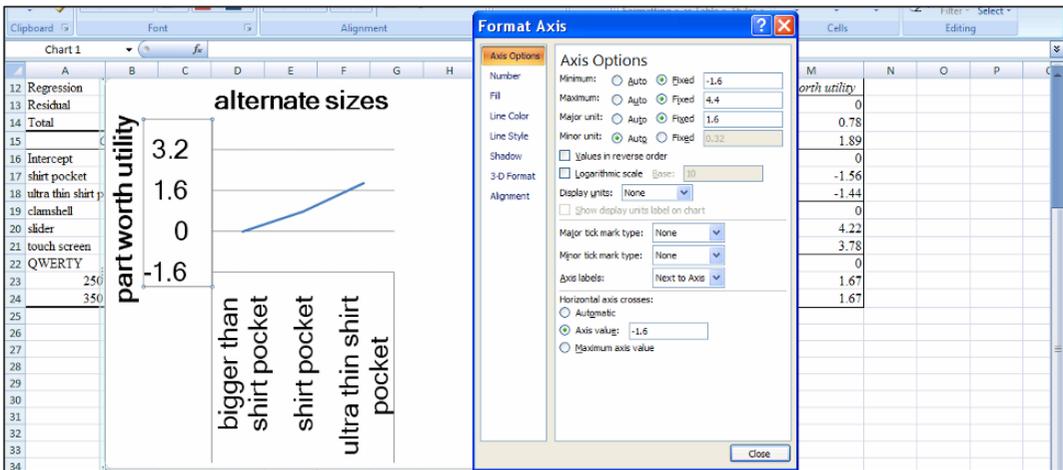
=SUM(B16,B18,B21,B24)									
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	1.00	0.95	1.1	0.3061	-0.99	2.99	-0.9945	2.99448	
shirt pocket	0.78	0.78	1.0	0.3290	-0.85	2.41	-0.8507	2.40626	
ultra thin shirt p	1.89	0.78	2.4	0.0254	0.26	3.52	0.26041	3.51737	
clamshell	-1.56	0.78	-2.0	0.0600	-3.18	0.07	-3.184	0.07293	
slider	-1.44	0.78	-1.9	0.0788	-3.07	0.18	-3.0729	0.18404	
touch screen	4.22	0.78	5.4	0.0000	2.59	5.85	2.59374	5.85071	
QWERTY	3.78	0.78	4.9	0.0001	2.15	5.41	2.14929	5.40626	
250	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515	
350	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515	8.78

**Attribute importances.** To find the attribute importances, first plot the part worth utilities for each attribute. In the regression sheet, enter the four attributes in **K**, the attribute options in **L**, and the part worth utilities in **M**. (Part worth utilities for baseline options are zero):

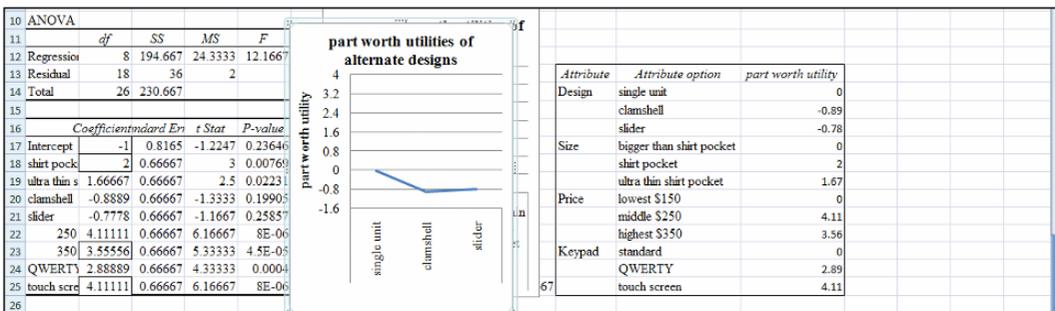
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
12	Regression	8	143.3	17.9	6.6	0.0004					Attribute	Attribute option	part worth utility			
13	Residual	18	48.7	2.7							Size	bigger than shirt pocket	0			
14	Total	26	192.0									shirt pocket	0.78			
15												ultra thin shirt pocket	1.89			
16	Intercept	1.00	0.95	1.1	0.3061	-0.99	2.99	-0.9945	2.99448		Design	single unit	0			
17	shirt pocket	0.78	0.78	1.0	0.3290	-0.85	2.41	-0.8507	2.40626			clamshell	-1.56			
18	ultra thin shirt p	1.89	0.78	2.4	0.0254	0.26	3.52	0.26041	3.51737			slider	-1.44			
19	clamshell	-1.56	0.78	-2.0	0.0600	-3.18	0.07	-3.184	0.07293		Keypad	standard	0			
20	slider	-1.44	0.78	-1.9	0.0788	-3.07	0.18	-3.0729	0.18404			touch screen	4.22			
21	touch screen	4.22	0.78	5.4	0.0000	2.59	5.85	2.59374	5.85071			QWERTY	3.78			
22	QWERTY	3.78	0.78	4.9	0.0001	2.15	5.41	2.14929	5.40626		Price	\$150	0			
23		250	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515			\$250	1.67		
24		350	1.67	0.78	2.2	0.0454	0.04	3.30	0.03818	3.29515			\$350	1.67		

To see the part worth utilities for alternate sizes, select **K12:M15**, and use shortcuts to make a line plot: **Alt NN**:

To compare across attributes, reformat the vertical y axis. Select the axis, right click, **Format Axis**, and set **Minimum** to -1.6, **Maximum** to 4.4, and **Major unit** to .8.

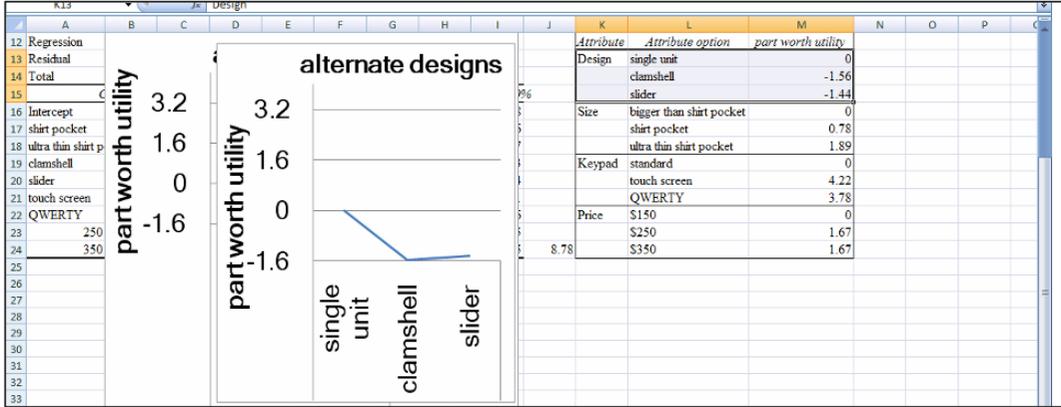


Set **Horizontal axis crosses** at -1.6:

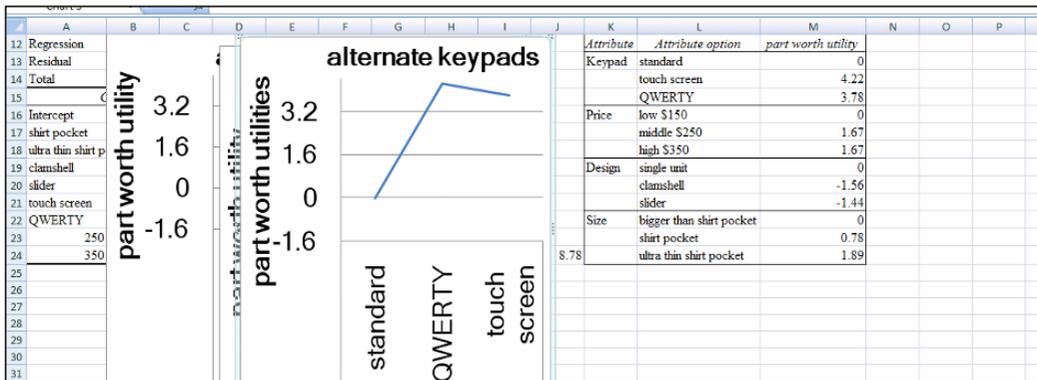


To see part worth utilities for alternate designs, use shortcuts to move the design cells up to rows 13 through 15: select **K16:M18**, **Cntd+X**, select **K13**, **Alt HIE**.

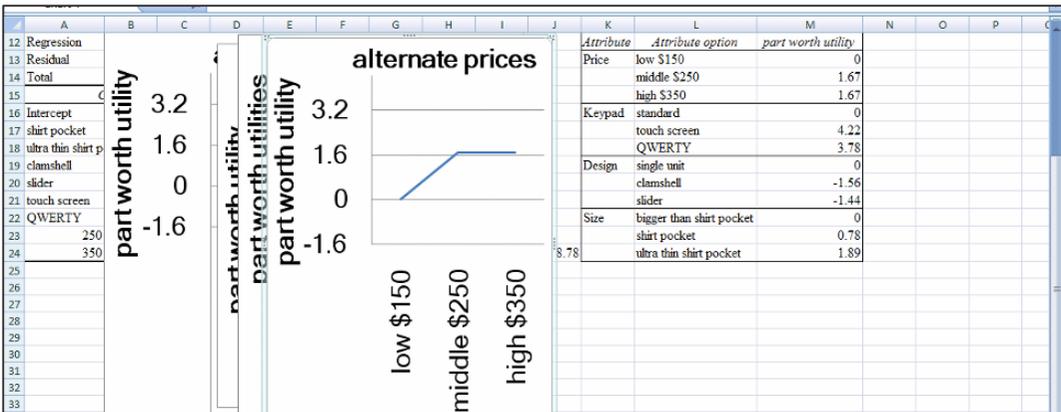
Plot the *Design* part worth utilities: select **K13:M16**, **Alt NN**. Reformat the y axis:



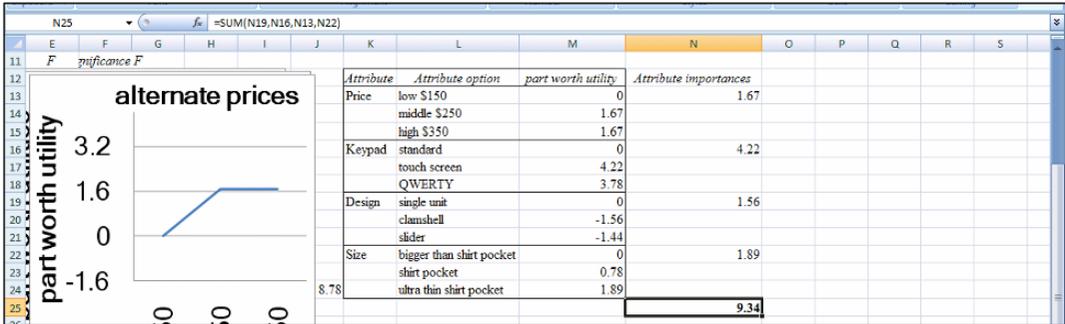
Move the *Keypad* cells up, plot, and reformat the y axis:



Add *lowest*, *middle*, and *highest* to the *price* options so that Excel will treat these cells as categories. Then move the *price* cells to **K14:M16**, plot, and reformat the y axis:



Find the attribute importances in N13, N16, N19, and N22. The importance of each attribute is the difference between the most and least preferred *attribute options*:

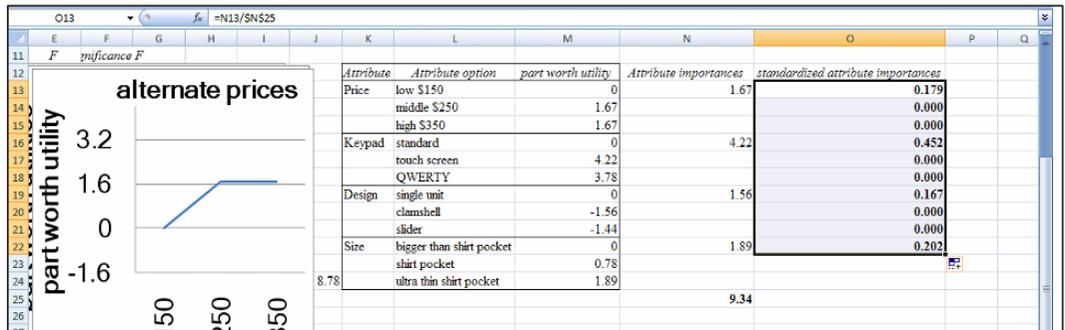


Find the *standardized attribute importances* in O. First sum four attribute importances, N14, N17, N20 and N23.

In N26 enter =SUM(N14,N17,N20,N23)[Enter].

Standardize the attribute importances by dividing by the sum.

In O13, enter =N13/\$N\$25 [Enter]. Select this new cell, grab, drag through O22:



Keypad is more than twice as important as size (.452/.202>2), and design is relatively unimportant.

### Excel 10.2 Add indicator variables to account for segment differences or structural shifts

*Indian Imports of U.S. Products.* We will build a model of India’s annual imports of U.S. products, using time series. A leading indicator of India’s economic productivity and political leadership are thought to drive imports. Party leadership alters import policies and is likely to affect India’s imports of U.S. products.

Data including time series *year*, *Indian Imports(B\$)* and *Indian GDP per capita (\$K)* are in **Excel 10.2 Indian Imports.xls**.

**Add Party leadership indicators.** To represent India’s political leadership, the earliest period of leadership under the BJP Party will be our baseline. To see how imports have differed under leadership of the alternate Congress Parties, add a *Congress* indicator variable:

In **D1** type in the label *Congress*, enter **0** in **D2**, select the new cell and double click to fill in column with zeros.

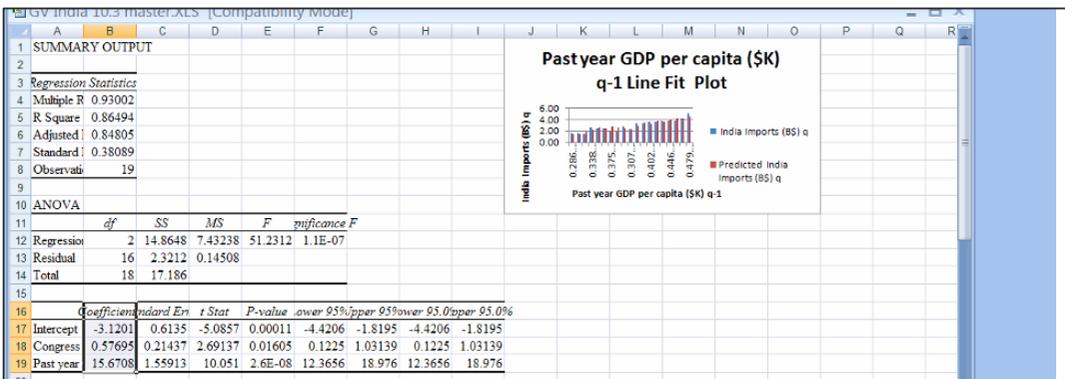
In years of *Congress* leadership, 1992-1996 and 2004-2007, **D9:D13** and **D21:D24**, change zeros to ones.

Year	India Imports (B\$) q	Past year GDP per capita (\$K) q-1	Congress Party
1985	1.64	0.28619	0
1986	1.54	0.29053	0
1987	1.46	0.31314	0
1988	2.50	0.33819	0
1989	2.46	0.36321	0
1990	2.49	0.35425	0
1991	2.00	0.37509	0
1992	1.92	0.32627	1
1993	2.78	0.32151	1
1994	2.29	0.30725	1
1995	3.30	0.34299	1
1996	3.33	0.38208	1
1997	3.61	0.4024	0

The indicator *Congress* will modify the baseline intercept, quantifying differences in the level of Indian imports from the baseline leadership under BJP. In our regression, the indicator will come first in the set of predictors, because it modifies the intercept.

Use shortcuts to rearrange the columns so that the indicator precedes the continuous predictor, *Indian GDP per capita*: Select **D**, **Cntl+X**, select **C**, **Alt HIE**.

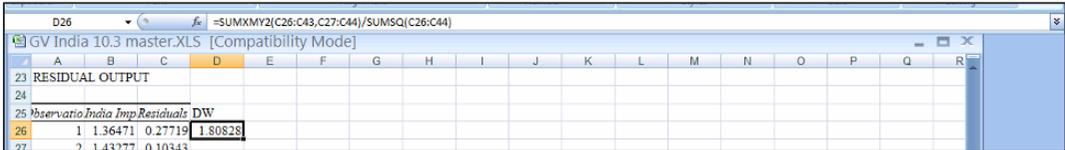
Run a regression, excluding the two most recent years, 2004 and 2005, with **B1:B20** for the **Input Y Range** and **C1:D20** for the **Input X Range**. (The two most recent years are excluded, since we want to test the model’s validity for reliable forecasting.)



The model is significant, and the coefficient sign for per capita GDP is positive, as expected. The standard error is \$.38B, making the forecast margin of error approximately \$.76B.

**Assess autocorrelation.** Since we are working with a time series, we must confirm that trend, cycles, and seasonality have been accounted for with the leading indicator. Find *DW* on the regression sheet.

Enter =SUMXMY2(C27:C44,C28:C45)/SUMSQ(C27:C45)[Enter].



The Durbin Watson statistic is 1.81.

Find the online tabled  $dU_{19,3}$ , and confirm that  $DW > dU$ .

We conclude that the residuals are free of unaccounted for trend or cycles.

**Model validation.** To test the model’s validity, select and copy the coefficient estimates **B16:B19** and paste them into the *Indian imports* worksheet **E1:E4**.

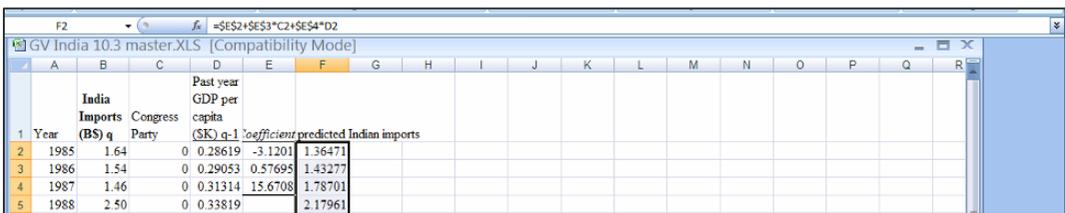
Use the regression equation to find *predicted Indian imports* in column **F**:

$$\text{Indian imports}_q = -3.12 + .58 \text{ Congress Party}_q + 15.7 \text{ GDP per capita}_{q-1}$$

In **F2** enter =E2 f4 + E3 f4 \* C2+E4 f4 \*D2 [Enter].

(f4 is the Excel function which locks the row and column of the coefficients in your equation so that as Excel moves through each row to find predicted *imports* from *Congress* and *past year GDP per capita* it uses the *coefficients* in rows **2** through **4**.)

Select the new cell, grab and drag through row **24**:



Copy the regression *standard error* in **B7** and paste into **G2**.

In **H2**, find the *t* value for 15 residual degrees of freedom: =TINV(.05, 15) [Enter].

Find the 95% *lower* and *upper prediction intervals* in **I** and **J**, by subtracting and adding *t* in **H2** x the *standard error* in **G2** from *predicted* values in **F**.

In **I2** enter =F2-H2 f4\*G2 f4 [Enter].

In **J2** enter =F2+H2 f4\*G2 f4 [Enter].

Select the two new cells, grab, and drag through row **24**:

Year	India Imports (BS) q	Congress Party	Past year GDP per capita (\$K) q-1	coefficient	predicted Indian imports	standard error	t	lower 95% prediction	upper 95% prediction
1985	1.64	0	0.28619	-3.1201	1.36471	0.38089	2.11991	0.557267	2.1721571
1986	1.54	0	0.29053	0.57695	1.43277			0.625325	2.2402152
1987	1.46	0	0.31314	15.6708	1.78701			0.979562	2.5944528
1988	2.50	0	0.33819		2.17961			1.372162	2.9870524

Confirm that the model is valid by comparing actual *Indian imports* in 2004 and 2005 in **B21:B22** with the *95% prediction intervals* for 2004 and 2005 in **I21:J22**:

Year	India Imports (BS) q	Congress Party	Past year GDP per capita (\$K) q-1	coefficient	predicted Indian imports	standard error	t	lower 95% prediction	upper 95% prediction
1998	3.56	0	0.43214		3.65186			2.844414	4.4593048
1999	3.69	0	0.42671		3.56685			2.759401	4.374291
2000	3.67	0	0.4465		3.87686			3.069415	4.6843056
2001	3.76	0	0.46003		4.08892			3.281472	4.8963624
2002	4.10	0	0.46455		4.15975			3.352304	4.9671942
2003	4.98	0	0.47955		4.39479			3.58735	5.20224
2004	6.11	1	0.54802		6.04478			5.24	6.85
2005	7.96	1	0.62241		7.21056			6.40	8.02
2006		1	0.68665		8.21714			7.409695	9.024585

Recalibrate by running the regression adding the two most recent rows **21** and **22** with years 2004 and 2005:

**Regression Statistics**

Multiple R	0.97358
R Square	0.94786
Adjusted R Square	0.94207
Standard Error	0.37997
Observations	21

**ANOVA**

	df	SS	MS	F	Significance F
Regression	2	47.2456	23.6228	163.619	2.8E-12
Residual	18	2.59878	0.14438		
Total	20	49.8444			

**Coefficients**

	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-3.5563	0.39259	-9.0587	4E-08	-4.3811	-2.7315
Congress	0.67138	0.17689	3.79552	0.00132	0.29976	1.04301
GDP per capita	16.795	0.9773	17.1851	1.3E-12	14.7418	18.8483

From model results, we can write the regression equation:

$$Indian\ imports\ (\$B)_q = -3.56^a + .67^a Congress_q + 16.8^a GDP\ per\ capita_{q-1}$$

Which becomes

- During the baseline years of BJP leadership:

$$\begin{aligned} \text{Indian imports } (\$B)_q &= -3.56^a + .67^a(0) + 16.8^a \text{GDP per capita}_{q-1} \\ &= -3.56 \qquad \qquad \qquad + 16.8^a \text{GDP per capita}_{q-1} \end{aligned}$$

- During Congress leadership:

$$\begin{aligned} \text{Indian imports } (\$B)_q &= -3.56^a + .67^a(1) + 16.8^a \text{GDP per capita}_{q-1} \\ &= -2.89 \qquad \qquad \qquad + 16.8^a \text{GDP per capita}_{q-1} \end{aligned}$$

**Recalibrated forecasts.** Copy and paste the recalibrated coefficient estimates **B17:B19** into the original *Indian imports* sheet Coefficient column to update *predicted Indian imports*.

Copy the recalibrated *standard error* from **B7** and paste into **G2**.

Change the error degrees of freedom in the *t* formula to 18 to update *95% lower Indian imports* and *95% upper Indian imports*.

1	Year	India Imports (BS) q	Congress Party	Past year GDP per capita (\$K) q-1	coefficient	predicted Indian imports	standard error	t	lower 95% prediction	upper 95% prediction
15	1998	3.56	0	0.43214		3.70145			2.90316	4.4997315
16	1999	3.69	0	0.42671		3.61033			2.812047	4.4086185
17	2000	3.67	0	0.4465		3.94259			3.144303	4.7408743
18	2001	3.76	0	0.46003		4.16986			3.371573	4.9681445
19	2002	4.10	0	0.46455		4.24577			3.447486	5.044058
20	2003	4.98	0	0.47955		4.49768			3.699395	5.2959665
21	2004	6.11	1	0.54802		6.31909			5.52	7.12
22	2005	7.96	1	0.62241		7.5685			6.77	8.37
23	2006		1	0.68665		8.6473			7.85	9.45
24	2007		1	0.72228		9.24569			8.45	10.04

In 2007, Indian imports are expected to reach \$8.5 to \$10.0 billion.

To plot and compare imports with the model forecasts under both leadership scenarios, insert three new columns **B**, **C** and **D** for *predicted Indian imports*, *predicted Indian imports under BJP* and *predicted Indian imports under Congress*.

Copy *predicted Indian imports* in **I2:I24** and use shortcuts to paste with **values and formats** (but not formulas) into **B2:B24**:

Select **I2:I24**, **Cntl+C**, select **B2**, **Alt HVSU**, **Ok**.

1	Year	predicted Indian imports	predicted Indian imports under BJP	predicted Indian imports under Congress	India Imports (BS) q	Congress Party	Past year GDP per capita (\$K) q-1	coefficient	predicted Indian imports	standard error	t	lower 95% prediction	upper 95% prediction	Congress Party
2	1985	1.25021			1.64	0	0.28619	-3.563	1.25021	0.37997	2.10092	0.451927	2.0484984	0
3	1986	1.32315			1.54	0	0.29053	0.67138	1.32315			0.524867	2.1214392	0
4	1987	1.7028			1.46	0	0.31314	16.795	1.7028			0.904519	2.5010906	0

Make *predicted Indian imports under BJP* by changing ones to zeros in column **F**, which would reflect ongoing leadership by the BJP Party.  
 (This will automatically change *predicted Indian imports* in **I** in years that were actually under *Congress* leadership.)

Use shortcuts to paste into *predicted Indian imports under BJP* in column **C**:

Select **I2:I24**, **Cntl+C**, select **C2**, **Alt HVSU**, **Ok**.

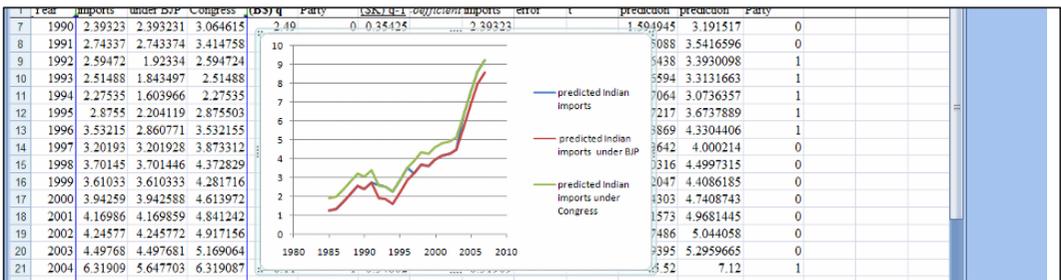
Make *predicted Indian imports under Congress* by changing zeros to ones in column **F**, which would reflect ongoing *Congress* leadership.

Use shortcuts to paste into *predicted Indian imports under Congress* in column **D**:  
 Select **I2:I24**, **Cntl+C**, select **D2**, **Alt HVSU**, **Ok**:

Year	predicted Indian imports	predicted Indian imports under BJP	predicted Indian imports under Congress	India Imports (BS) q	Party	Past year GDP per capita (SK) q-1	coefficient	predicted Indian imports	standard error	t	lower 95% prediction	upper 95% prediction	Congress Party
1990	2.39323	2.393231	3.064615	2.49	0	0.35425	2.39323	2.39323	1.594945	3.191517	0		
1991	2.74337	2.743374	3.414758	2.00	0	0.37509	2.74337	2.74337	1.945088	3.5416596	0		
1992	2.59472	1.92334	2.594724	1.92	1	0.32627	2.59472	2.59472	1.796438	3.3930098	1		
1993	2.51488	1.843497	2.51488	2.78	1	0.32151	2.51488	2.51488	1.716594	3.3131663	1		

Make a scatterplot to compare predictions under the two Parties:

Select *year*, *predicted Indian imports*, *predicted Indian imports under Congress*, and *predicted Indian imports under BJP* in **A1:D24**, **Alt ND**:



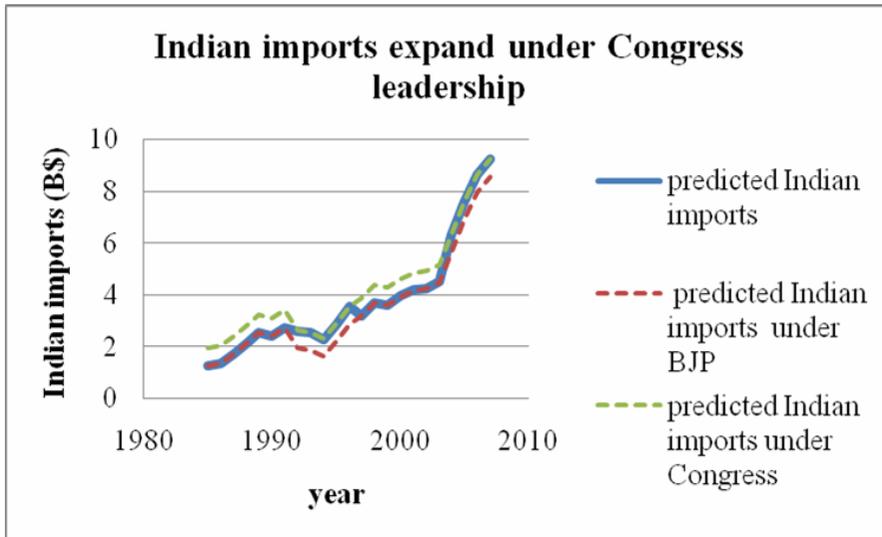
You will see only two prediction lines, since the model's predictions are under the *BJP* forecast in some years and under the *Congress* forecast in other years.

To reveal the model predictions, select the *predicted Indian imports under BJP* line in the legend, right click, **Format Data Series**, then change **Line Style** to dashed.

Select the *predicted Indian imports under Congress* line in the legend, right click, **Format Data Series**, and then change **Line Style** to dashed.

Select the *predicted Indian imports* line in the legend, right click, **Format Data Series**, and then change **Line Style**, to a wider line.

Now we can see how the indicator *Congress* shifts the regression line upward in years when the Congress Party assumes leadership, then back down when leadership reverts to the BJP Party:



## Lab Practice 10

### Conjoint Analysis of PDA Preferences

Rate the nine hypothetical PDAs in **Table 10.3**, then replace the third customer's ratings in **G20:G28** of **Lab Practice 10 PDA conjoint.xls** with *your* ratings.

Follow the steps in **Excel 10.1** to find the

- *Part worth utilities*
- *Standardized attribute importances*

in your regression sheet For PDAs.

Describe the preferred PDA:

---

Which PDA attribute is most important? \_\_\_\_\_

Which PDA attributes do not significantly affect preferences, if any? \_\_\_\_\_

Attach a printout of your regression sheet with the table of part worth utilities and standardized attribute importances.

### *The Climate for a Joint Venture in China*

A coalition of U.S. business leaders is interested in investing in a joint venture in China to produce and sell commercial vehicles there. They require a forecast of commercial vehicle sales in China over the next five years. They are particularly interested in learning

- the degree to which structural shifts from political shocks affect commercial vehicle sales growth, and
- the influence of growth in China's GDP on growth in commercial vehicle sales. Several structural shifts have altered the Chinese political and economic climate in the past twenty-five years.

- **Third Generation Leadership.** In 1989, following rampant inflation and alleged government corruption, students and intellectuals staged protests in Beijing's Tiananmen Square, which spread to major cities throughout China. The Chinese government instituted martial law and silenced protestors. Following Tiananmen Square, Deng Xiaoping stepped down from leadership, though the new *Third Generation* leadership, followed his policies and endorsed his proposals for reform toward a more market-driven economy. Steps were initiated to open China's economy to international trade. **Third Generation Leadership policies were in effect from 1991 through 1996.** Following Deng Xiaoping's death in 1997, a new group of Third Generation leaders assumed power. The Fourth Generation of leaders, assumed power in 2003, led by President Hu Jintao, who is openly committed to trade with the U.S.

The dataset **Lab Practice 10 China JV.xls** contains time series of *China GDP* and *Commercial Vehicle Sales in China* for the period 1990 through 2005, including forecast *China GDP* through 2010.

Follow the steps in **Excel 10.2** to build a model of commercial vehicle sales in China, including

- *an indicator of Third Generation leadership following Tiananmen Square until Deng's death (1991 through 1996)*
- *past year Chinese GDP*

Since this is a time series model, assess the model Durbin Watson statistic to discover whether or not unaccounted for trend or cycles remain.

To confirm that your model produces reliable forecasts, assess your model validity by holding out the two most recent observations, forecasting those, then looking to see whether or not the 95% mean prediction intervals contain the holdout data.

Following validation, recalibrate your model.

Write your model equations for *commercial vehicle sales* in China:

- during **baseline years 1989-1990 and 1997-present**
- during Third Generation Leadership 1991-1996

What is your 95% prediction interval for commercial vehicle sales in 2010? \_\_\_\_\_

To compare the impact of Third Generation Leadership on commercial vehicle sales, plot *predicted commercial vehicle sales* by year

- with ***Third Generation Leadership in 1991-1996*** and
- assuming that ***Third Generation Leadership had continued through 2010***

If Third Generation Leadership had not changed following Deng's death, and had remained in power, what would be the estimated impact on commercial vehicle sales in 2010?

---

Attach a printout of your plot.

## Assignment 10-1 Conjoint Analysis of PDA Preferences

Dell is considering introduction of a new PDA which would be sold at a competitive price through WalMarts. New product development managers believe that customers would choose brightly colored Dell PDAs at competitive prices.

Choose four attributes of PDAs that you believe to be influences on college students' preferences. Identify three alternative options for each attribute and fill in the orthogonal array table, below, to make nine hypothetical PDAs:

<i>Hypothetical PDA</i>	<i>Brand</i>	<i>color</i>	<i>keypad</i>	<i>price</i>
1	Dell	silver	standard	\$150
2	Dell	white	QWERTY	\$250
3	Dell	lime green	touch screen	\$350
4	Apple	silver	QWERTY	\$350
5	Apple	white	touch screen	\$150
6	Apple	lime green	standard	\$250
7	Palm	silver	touch screen	\$250
8	Palm	white	standard	\$350
9	Palm	lime green	QWERTY	\$150

Rate the nine hypothetical PDAs, using a scale from 1 (“undesirable”) to 9 (“very desirable”). Ask two friends or classmates to rank the nine hypotheticals also.

Enter your ratings in the **Assignment 10-1 Dell PDA conjoint.xls**. The file contains 27 rows, nine rows for each person in your sample, and seven columns, *customer*, *hypothetical PDA*, *brand*, *color keypad*, *price*, and *rating*.

Identify the baseline PDA, then make eight indicator variables to designate options other than baseline.

Run a regression to find the preferred PDA configuration, the part worth utilities, and the relative importances of attributes.

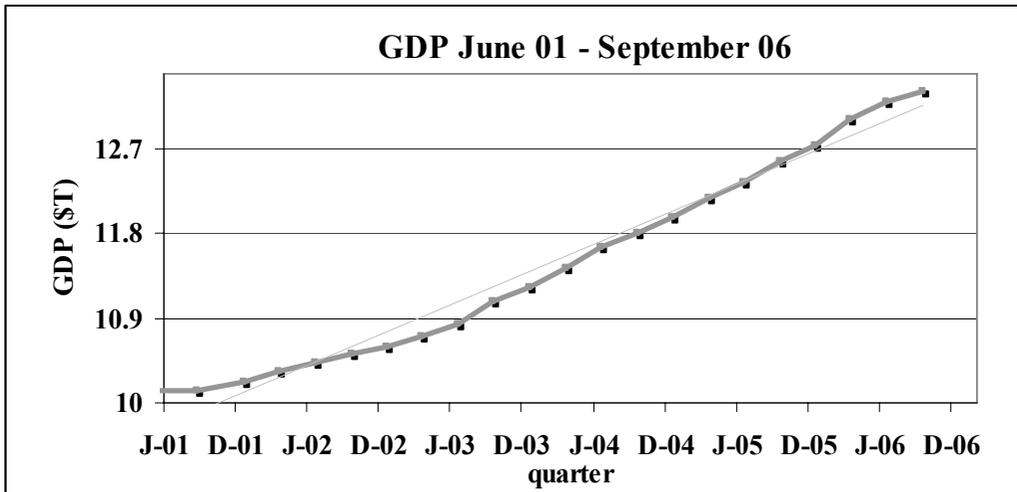
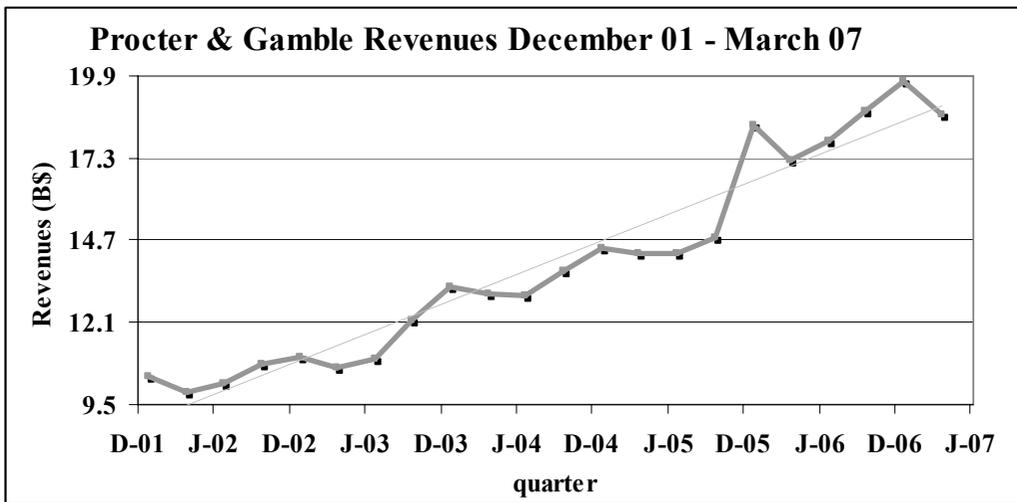
**Deliverables:** Write a paragraph to management, summarizing your results, with recommendations for the new product development team.

Attach a copy of your regression sheet with a table and plots of part worth utilities, and a table of attribute importances.

### CASE 10-1 Modeling Growth: Procter & Gamble Quarterly Revenues

Procter & Gamble revenues are growing as the company’s managers innovate and forge into new markets, and as the company acquires complementary businesses. Procter & Gamble management want to quantify the impact on revenues of the acquisition of Gillette late in 2005. They have asked for a model which quantifies quarterly revenue drivers, including the Gillette acquisition, which can also be used to forecast.

Like the revenues of many firms, Procter & Gamble revenues fluctuations follow movement in GDP. The impact seems to occur fairly quickly, after about two quarters:



The terrorist incident of September 11, 2001 affected business performance in many industries, and P&G executives believe that revenues were unusually low in the seven quarters following the incident.

Procter & Gamble acquired Gillette in 2005. The first quarter of the combination is December 2005. Revenues in that quarter were nearly \$4 billion greater than in the preceding quarter.

Build a time series model of P&G revenues, including the *9/11 shock*, the *Gillette acquisition*, and the Leading Indicator, *past year GDP*.

Quantify the impact of the 9/11 shock and estimate how damaging a similar incident could be in the future by adding an indicator:

- *9/11*, equal to one the last quarter of 2001 through the second quarter in 2003 and equal to zero in other quarters.

Add an indicator of the *Gillette boost*, equal to zero in quarters before December 2005 and equal to one in December 2005 and quarters after.

Assess the Durbin Watson statistic to decide whether or not your model has accounted for trend, cycles and seasonality in the quarterly data.

Validate your model, then add the two most recent quarters and recalibrate.

### **Sensitivity analysis to find expected response under alternate scenarios.**

Find forecasts with the *9/11* indicator set to zero to determine what *revenues* would have been had there not been a terrorist incident.

### **Deliverables.**

1. Write your model equations for

- The baseline before *9/11*
- Following *9/11*
- After the *Gillette acquisition*

2. What is the margin of error in your forecasts? \_\_\_\_\_

3. What are the *95% prediction intervals* for revenues in June and September 2007?

June 2007: \_\_\_\_\_ September 2007: \_\_\_\_\_

4. What is the expected percent increase in *revenues* in these two quarters, relative to the same quarters in 2006?

June 2007 relative to June 2006: \_\_\_\_\_

September 2007 relative to September 2006: \_\_\_\_\_

5. Make a table to show

- *revenue* lost in each of the seven quarters following 911
- The percent reduction from expected *revenues* had there been no incident

6. Make a table to show

- how much the Gillette acquisition has enhanced *revenues* in each of the quarters since December 2005.
- The percent of *revenues* contributed by Gillette relative to what *revenues* would have been without Gillette in each of the quarters since December 2005

7. Illustrate your model fit and sensitivity analysis with a scatterplot of

- *revenue predictions*, December 2001 through September 2007
- *actual revenues*
- *revenues predictions* without the *Gillette acquisition* from December 2006 through September 2007

*CASE 10-2 Store24 (A): Managing Employee Retention\**  
*and Store24 (B): Service Quality and Employee Skills\*\**

Use the accompanying data in **Case 10-2 store24.xls** for your analyses and preparation for class discussion.

---

\*Harvard Business School case 9602096

\*\*Harvard Business School case 9602097

# 11

## Nonlinear Multiple Regression Models

In this chapter, we consider the use of nonlinear transformations that allow us to use multiple linear regression to model situations in which marginal responses are either increasing or decreasing, rather than constant. We will explore Tukey's Ladder of Powers to identify particular ways to rescale variables to produce valid models with superior fit.

### 11.1 Consider a Nonlinear Model When Response Is Not Constant

To decide whether or not to use a nonlinear model, first rely on your logic:

- Do you expect the response, or change in the dependent, performance variable, to be constant, regardless of whether a change in an independent variable is at minimum values or at maximum values? Linear models assume constant response.
- Is the dependent variable limited or unlimited?

Linear models are unlimited. If your dependent variable couldn't be negative, because it is measured in dollars, purchases, people, or uses, a nonlinear model is logically more appropriate.

After consulting your logic, plot your data, then fit a line as well and examine the residuals. You will see just how well a linear model fits.

### 11.2 Tukey's Ladder of Powers

Tukey offered a simple heuristic to quickly suggest ways to rescale variables when residuals from linear regression would be either skewed or heteroskedastic. Scales are chosen which reduce skewness of both independent and dependent variables.

If a variable is positively skewed, as the variable on the left in Figure 11.1, shrinking it by rescaling in square roots, natural logarithms, or inverses (reciprocals) will *Normalize*. Square roots are lower absolute power, .5, than inverses, -1, and are less radical. Natural logarithms are moderate, making a bigger difference than square roots and a smaller difference than inverses.

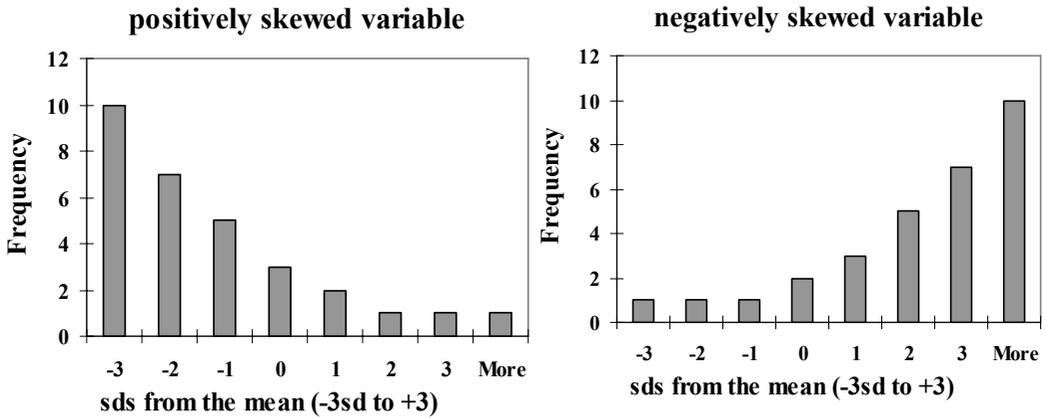


Figure 11.1 Positively and negatively skewed variables

When a variable is negatively skewed, as is the variable on the right in Figure 11.1, expanding it by rescaling to squares or cubes will *Normalize*. The higher power, cubes, will make a bigger difference.

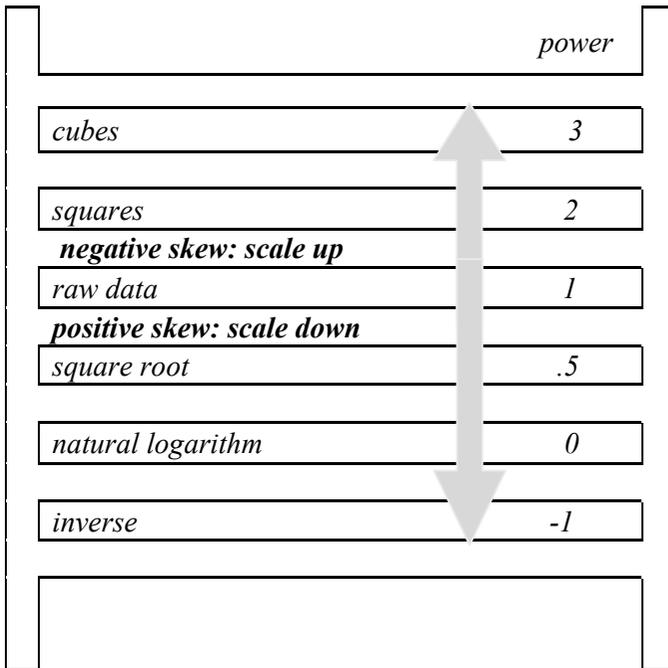


Figure 11.2 Tukey's Ladder of Powers

Moving from the center up or down the *Ladder of Powers*, Figure 11.2, by using a higher power, changes the data more. More skewness calls for rescaling with a higher power.

### 11.3 Rescaling $y$ Builds in Synergies

Sometimes one driver is particularly strong when a second driver is included in the model. Jointly, two drivers may make a larger difference than the sum of their individual influences. For example, advertising levels may be more effective when sales forces are larger. The impact of population growth in a country may influence imports more if growth in GDP has been relatively high. When we rescale the dependent variable, we build in synergies between predictors. To this potential benefit of improved fit and validity, comes the cost of transforming predictions in rescaled units back to the original units.

*Example 11.1 Executive Compensation.* The Board of a large corporation in the Financial industry is courting a new CEO candidate. To more precisely craft their offer, they would like to be able to relate executive compensation to performance in the industry. They have asked for a model relating executive compensation to firm sales, profits, and returns in similar large corporations. Forbes has published a dataset containing executive compensation, firm performance, and demographics from a sample of five hundred large corporations. Using this dataset, we will build a model to help The Board more confidently quantify their offer.

Board members believe that executives from larger, more profitable firms earn more, and that older, more experienced executives are better compensated. They also believe that there may be noticeable differences across industries. We will include in the model

- *Revenues in billion (B) dollars,*
- *Profits in million (MM) dollars,*
- *Percent return over five years,*
- *Age in years,*
- *Indicators to distinguish industries*

Complete data on these measures are available for 434 firms in six major industries: Computers, Energy, Financial, Food, Health and Utilities. The best paid executives are compensated well beyond most. Consequently, approximately ten percent of the total compensation packages are outliers within each of the six industries and will be excluded, leaving a sample of 402 CEOs of large corporations.

Four of the five continuous variables are positively skewed, as Figure 11.3 and Table 11.1 illustrate. A relatively small proportion of executives are better compensated, and a relatively small proportion of firms have *higher revenues, profits, and five year returns*. *Age* is approximately *Normally* distributed.

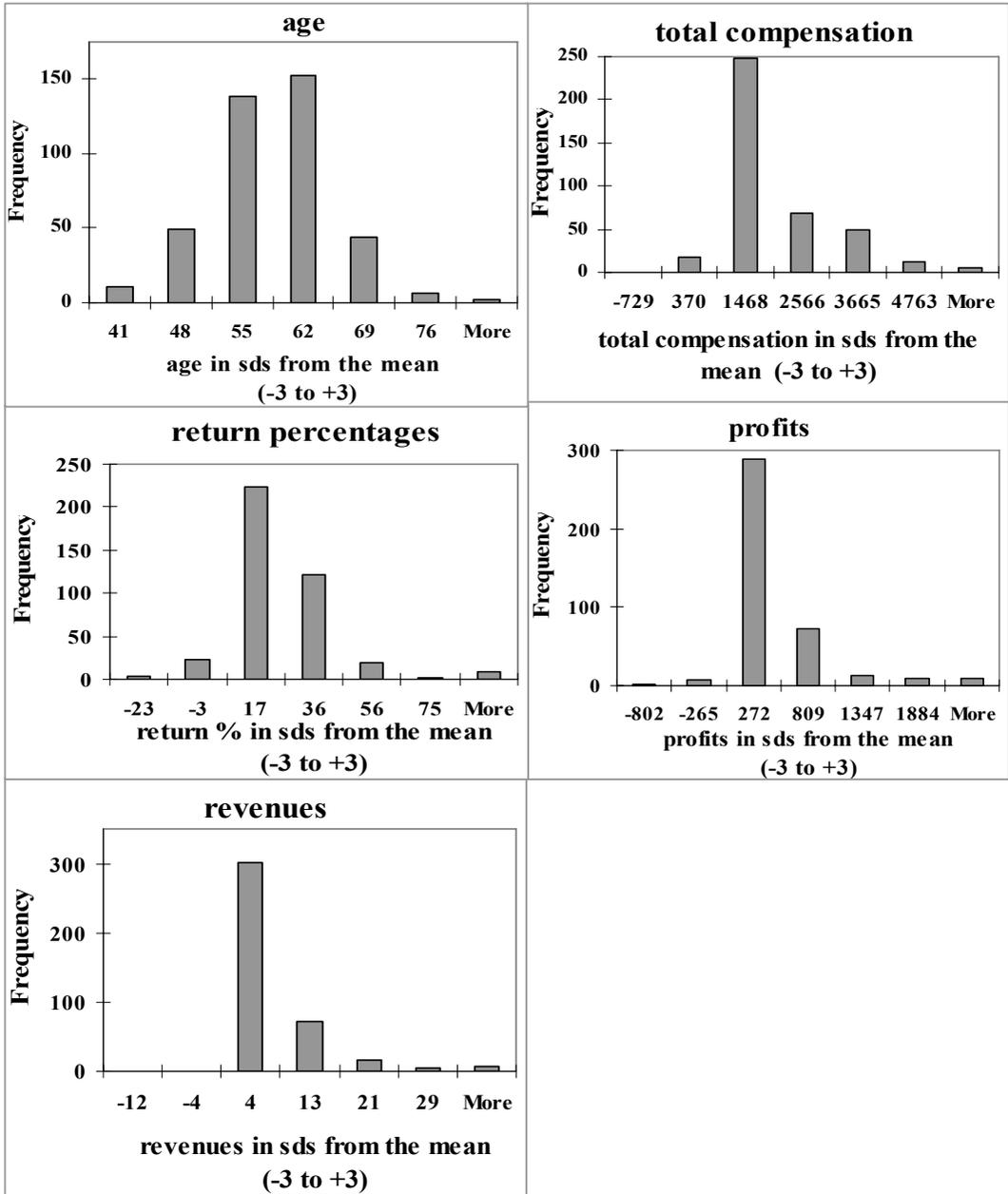


Figure 11.3 Skewness of variables in the executive compensation data

	<i>age</i>	<i>total compensation</i> (\$K)	<i>5-year</i> <i>return %</i>	<i>profits</i> (\$MM)	<i>revenues</i> (\$B)
<i>Skewness</i>	-2	1.5	3.0	4.4	6.1

Table 11.1 Skewness of executive compensation variables

To *Normalize* positively skewed total compensation, the square roots or natural logarithms, shown in Figure 11.4 are effective.

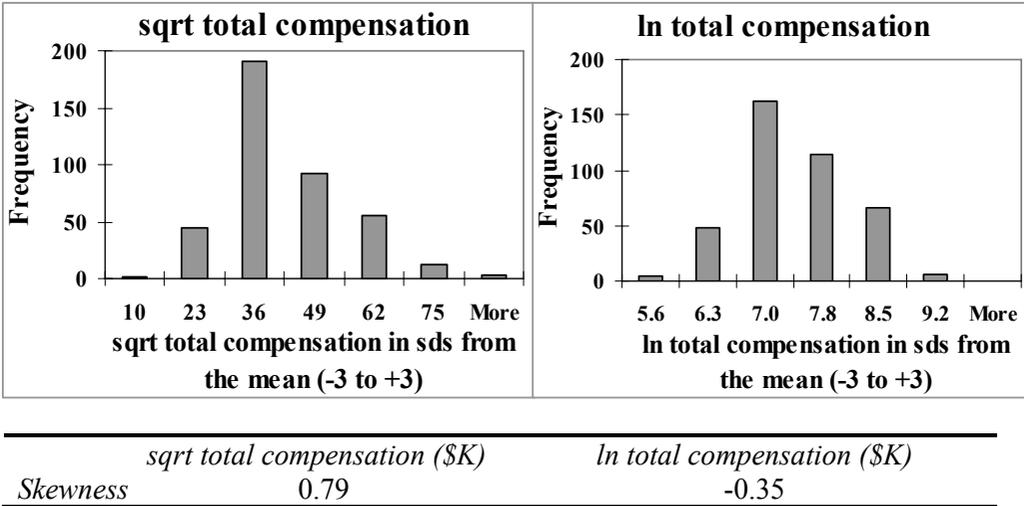


Figure 11.4 Rescaled total compensation and revenues

Revenues are more positively skewed, and the square roots, shown in left panel of Figure 11.5, aren't enough correction. The natural logarithms, shown in the right panel of Figure 11.5, are needed to remove the positive skew:

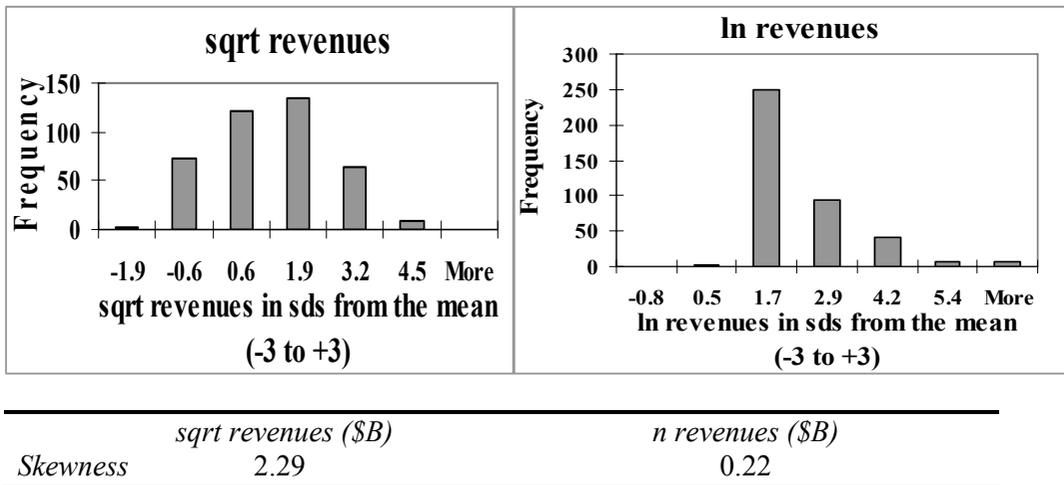
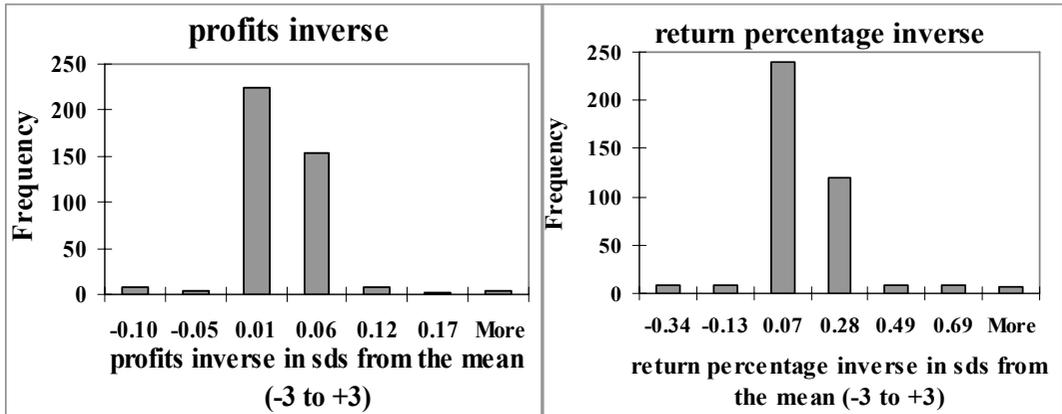


Figure 11.5 Rescaled revenues

With profits and five year return, square roots and natural logarithms are not options, since some firms reported negative profits and negative returns. The available option for positively skewed variables with negative values is to invert, scaling in inverses, which are shown in Figure 11.6.



	<i>profits(\$MM) inverse</i>	<i>5-year return % inverse</i>
<i>Skewness</i>	-0.04	-0.21

Figure 11.6 Rescaled profits and returns

Inverses are fairly drastic and produce peaked distributions where most cases are close to the mean. We will retain the original scales of profits and five year return percentage. The nonlinear multiple regression model results are in Table 11.2.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.602					
R Square	0.362					
Adjusted R Square	0.347					
Standard Error	10.575					
Observations	402					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	9	24878	2764	24.7	.0000	
Residual	392	43837	112			
Total	401	68714				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.78	4.69	1.0	.31	-4.45	14.01
<i>computers</i>	13.3	2.0	6.6	.0000	9.4	17.3
<i>energy</i>	7.7	2.2	3.4	.0007	3.3	12.1
<i>financial</i>	11.3	1.7	6.6	.0000	7.9	14.6
<i>food</i>	4.7	2.1	2.3	.02	0.7	8.8
<i>health</i>	12.7	2.1	5.9	.0000	8.4	16.9
<i>ln revenues (B\$)</i>	4.03	0.65	6.2	.0000	2.76	5.31
<i>profits(MM\$)</i>	.0040	.0013	3.1	.002	.0015	.00650
<i>5-year return %</i>	.107	.030	3.6	.0004	.048	.165
<i>age</i>	.307	.080	3.8	.0001	.150	.464

Table 11.2 Executive compensation is driven by industry, firm performance and executive age

From regression output, the nonlinear model equation is:

$$\begin{aligned}
 \text{TotalCompensation}(\$K)^5 = & 4.78 + 13.3^a \text{ computers} + 7.7^a \text{ energy} + 11.3^a \text{ financial} \\
 & (4.69) \quad (2.0) \quad (2.2) \quad (1.7) \\
 & + 4.7^b \text{ food} + 12.7^a \text{ health} + 4.03^a \ln(\text{revenues}(\$B)) + .307^a \text{ age} \\
 & (2.1) \quad (2.1) \quad (.65) \quad (.080) \\
 & + .0040^a \text{ profits}(\$MM) + .107^a \text{ return}\% \\
 & (.0013) \quad (.030)
 \end{aligned}$$

*RSquare*: 36%<sup>a</sup>

<sup>a</sup>Significant at .01 or better

<sup>b</sup>Significant at .05

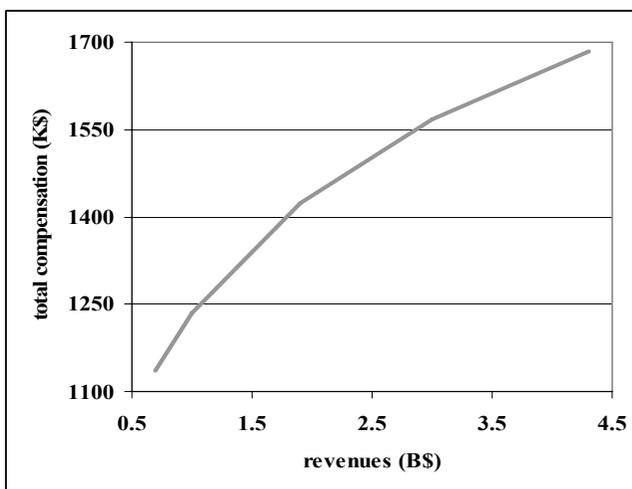
This equation is in square roots. To see the equation in the original scale of hundred thousand dollars, we square both sides:

$$\begin{aligned} TotalCompensation(\$K) = & [4.78 + 13.3^a computers + 7.7^a energy + 11.3^a financial \\ & + 4.7^b food + 12.7^a health + 4.03^a Ln(revenues(\$B)) + .307^a age \\ & + .0040^a profits(\$MM) + .107^a return\%]^2 \end{aligned}$$

Variation in firm performance, industry and age differences account for 36% of the variation in CEO compensation. Better performing firms pay their executives more ( $b_{Ln\text{Revenues}} = 4.03 > 0$ ). Older, more experienced executives earn more ( $b_{Age} = .307 > 0$ ), and compensation is higher in computer, health and financial industries, ( $b_{Computers} = 13.3; b_{Health} = 12.7; b_{Financial} = 11.3; b_{Energy} = 7.7 \equiv 1$ ), lower in food industries  $0 < b_{Food} = 4.72 < 1$ , and lowest in (the baseline) industry, utilities.

#### 11.4 Sensitivity Analysis Reveals the Relative Strength of Drivers

By comparing expected total compensation for hypothetical firms, we can compare the relative impact of each of the drivers. For example, within the Financial industry (with Computers, Energy, Food, Health indicators each equal to zero and the Financial indicator equal to one), the impact of a difference in firm size, the difference between lower and higher revenues, can be estimated by comparing predicted compensation with other drivers set at their mean or median values. For Normally distributed variables, such as age, choose the mean as a representative value. For skewed variables, such as profits and return percentage, choose the medians as representative values.



50% of the sample firms earned revenues between .7 and 4.3 (B\$) billion dollars, the inter-quartile revenue range.

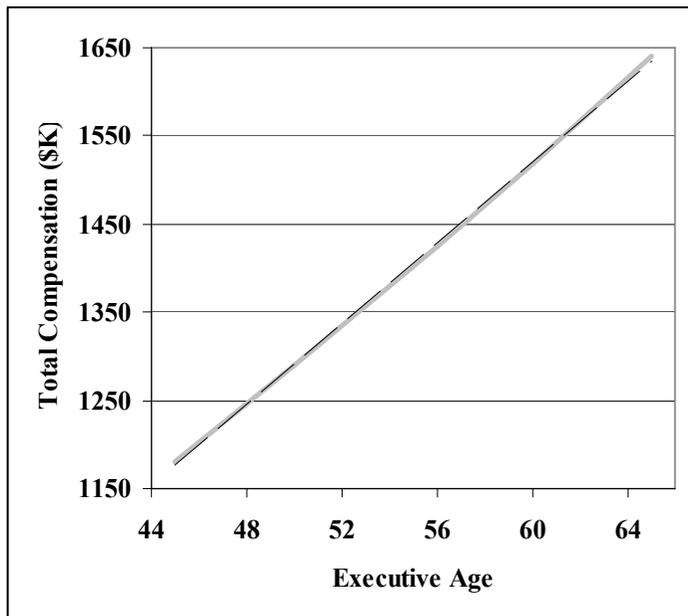
From Figure 11.7, we see that an executive (at the average age of 55, whose firm earned median profits of \$114 MM with a five year median return of 14%) could expect to earn \$1,120 to \$1,660 (K), depending on firm revenues.

Figure 11.7 Revenues drive expected total compensation

If all continuous drivers except *revenues* (*age*, *profits*, and *five year return*) are at median levels, *revenue* differences make an expected difference of about \$600 (K) (= \$600,000) to the executives in the Financial industry.

*Total compensation* response to *revenues* increases at a *decreasing* rate. *Executive compensation* differences are greater for firm *revenue* differences among smaller firms than among larger firms: *revenues* influence *executive compensation* more when *revenues* are lower.

To compare compensation of younger, less experienced executives with older, more experienced executives in the Financial industry, we set the performance variables (*revenues*, *profits*, and *five year return*) at median levels. Then we observe that the difference of ten years in executive *age* is associated with an expected difference in *compensation* of about a quarter million dollars:



More than half of executives in this sample of large corporations are between the ages of 51 and 60.

From Figure 11.8, we see that across this fairly narrow range of executive ages, a difference of nine years, makes an expected difference of \$210 (K) (= \$210,000).

Figure 11.8 Executive age drives compensation

Repeating this process for each of the independent variables, we see from Table 11.3 that industry, firm size (*revenues*), and executive *age* are the three most important drivers of executive compensation, and in that order.

Longer term firm performance indicators, *profits* and *five year return*, are less influential.

<i>Driver</i>	<i>Expected Compensation</i>		<i>Expected</i>
	<i>Range (\$K)</i>		<i>Difference (\$K)</i>
<i>Industry (Utilities to Computers)</i>	742	1,670	929
<i>Revenues (.7 to 4.28 B\$)</i>	1,120	1,660	540
<i>Age (51 to 60)</i>	1,320	1,530	210
<i>Returns (8 to 20%)</i>	1,360	1,460	100
<i>Profits (50 to 300 MM\$)</i>	1,390	1,460	70

Table 11.3 Industry and firm revenue are the most influential drivers of executive compensation

Results of the analyses are summarized in the memo to The Board, below:

## MEMO

**Re: Executive Compensation Driven by Firm Performance and Age**

**To:** The Board

**From:** James Melton, Director, Econometric Analysis

**Date:** June 2007

Analysis of 402 executive compensation packages offered by firms surveyed by Forbes Magazine reveals that industry, firm performance and executive age are key drivers.

**Compensation Model.** Using *Forbes* data from 402 of firms in six broad industries, a model linking industry, executive age and firm performance measures with compensation was built.

**Model Results.** Industry, executive age, firm revenues, profits, and return percent over five years account for 36% of the variation in compensation.

Executives in the financial industry are better rewarded than those in food, energy, or utilities, but paid less than those in computer, communications or health sectors.

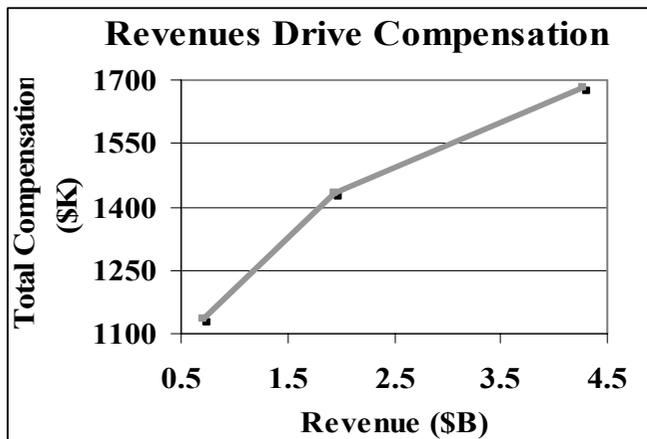
Aside from industry differences, firm revenues are the strongest driver of compensation, particularly for smaller firms with revenues less than the median of \$2 billion.

Older, more experienced executives and those heading more profitable firms with higher returns are better compensated.

Each year in age adds an average of \$200 to compensation packages.

On average, among financial firms, revenue differences make \$.5 million difference in compensation packages, return percentage differences make a \$100 thousand difference, and profit differences make a \$70 thousand difference in compensation packages.

**Conclusions.** In similar financial firms, executive compensation is tied to experience and firm performance. More experienced, more successful executives are better rewarded, particularly for growth in firm revenues.



$$\begin{aligned} \text{Compensation}(\$K) = & [4.8 + 13.3^a \text{ computers} \\ & + 7.7^a \text{ energy} + 11.3^a \text{ financial} \\ & + 4.7^b \text{ food} + 12.7^a \text{ health} \\ & + 4.0^a \ln(\text{revenues}(\$B)) \\ & + .0040^a \text{ profits}(\$MM) \\ & + .30^a \text{ age} + .11^a \text{ return}\%]^2 \end{aligned}$$

*RSquare: .36<sup>a</sup>*

*<sup>a</sup>Significant at .01; <sup>b</sup>Significant at .05.*

### 11.5 Gains from Nonlinear Rescaling Are Significant

What did we gain by building a nonlinear model instead of a simpler linear model? The linear model of total compensation using the same variables and Forbes sample is:

$$\begin{aligned}
 \widehat{Total\ Compensation}(K\$) = & -1030^a + 1330^a \textit{computers} + 678^a \textit{energy} + 654^a \textit{financial} \\
 & (410) \quad (180) \qquad \qquad (190) \qquad \qquad (140) \\
 & + 563^a \textit{food} + 1050^a \textit{health} + .75^a \textit{profits}(MM\$) + 29^a \textit{age} \\
 & (180) \qquad (190) \qquad (.09) \qquad \qquad (7)
 \end{aligned}$$

*RSquare*: 29%<sup>a</sup>

<sup>a</sup>Significant at .01 or better

In the linear model, firm *revenues* and *5-year return percentage* are not significant, and have been removed, accordingly. The remaining predictors, industry indicators, firm *profits*, and executive *age* account for 29% of the variation in executive *compensation*. Relying on a linear model, The Board would ignore the particularly important links between firm *revenues*, firm *return percentage* over five years, and *total compensation* reducing potential performance incentives.

Comparing residuals from the nonlinear and linear models, shown in Figure 11.9, we see that the nonlinear model residuals are less skewed and better satisfy multiple linear regression assumptions:

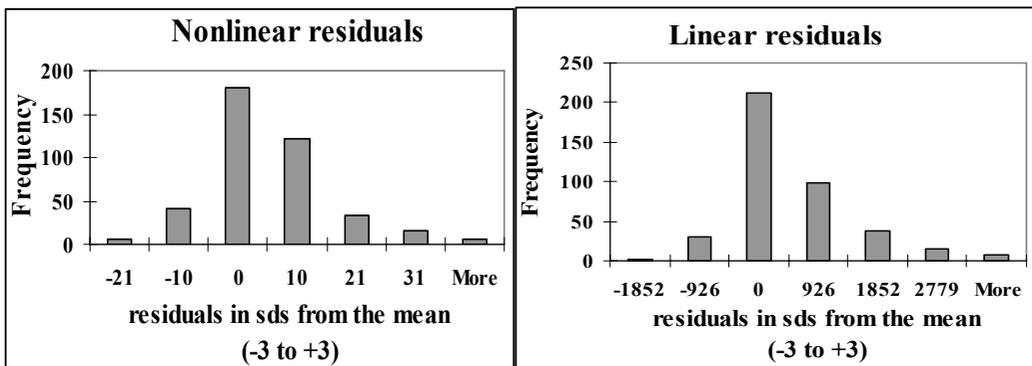


Figure 11.9 Residuals from the nonlinear model (left) are closer to normal

## 11.6 Nonlinear Models Offer the Promise of Better Fit and Better Behavior

It is a challenge to think of an example of truly linear (constant) response. Responses tend to be nonconstant, or nonlinear. We consume and invest in nonlinear ways. The fifth dip of ice cream is less appetizing than the first. Consumers become satiated at some point, and beyond that point, additional consumption is less valuable. Adding the twentieth stock to a portfolio makes less difference to diversification than adding the third. A second ad insertion in a magazine enhances recall more than a tenth ad insertion. As a consequence of nonconstant, changing marginal response, nonlinear models tend to offer the promise of superior fit and better behaved models, with more nearly random residuals. Nonlinear models do carry the cost of transformation to and back from logarithms, square roots, inverses or squares. In some cases, a linear model fits data quite well and is a reasonable approximation. Thinking logically about the response that you've set to explain and predict, and then looking at the distribution and skewness of your data and your residuals, will sometimes lead you toward the choice of a nonlinear alternative.

Tukey's Ladder of Powers can help quickly determine the particular nonlinear model which will fit a dataset best. When a variable is positively skewed, rescaling to square roots, natural logarithms, or inverses often reduces the positive skew. Negatively skewed variables are sometimes Normalized by squaring or cubing. The amount of difference corresponds to the power—square roots with power .5 are less radical than inverses with power (-)1 and squares with power 2 are less extreme than cubes with power 3.

## Excel 11.1 Rescale to build and fit nonlinear regression models with linear regression

*Executive Compensation.* Executive compensation, including salary, stock options, and bonuses, probably depends on the industry, executive age (reflecting experience), and company performance. Company performance measures include revenues, profits, and five-year return percentage.

Since the fewer, exceptional executives are probably compensated more, we expect total executive compensation to be positively skewed. Because unsuccessful firms exit markets, we expect company performance measures to be positively skewed, as well.

To assess skewness and to choose how to rescale, we will look at the skew in the distributions of *total compensation*, *age*, *revenues*, *profits*, and *five-year return percentage*. These data for 402 firms surveyed by Forbes magazine are in **Excel 11.1 Executive Compensation.xls**.

**Assess skewness and choose scales.** Use Excel's **SKEW(array)** function to assess skewness.

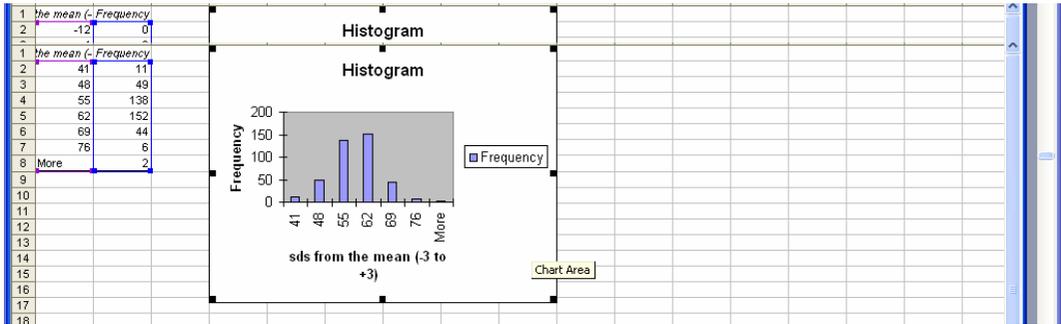
In row **405**, column **A**, type in the label *skewness*, and in column **B** enter = **SKEW(B2:B403)[Enter]**.

Select the new cell **B405**, **Shift+→** through **F405**, **Cntl+R** to fill in skewness values:

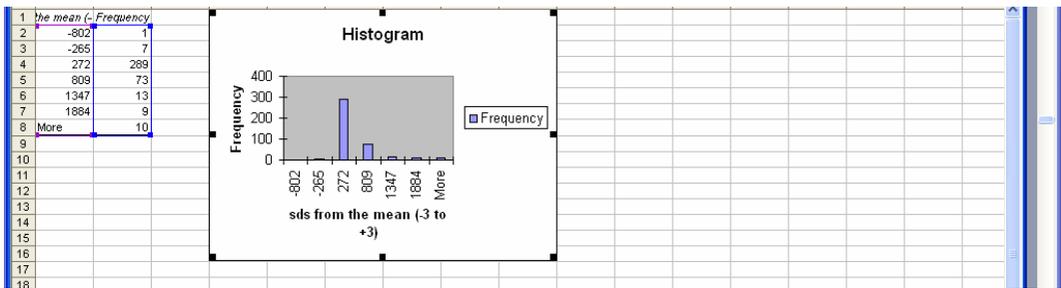
B405		=SKEW(B1:B403)														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O		
Wide Industry	total compensation (K\$)	revenues (\$B)	Age	Profits (MM\$)	Return % Over 5 Yrs	computers & communication	energy	financial	health	food						
393 Utility	476.785	1.581	56	-55.8	10	0	0	0	0	0						
394 Utility	457.096	2.727	51	348.1	13	0	0	0	0	0						
395 Utility	444.847	0.845	61	85.5	14	0	0	0	0	0						
396 Utility	434.984	0.971	51	111.1	11	0	0	0	0	0						
397 Utility	421.197	0.826	55	104.5	10	0	0	0	0	0						
398 Utility	415.552	1.8	61	166	11	0	0	0	0	0						
399 Utility	397.721	0.866	55	95.3	12	0	0	0	0	0						
400 Utility	391.101	2.474	58	-943	1	0	0	0	0	0						
401 Utility	324.434	0.54	55	84.5	11	0	0	0	0	0						
402 Utility	314.382	2.066	55	297.2	15	0	0	0	0	0						
403 Utility	280.396	1.076	60	107.2	14	0	0	0	0	0						
404																
405 skewness	1.542100737	6.13373	-0	4.435	3.04											

*Total compensation* and the three firm performance measures are positively skewed. Executive *age* is slightly negatively skewed.

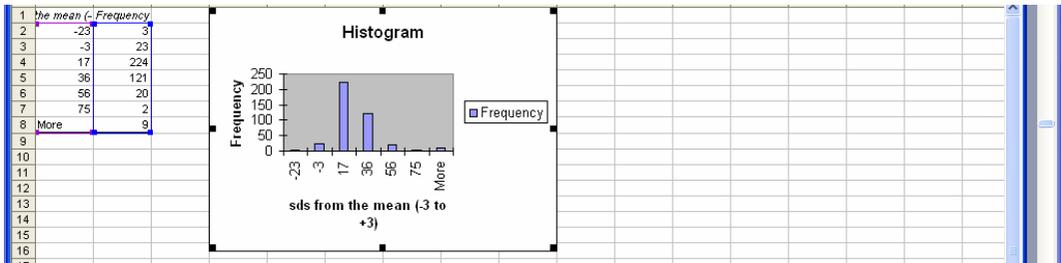
To see the skewness, make histograms for *revenues*,



*age*,



*profits*,



and *return percentage*:

To *Normalize* the positively skewed variables we shrink. For *total compensation* and *revenues*, which are never zero and never negative, we will consider the square roots and the natural logarithms, which have powers .5 and 0 on Tukey's Ladder.

Use shortcuts to add four columns: select **O** through **R**, **Alt HIC**.

Make  $\sqrt{\text{total compensation}} (\$K)$ ,  $\ln \text{total compensation} (\$K)$ ,  $\sqrt{\text{revenues}} (B\$)$ , and  $\ln \text{revenues} (B\$)$  in **O** through **R**.

- In **O2**, enter  $=\sqrt{\text{B2}}$  [Enter].
- In **P2**, enter  $=\ln(\text{B2})$  [Enter].
- In **Q2**, enter  $=\sqrt{\text{C2}}$  [Enter].
- In **R2**, enter  $=\ln(\text{C2})$  [Enter].

For *profits* and *five-year return*, which are sometimes negative, we cannot use either square roots or logarithms. We will consider the inverse of both of these, which has power -1.

Add two columns by selecting **S** and **T**, **Alt HIC**.  
Enter *profit (MM\$) inverse* and *five year return % inverse* in **S** and **T**.

- In **S2** enter  $=1/\text{E2}$  [Enter].
- In **T2** enter  $=1/\text{F2}$  [Enter].

Select the six new cells **O2:T2** and double click the lower right corner to fill in the rows:

T402		=1/F402																		
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Industry	total compensation (K\$)	Age	revenues (\$B)	Profits (MM\$)	Return % 5 Yrs	computers & communication	energy	financial	health	food	mean	standard deviation	sds from the mean (-3 to +3)	sqrt total compensation (K\$)	ln total compensation (K\$)	sqrt revenues (\$B)	ln revenues (\$B)	profits (MM\$) inverse	return % 5 yrs inverse
402	Utility	314.38	55	2.07	297.2	15	0	0	0	0	0			17.731	5.75061	1.43736	0.72561	0.0034	0.0667	
403	Utility	280.4	60	1.08	107.2	14	0	0	0	0				16.745	5.6362	1.0373	0.07325	0.0093	0.0714	
404																				

Use shortcuts to fill in skewness, means, and standard deviations for the rescaled variables:

Select **F405:F407**, **Shift+>** through **T405:407**, **Cntl+R**.

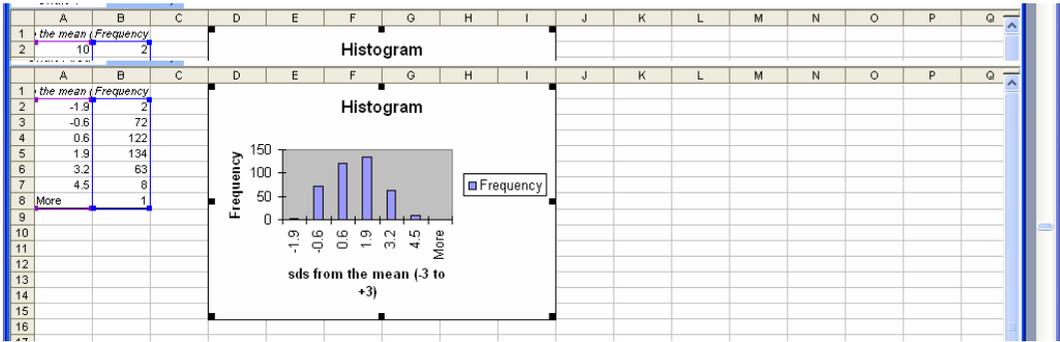
T405		=SKEW(T1.T403)																		
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Industry	total compensation (K\$)	Age	revenues (\$B)	Profits (MM\$)	Return % 5 Yrs	computers & communication	energy	financial	health	food	mean	standard deviation	sds from the mean (-3 to +3)	sqrt total compensation (K\$)	ln total compensation (K\$)	sqrt revenues (\$B)	ln revenues (\$B)	profits (MM\$) inverse	return % 5 yrs inverse
402	Utility	314.38	55	2.07	297.2	15	0	0	0	0	0			17.731	5.75061	1.43736	0.72561	0.0034	0.0667	
403	Utility	280.4	60	1.08	107.2	14	0	0	0	0	0			16.745	5.6362	1.0373	0.07325	0.0093	0.0714	
404																				
405	skewness	1.5421	-0	6.13	4.435		2	2.78	0.5	2.64	2.22			0	0.7885	-0.3485	2.28923	0.21731	-0.037	-0.206
406	mean	1.469	55	4.38	272.2	16.57	0.1	0.00	0.28	0.10	0.13	0.1	0.2069	0.19	26.015	2.0270	1.60905	0.63265	0.0096	0.0743

The skewness of the square roots and natural logarithms of *total compensation* are closer to Normal.

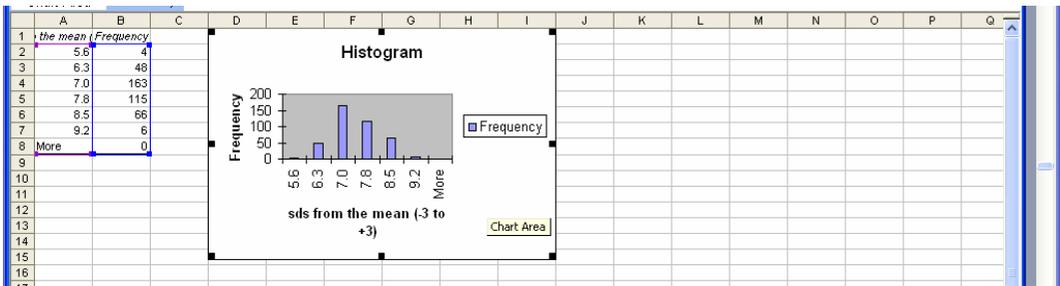
Skewness in the square roots of *revenues* is positive but greater than one, and skewness in the natural logarithms is close to zero.

Skewness of inverses of *profits* and *returns* are close to zero.

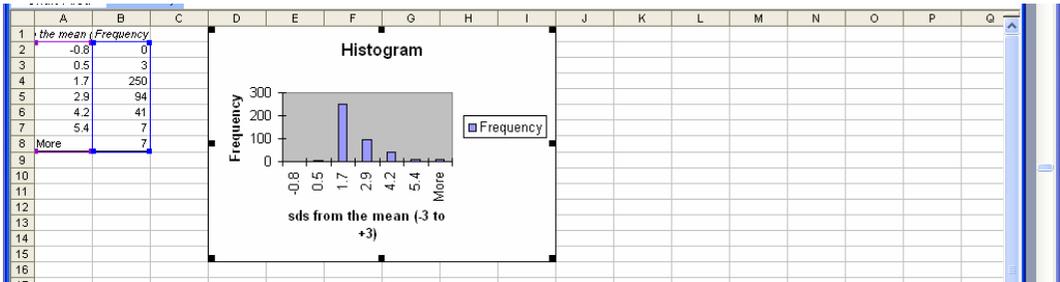
Reset the mean and standard deviation to make histograms of *sqrt total compensation*,



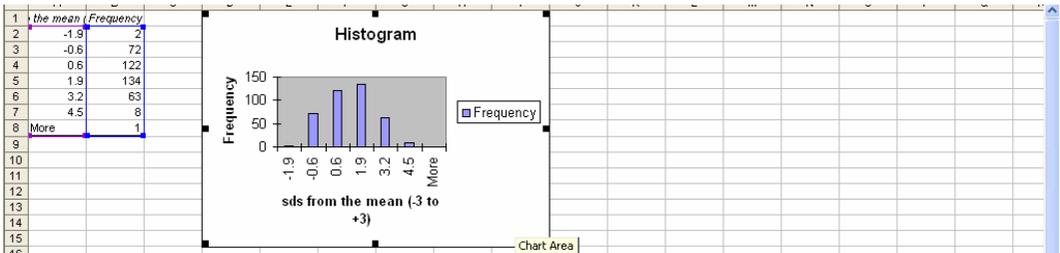
In total compensation,



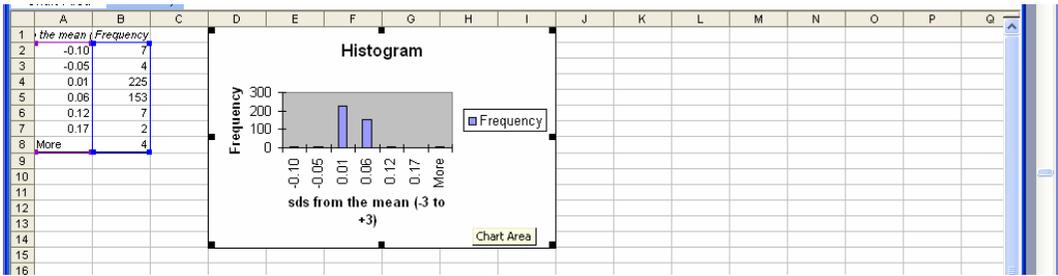
sqrt revenues,



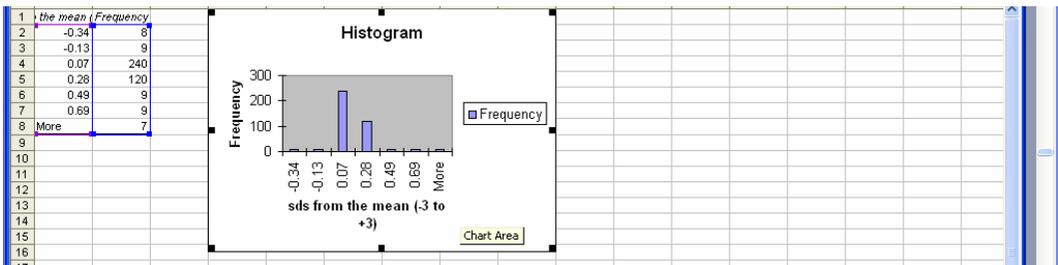
In revenues,



profit inverse,



and return inverse:



We will use the square roots of *total compensation* with the natural logarithms of *revenues*, leaving *profits* and *returns* in their original scales.

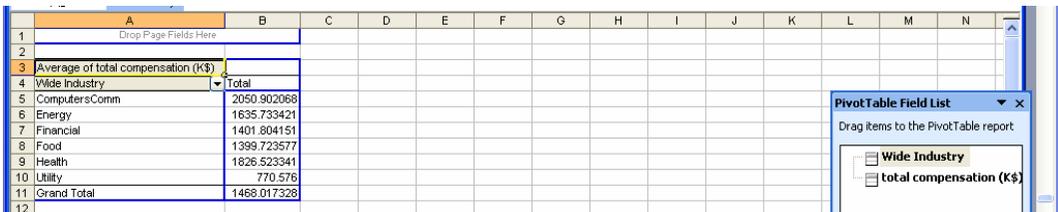
**Add indicators.** To account for industry differences in executive compensation, add industry indicators. There are six industries represented in the dataset. It will simplify interpretation if we choose the industry with lowest average executive compensation for our baseline. Coefficient estimates for the five other industry indicators will reflect the average difference from the least well compensated baseline.

Find average total compensation by industry with a PivotTable.

Select **A1:B403**, **Alt NVT**.

Drag *Wide Industry* to the **ROWS** and *total compensation (\$K)* to **DATA**.

Double click **Count of total compensation** and choose **Summarize by Average**, **Ok**.



Executives in the *utility* industry are least well compensated, on average.

Designate *utility* as the baseline industry, using indicators for each of the remaining five.

Select all of the rows and columns, then use shortcuts to sort the dataset by industry:

Select **A1**, **Cntl+Shift->**, **Cntl+Shift** down through row **403**.

**Alt AS**, **Sort By Wide Industry**, **Header Row**:

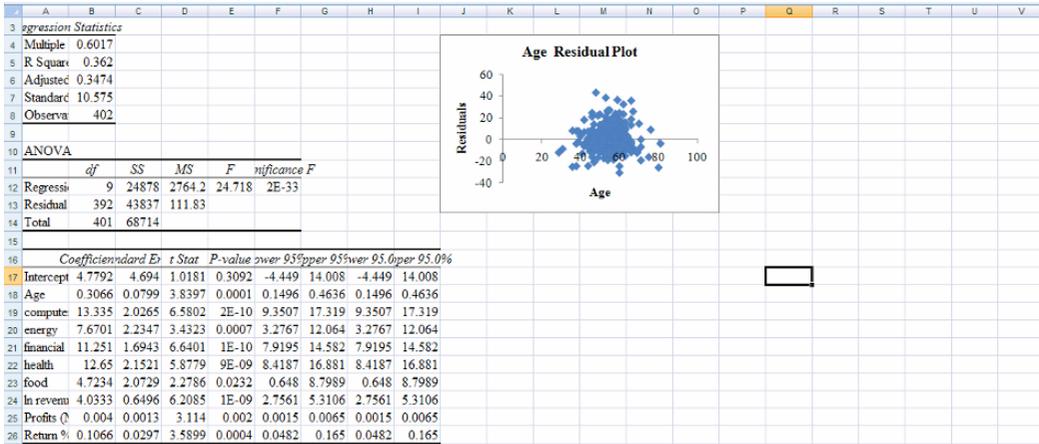
The five indicators *computers & communication*, *energy*, *financial*, *food* and *health* are in columns **G** through **K**. Confirm that in the rows **1:60**, *computers & communication* cells in **G** are one and other indicator column cells **H:K** are zero.

Confirm that the last group, *utilities*, has zeros in each of the five columns **H:L**, since *utilities* are our baseline.

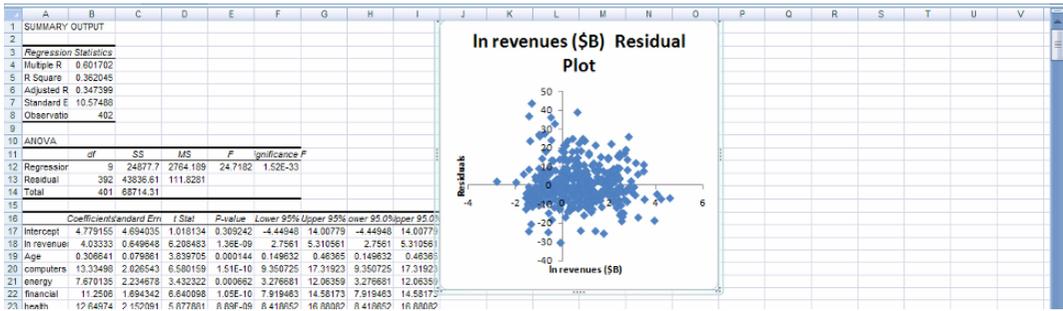
To see the residual plot of the residuals, a continuous variable must come first in the predictor list. Excel plots the residuals by the first predictor in the list. Rearrange columns so that *age* appears first in **E**, followed the five indicator columns:

The dependent variable, *sqrt total compensation* in **D** is followed by age in **E**, the five indicators in **F** through **J**, and the three firm performance variables, *ln revenues*, *profits* and *returns* in **K** through **M**.

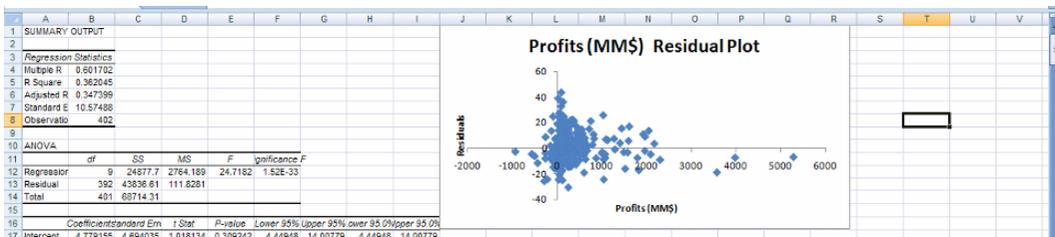
Run regression using the rescaled variables.



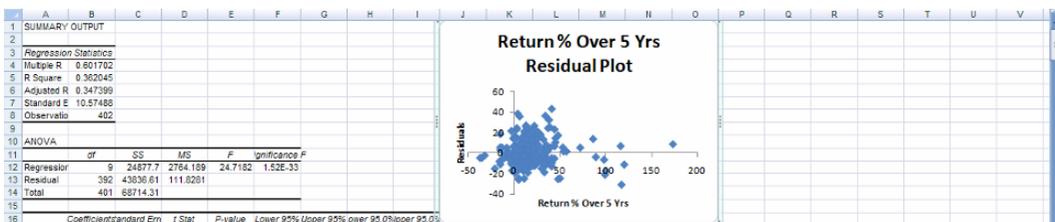
To see the plot of residuals by *ln revenues*, rearrange columns, placing *ln revenues* in **E**, and re-run the regression:



To see the plot of residuals by *profits*, rearrange columns, placing *profits* in **E**, and re-run:

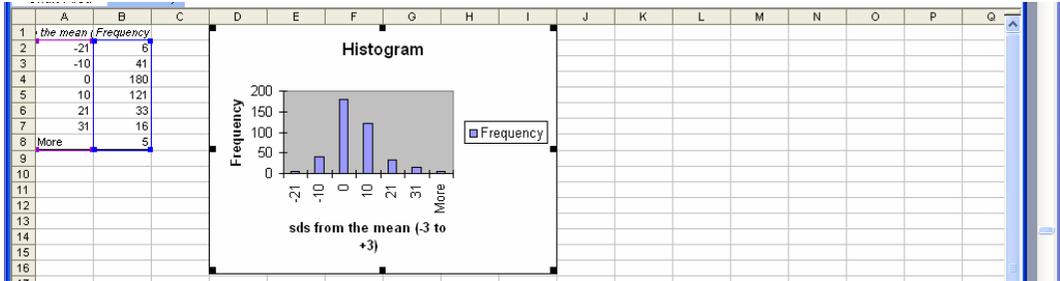


Move *return %* to **E** and re-run:



Residual line fit plots by the four continuous variables are each “cloud-like” and free of heteroskedasticity or patterns.

To assess the *Normality* of the residuals, make their histogram:



The residuals are approximately *Normal*.

From the model coefficient estimates, the regression is:

$$\begin{aligned}
 \text{TotalCompensation}(\$K)^5 &= 4.78 + 13.3\text{Computers \& Communication} + 7.67\text{energy} \\
 &\quad (4.69) \quad (2.03) \qquad\qquad\qquad (2.23) \\
 &\quad + 11.3\text{financial} + 12.6\text{health} + 4.72\text{food} + .307\text{age} \\
 &\quad (1.7) \qquad\qquad (2.2) \qquad\qquad (2.07) \\
 &\quad + 4.03 \ln \text{Revenues}(\$B) + .00399\text{profits}(\$MM) + .107\text{return}\% \\
 &\quad (.65) \qquad\qquad\qquad (.00128) \qquad\qquad\qquad (.030)
 \end{aligned}$$

To see predicted compensation values, add two columns following the regression variables.

Select **N** and **O**, **Alt HIC**, then add the label *predicted sqrt total compensation (\$K)* in **N1**.

Select the *Coefficients* from **B16:B26** of the regression worksheet, copy and paste into **O**.

Use the regression equation to enter the formula in **N2**.

Select the new cell **N2** and double click the lower right corner to fill in the column:

N2		= \$O\$2+\$O\$3*E2+\$O\$4*F2+\$O\$5*G2+\$O\$6*H2+\$O\$7*I2+\$O\$8*J2+\$O\$9*K2+\$O\$10*L2+\$O\$11*M2														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	F
	Wide	total	revenues	sqrt total	comput					ln	Profits	Return	predicted sqrt			
1	Industry	(K\$)	(B\$)	(K\$)	ers & commu	energy	financial	health	food	Age	(M\$)	5 Yrs	total	total	Coefficients me	
2	Compute	550.983	2.873	23.47302707	1	0	0	0	0	29	1.05536	-36	40	35.3852206	4.7791553	
3	Computer:	6398.7	16.963	79.99187459	1	0	0	0	0	54	2.83103	1022	39	54.32323238	13.334978	
4	Computer:	5805.38	1.909	76.1930443	1	0	0	0	0	53	0.64658	14.8	4	37.45943424	7.6701345	
5	Computer:	5739.2	0.274	75.75750788	1	0	0	0	0	48	-1.2946	81.2	41	32.30615117	11.250597	

The predictions are in square roots. To rescale back to the original scale in thousand dollars, square the square roots:

$$\begin{aligned}
 TotalCompensation(\$K) &= [TotalCompensation(\$K)^{-5}]^2 \\
 &= [4.78 + 13.3Computers \& Communication + 7.67energy \\
 &\quad + 11.3financial + 12.6health + 4.72food + .307age \\
 &\quad + 4.03\ln(revenues(\$B)) + .00399profits(\$MM) + .107return\%]^2
 \end{aligned}$$

Add a new column **O**: predicted total compensation (\$K)

In **O2** enter =N2^2.

Select the new cell **O2** and double click the lower right corner to fill in the column:

O2		=N2^2																	
	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		
1	energy	financial	health	food	Age	ln revenues (\$B)	Profits (MM\$)	Over 5 Yrs	Return %	predicted total compensation (K\$)	sqrt total compensation (K\$)	predicted total compensation (K\$)	Coefficients	mean	standard deviation	from the mean (-3 to 3)	ln total compensation (K\$)	sqrt revenues (\$B)	profits (MM\$) inverse
2	0	0	0	0	29	1.05536	-36	40	35.3852206	1252.113834	4.7791553	0.07	0.206868	-0.339	6.3117	1.69499	-0.028		
3	0	0	0	0	54	2.83103	1022	39	54.32323238	2951.013576	13.334978			-0.133	8.76385	4.11862	0.001		
4	0	0	0	0	53	0.64658	14.8	4	37.45943424	1403.209213	7.6701345			0.0743	8.66654	1.38167	0.0676		
5	0	0	0	0	48	-1.2946	81.2	41	32.30615117	1043.687403	11.250597			0.2812	8.65508	0.52345	0.0123		

The first executive in the dataset, Michael Dell, from the *computer & communications* industry, at age 29, from a firm (Dell) that reported revenues of \$2.87 billion, profits of -\$35.8 million, and a five year return of 40% is expected to earn total compensation of \$1,250 thousand, or \$1.25 million. This executive actually earned about half this amount, \$551 thousand, reminding us that we have accounted for just over a third of the variation in compensation packages.

### Excel 11.2 Consider synergies in sensitivity analysis with a nonlinear model

To isolate the importance of a dimension in driving compensation, we will compare expected total compensation of hypothetical executives which differ along only that dimension.

**Marginal impact of revenues.** To determine the difference in compensation driven by firm *revenues* in an industry, add three new rows to the dataset which describe three hypothetical executives

- from the *same* industry,
- of the *same* (median) age,
- from firms with *identical* (median) profits and returns.

The three hypothetical executives will differ only with respect to their firm's *revenues*.

- One will head a smaller firm with *revenues* at the 25% in the sample;
- the second will lead a larger firm with *revenues* at the 75%, and
- the third will manage a firm with median *revenues*.

We will use the *financial* industry as an example.

**Find representative values of predictors.** First, find the

- 75% largest revenues (\$B) with the Excel function **PERCENTILE(array, percentile)**, entering .75 for **percentile**,
- median revenues (\$B), with the Excel function **MEDIAN(array)**, and
- 25% largest revenues (\$B) with Excel function **PERCENTILE(array, percentile)**, entering .25 for **percentile**.

In **A408:A410**, type in labels *75%*, *median*, and *25%*.

In **C408**, enter **=PERCENTILE(C2:C403, .75)[Enter]**.

In **C409**, enter **=MEDIAN(C2:C403)[Enter]**.

In **C410**, enter **=PERCENTILE(C2:C403, .25)[Enter]**.

Select the new cells **C408:C410**, **Shift+→** through column **M**, **Cntl+R** to fill in the statistics:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	F
1	Industry	total compensation (K\$)	revenues (\$B)	sqrt total compensation (K\$)	computers & communication	energy	financial	health	food	Age	In revenues (\$B)	Profits (MM\$)	Return % 5 Yrs	predicted total compensation (K\$)		
403	Utility	280.396	1.076	16.74502911	0	0	0	0	0	60	0.07325	107.2	14	25.39300304		
404																
405	skewness	1.542100737	6.13373	0.788527141	2.0039	2.782	0.50461	2.64	2.22	-0	0.21731	4.435				
406	mean	1468.017328	4.37984	36.01508293	0.1468	0.095	0.37811	0.102	0.13	55	0.63365	272.2	16.57			0
407	sd	1098.362619	8.35835	13.09035492	0.3543	0.293	0.48552	0.303	0.34	7	1.2764	537.2	19.55			#
408	75%		4.28075	44.49535632	0	0	1	0	0	60	1.45413	292.5	20			
409	median		1.945	33.07680161	0	0	0	0	0	56	0.66524	114.2	14			
410	25%		0.70825	26.4053167	0	0	0	0	0	51	-0.345	51.33	8			
411																

**Add hypotheticals.** Describe three hypothetical executives in the financial industry.

Use shortcuts to add three new rows **404:406**: select rows **404:406**, **Alt HIR**.

To describe identical executives, enter

- identical values for industry indicators, and
- identical, median values for
  - *age*,
  - *profits*, and
  - *return*.

Make the *financial* column indicator values in **G404:G406** equal to one and enter zeros for the remaining indicators in **E, F, H, and I** in rows **404:406**.

In **J404:J406**, enter median age, 56.  
 In **L404:L406**, enter median profits (\$MM), 114.  
 In **M404:M406**, enter median return %: 14.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Wide Industry	total compensation (K\$)	revenues (\$B)	predicted total compensation (K\$)	sqrt total compensation (K\$)	computers & communication	energy	financial	health	food	Age	In revenues (\$B)	Profits (MM\$)	Return % Over 5 Yrs	predicted sqrt total compensation (K\$)
402	Utility	314.382	2.066	748.2965052	17.73082062	0	0	0	0	0	55	0.72561	297.2	15	27.35500878
403	Utility	280.396	1.076	644.8046032	16.74502911	0	0	0	0	0	60	0.07325	107.2	14	25.39300304
404	financial					0	0	1	0	0	56		114	14	
405						0	0	1	0	0	56		114	14	
406						0	0	1	0	0	56		114	14	

Allow the hypothetical executives' firm revenues to vary, from large (75%) to small (25%).

In **C404**, enter the 75% of revenues (\$B): 4.28,  
 In **C405**, enter median revenues (\$B): 1.95, and  
 In **C406**, enter the 25% of revenues (\$B): .707.

Drag *In revenues (B\$)* in **K** down through the three new rows.

L406		fx =LN(C406)													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Wide Industry	total compensation (K\$)	revenues (\$B)	predicted total compensation (K\$)	sqrt total compensation (K\$)	computers & communication	energy	financial	health	food	Age	In revenues (\$B)	Profits (MM\$)	Return % Over 5 Yrs	predicted sqrt total compensation (K\$)
402	Utility	314.382	2.066	748.2965052	17.73082062	0	0	0	0	0	55	0.72561	297.2	15	27.35500878
403	Utility	280.396	1.076	644.8046032	16.74502911	0	0	0	0	0	60	0.07325	107.2	14	25.39300304
404	financial		4.28			0	0	1	0	0	56	1.45395	114	14	
405			1.95			0	0	1	0	0	56	0.66783	114	14	
406			0.7			0	0	1	0	0	56	-0.3567	114	14	

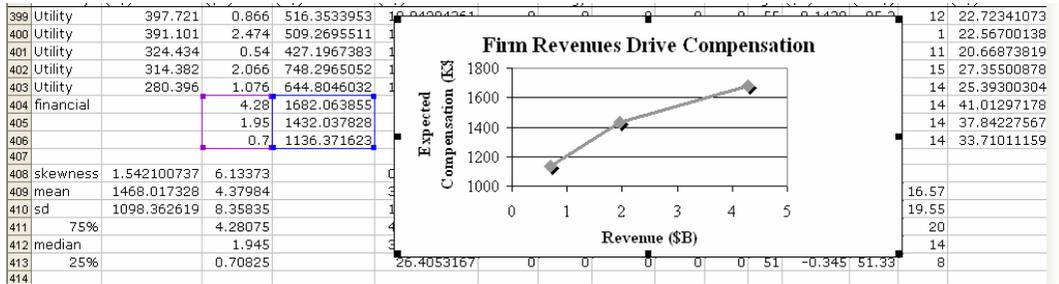
Drag *predicted sqrt total compensation (\$K)* and *predicted total compensation (\$K)* in **N** and **O** down through the three new rows.

O406		fx =N406^2													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Wide Industry	total compensation (K\$)	revenues (\$B)	sqrt total compensation (K\$)	computers & communication	energy	financial	health	food	Age	In revenues (\$B)	Profits (MM\$)	Return % Over 5 Yrs	predicted sqrt total compensation (K\$)	predicted total compensation (K\$)
399	Utility	397.721	0.866	19.94294361	0	0	0	0	0	55	-0.1439	95.3	12	22.72341073	516.3533953
400	Utility	391.101	2.474	19.77627366	0	0	0	0	58	0.90584	-943	1	22.56700138	509.2695511	
401	Utility	324.434	0.54	18.01205152	0	0	0	0	55	-0.6162	84.5	11	20.66873819	427.1967383	
402	Utility	314.382	2.066	17.73082062	0	0	0	0	55	0.72561	297.2	15	27.35500878	748.2965052	
403	Utility	280.396	1.076	16.74502911	0	0	0	0	60	0.07325	107.2	14	25.39300304	644.8046032	
404	financial		4.28		0	0	1	0	0	56	1.45395	114	14	41.01297178	1682.063855
405			1.95		0	0	1	0	0	56	0.66783	114	14	37.84227567	1432.037828
406			0.7		0	0	1	0	0	56	-0.3567	114	14	33.71011159	1136.371623

For a financial industry executive of the median age of 56, heading a firm with median profits and return percentage, the revenue difference of \$3.58B between small and large firms makes an expected difference of \$546K (= \$1,682K - \$1,136K) in compensation. Executives from larger firms earn as half a million dollars more.

**Illustrate the marginal response.** To see this expected compensation response to differences in revenues, rearrange columns so that *predicted total compensation (\$K)* follows *revenues (\$B)*.

Select the six cells **C404:D406**, **Alt ND**, and add a title and axes labels:



## *Lab Practice 11*

The Board of a firm in the Communications and Computer industry would like to know whether executive compensation packages in their industry are tied to firm performance or executive age. **Lab Practice 11 Executive Compensation CC.xls** contains data on the largest firms in the industry. Follow the steps in **Excel 11.1** and **Excel 11.2** to build a model of executive compensation for The Board. Since all firms are in the same industry, you will not need to add industry indicators.

Which variables are positively skewed? \_\_\_\_\_

Which variable is negatively skewed? \_\_\_\_\_

Which scale, square roots or natural logarithms, *Normalizes* each of the positively skewed variables better?

\_\_\_\_\_

Does rescaling the negatively skewed variable to squares make the variable more *Normal*?  
Y or N

Write your model equation in thousand dollars (\$K) of *compensation*:

Make a table of the marginal impacts of the two significant drivers.

Make a scatterplot to illustrate the marginal impact of the most important driver on *compensation* and attach to your lab practice worksheet.

## *CASE 11-1 Global Emissions Segmentation: Markets Where Hybrids Might Have Particular Appeal\**

Carbon emissions policies are being watched carefully by Ford Motor Company. Ford executives believe that major markets for new hybrid models will arise in developing countries where increased economic productivity and growing population stimulate demand for vehicles.

To reduce carbon emissions, the Kyoto Protocol went into effect Feb. 16, 2005, with 141 countries signing on, including every major industrialized country, except the United States, Australia and Monaco. The Protocol stipulates conditions for systematically reducing carbon emissions. Some of the world's biggest and fastest growing polluters, including China and India, have not signed the Kyoto Protocol. Because they are considered developing countries, they are outside the Protocol's framework. Yet the publicity about the Kyoto Protocol has heightened interest in Carbon Emissions Reductions (CERs). A number of countries have publicized their expected CERs, shown in the table below:

**Case 11-1 Global Carbon Emissions.xls** contains data from 68 countries with measures of

- *Carbon Emissions,*
- *GDP,*
- *Population,*
- *Vehicle Registrations,* and
- *Barrels of Crude Oil Produced per Day,*
- *two indicators of global region: Indo Asia (India, Pakistan and Bangladesh) and Asia.*

(*Other global regions is the **baseline.***)

Ford executives have asked you to confirm that *vehicle registrations* affect *carbon emissions*. They would like to know, specifically, how important the influence of *vehicle registrations* is, relative to *GDP*, *Population* and *oil* production in the global regions which include India and China. If you can confirm that vehicle registrations are an important influence on carbon emissions, Ford will use that information to promote the manufacture and marketing of their hybrid models in China and India.

Build a model of carbon *emissions* to provide Ford with answers.

---

\*This example is a hypothetical scenario using actual data.

<b>Expected Average Annual CERs from registered projects by host party.</b> (Source: Clean Development Mechanism (CDM), <a href="http://cdm.unfccc.int">cdm.unfccc.int</a> , 10 Feb 07)					
Country	Average Annual Reduction Expected	Country	Average Annual Reduction Expected	Country	Average Annual Reduction Expected
<b>China</b>	46,500,229	Colombia	414,205	Costa Rica	162,515
Brazil	15,846,288	El Salvador	360,268	Dominican Republic	123,916
<b>India</b>	15,534,244	Ecuador	357,900	Sri Lanka	109,619
Korea	12,362,308	Nicaragua	336,723	Israel	101,617
Mexico	5,566,398	Guatemala	279,694	Panama	96,469
Chile	2,183,123	Papua New Guinea	278,904	Nepal	93,883
Argentina	1,765,007	Philippines	247,885	Bolivia	82,680
Malaysia	1,682,653	South Africa	225,446	Cyprus	72,552
Indonesia	1,557,100	Morocco	223,313	Jamaica	52,540
Nigeria	1,496,934	Honduras	205,251	Cambodia	51,620
Egypt	1,436,784	Peru	199,265	Moldova	47,343
Pakistan	1,050,000	Armenia	197,832	Fiji	24,928
Tunisia	687,573	Bangladesh	169,259	Mongolia	11,904
Viet Nam	681,306	South Africa	225,446	Bhutan	524

1. Which variables are positively skewed? \_\_\_\_\_
2. Which scale, square roots or natural logarithms, is the better choice for each positively skewed variable? (A better scale will produce fewer outliers.)  
  
\_\_\_\_\_
3. Write your model equations in the original scale of carbon *emissions* for *Indo Asia*, *Asia*, and *Other* global regions outside Asia.

4. The segment with greatest potential for hybrid sales will be those countries with **high GDP**, **high population**, and **high vehicle registrations**.

Which countries have:

- *GDP* at or above the **75<sup>th</sup> percentile**
- *Population* at or above the **75<sup>th</sup> percentile**

AND

- *Vehicle registrations* at or above the **75<sup>th</sup> percentile**
- 

5. Make a table comparing the marginal impacts of *GDP*, *Population*, and *Vehicle Registrations* for each of the three global regions, *Asia*, *Indo Asia*, and *Other*, and explain the table:

- Assuming median *population* and *vehicle registrations*, what difference in *emissions* is expected between countries with **lowest and highest GDP** in *Asia*? In *Indo Asia*?
- Assuming median *GDP* and *vehicle registrations*, what difference in *emissions* is expected between countries with **lowest and highest population** in *Asia*? In *Indo Asia*?
- Assuming median *GDP* and *population*, what difference in *emissions* is expected between countries with **lowest and highest vehicle registrations** in *Asia*? In *Indo Asia*?

Attach a scatterplot showing the marginal impacts of *vehicle registrations* on carbon *emissions* in each of the three global regions, *Asia*, *Indo Asia*, and *Other*.

Write a paragraph to summarize your key results and their implications for Ford management.

## 12

# Indicator Interactions for Structural Differences or Changes in Response

In this chapter, we explore indicator interactions with predictors. Adding this type of interaction to models allows us to capture differences in response between segments or changes in response following structural changes or shocks. Indicator interactions alter partial slopes, in the way that indicators alter intercepts.

### 12.1 Indicator Interaction with a Continuous Influence Alters Its Partial Slope

At times, segment average response levels, the intercepts, *and* responses to an influence, the partial slopes, differ. Two segments may respond differently to an influence. In marketing, segmentation is a basic principal. Customer segments respond differently to prices, advertising and product characteristics.

In time series models, a structural shift may alter the partial response to a continuous influence. The impact of economic productivity on business performance may differ by Party leadership. More households may donate to charitable organizations following a natural disaster.

In such cases, where segment responses differ, or structural shifts alter responses, we add one or more interactions, each equal to the product of an indicator and a continuous predictor.

To model differences between two segments' responses to a driver  $X$ , we add an indicator for one of the two segments, and make a new interaction variable which is the product of the indicator and the driver  $X$ :

$$\hat{Y} = b_0 + b_1 \text{Segment}_1 + b_2 X + b_3 \text{Segment}_1(X)$$

To model change in response following a structural shift, we add an indicator of the structural shift and make a new interaction variable which is the product of the shift indicator and the driver  $X_t$ :

$$\hat{Y}_t = b_0 + b_1 \text{Shift}_t + b_2 X_t + b_3 \text{Shift}_t(X_t)$$

When the indicator is zero, representing baseline segment response in a cross-sectional model, or baseline response before a structural shift in a time-series model, the equations are:

- In a cross-sectional model:

$$\begin{aligned}\hat{Y} &= b_0 + b_1(0) + b_2X + b_3(0)X \\ &= b_0 + b_2X\end{aligned}$$

- In a time series model:

$$\begin{aligned}\hat{Y}_t &= b_0 + b_1(0) + b_2X_t + b_3(0)X_t \\ &= b_0 + b_2X_t\end{aligned}$$

When the indicator is one, representing a second segment's response in a cross-sectional model, or response following a structural shift in a time-series model, the equations become:

- In a cross-sectional model:

$$\begin{aligned}\hat{Y} &= (b_0 + b_1(1)) + (b_2 + b_3(1))X \\ &= (b_0 + b_1) + (b_2 + b_3)X\end{aligned}$$

- In a time series model:

$$\begin{aligned}\hat{Y}_t &= (b_0 + b_1(1)) + (b_2 + b_3(1))X_t \\ &= (b_0 + b_1) + (b_2 + b_3)X_t\end{aligned}$$

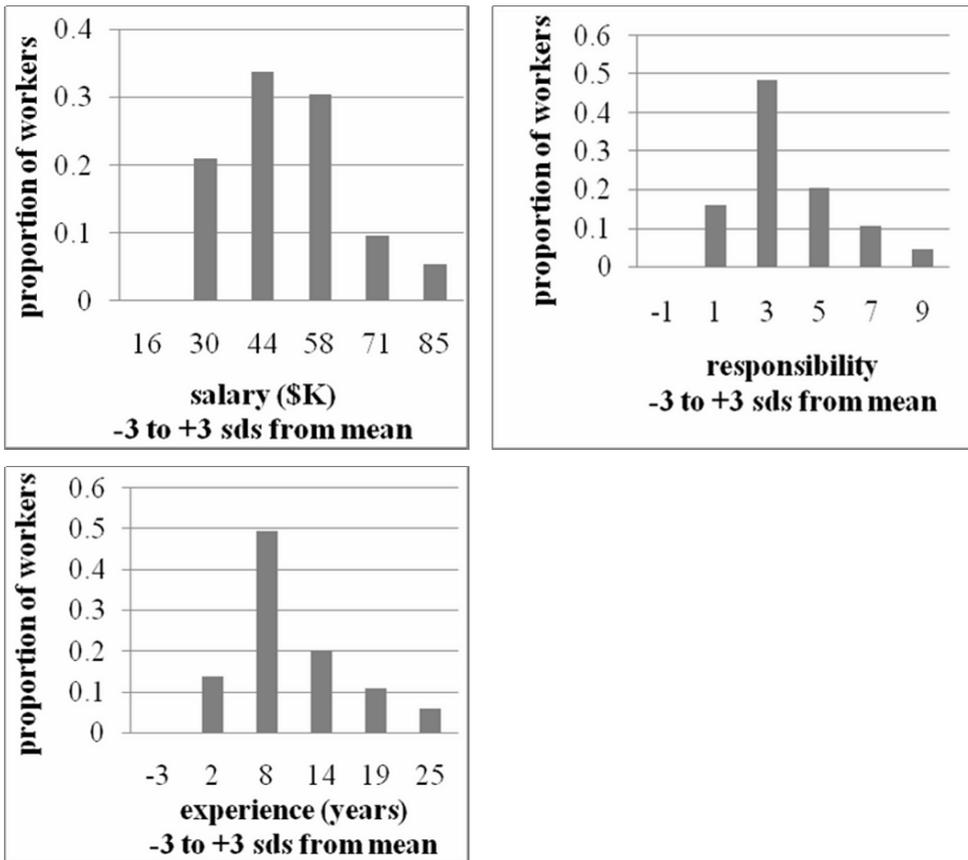
The indicator alters the average level of response, by adjusting the intercept from  $b_0$  to  $b_0 + b_1$ , and the indicator interaction alters the response to variation in the predictor, by adjusting the partial slope from  $b_2$  to  $b_2 + b_3$ .

*Example 12.1 Gender Discrimination at Slams Club.* A disgruntled Slam's Club employee resigned and decided to sue the firm on grounds of gender discrimination. She alleges that Slams Club pays female employees less than male employees. The Slam's Club Board asked a consultant, Morey Furless, to build a model to assess gender discrimination.

Slams Club executives admitted that women were encouraged to work part time and focus on their roles as homemakers, rather than pursuing long term careers. They maintain that women are paid equally to men in similar positions. Following the meeting with executives, Morey made a note to be sure to include level of responsibility in the model.

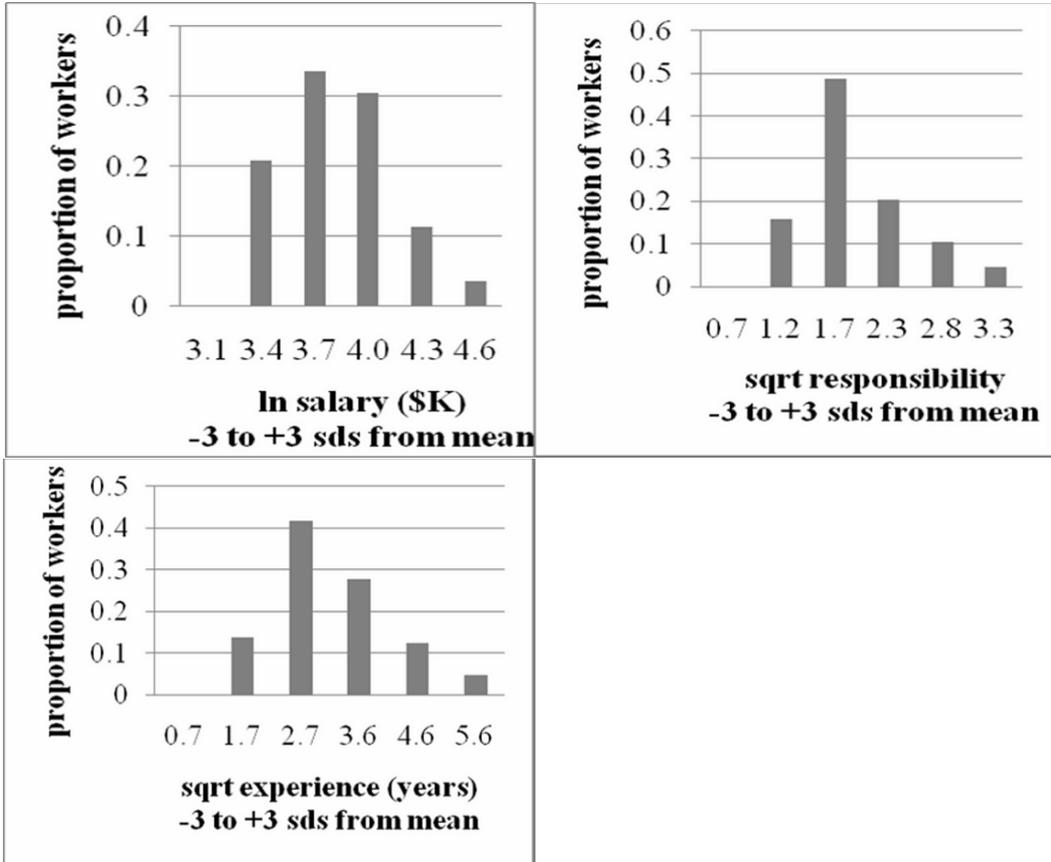
From a random sample of 220 employee records, Morey built a model of salaries, including level of responsibility, years of experience, and an indicator for gender. Since it is possible that the value of responsibility and gains from experience each differ across the genders, interactions between the gender indicator and these two continuous variables were included.

**Examine skewness of the model variables to choose scales.** Examining the distributions of *Responsibility*, *Experience* and *Salary*, shown in Figure 12.1, Morey found that *salary*, *responsibility*, and *experience* were positively skewed:



	<i>salary</i> (\$K)	<i>responsibility</i> (1 to 9)	<i>experience</i> (years)
<i>Skewness</i>	1.01	1.01	1.05

Figure 12.1 Distribution of responsibility, experience and salary



	<i>ln salary(K\$)</i>	<i>Sqrt responsibility</i>	<i>Sqrt experience</i>
<i>Skewness</i>	0.32	0.45	0.39

Figure 12.2 Rescaled variables

To reduce positive skew, we shrink, rescaling to square roots or natural logarithms. The natural logarithms better *Normalize salary*, but are too extreme for *responsibility* and *experience*. The square roots of *responsibility* and *experience* *Normalize* without overcorrecting. These are shown in Figure 12.2.

When a dependent variable is rescaled, the model features built-in synergies. With rescaled *salary(\$K)*, this salary model will feature built-in synergies between gender, *responsibility* and years of *experience*. Regression results are in Table 12.1.

## SUMMARY OUTPUT

<i>Regression Statistics</i>						
Multiple R	0.909					
R Square	0.827					
Adjusted R Square	0.823					
Standard Error	0.125					
Observations	220					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	5	15.87	3.17	204.6	0.0000	
Residual	214	3.32	0.02			
Total	219	19.19				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2.738	0.055	49.4	0.0000	2.629	2.847
<i>male</i>	-0.087	0.073	-1.2	0.2376	-0.231	0.058
<i>Responsibility</i> <sup>5</sup>	0.270	0.031	8.6	0.0000	0.208	0.332
<i>Experience</i> <sup>5</sup>	0.208	0.037	5.6	0.0000	0.135	0.282
<i>male x responsibility</i> <sup>5</sup>	0.253	0.023	11.2	0.0000	0.208	0.297
<i>male x experience</i> <sup>5</sup>	-0.172	0.025	-6.8	0.0000	-0.222	-0.122

Table 12.1 Gender differences in the value of responsibility and experience at Slam's Club

The *male* indicator is not significant, though the interactions between *male* and *responsibility* and *experience* are significant, so *male* remains in the model. We cannot include an interaction without its components, the indicator and the *main effect*, since the interaction is relative to the baseline main effect.

Morey's model is:

$$\ln(\hat{\text{salary}}(\$K)) = 2.74^a - .087 \text{ male} + (.27^a + .253^a \text{ male}) \text{ responsibility}^5 \\
\quad \quad \quad (.06)(.073) \quad \quad \quad (.03) (.023) \\
+ (.208^a - .172^a \text{ male}) \text{ experience}^5 \\
\quad \quad \quad (.037) (.025)$$

*RSquare*: .83<sup>a</sup>

<sup>a</sup>Significant at .01.

To rescale back to the original thousand dollars, we use the exponential function to undo the natural logarithms:

$$\exp(\ln(\hat{salary}(\$K))) =$$

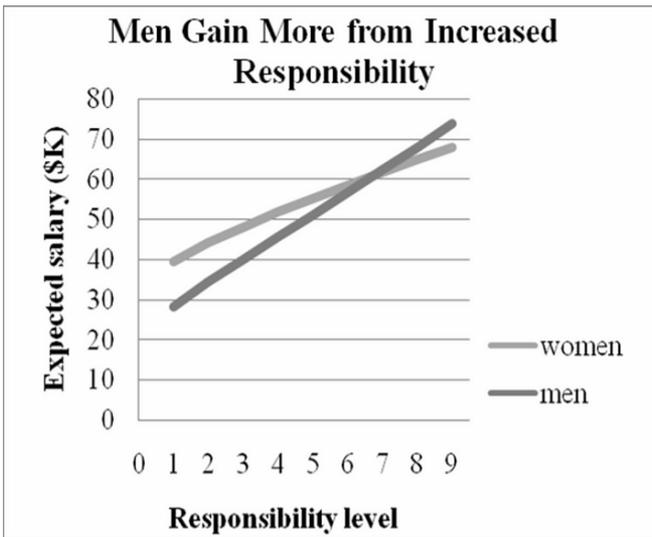
$$\hat{salary}(\$K) = \exp[ 2.74^a - .087 \text{ male} + (.27^a + .253^a \text{ male}) \text{ responsibility}^5 \\ + (.208^a - .172^a \text{ male}) \text{ experience}^5 ]$$

By setting *male* to zero, the model for women can be written as:

$$\hat{salary}(\$K) = \exp[ 2.74^a + .27^a \text{ responsibility}^5 + .208^a \text{ experience}^5 ]$$

and by setting *male* to one, the model for men can be written as:

$$\hat{salary}(\$K) = \exp[ 2.74^a - .087 + (.27^a + .253^a) \text{ responsibility}^5 \\ + (.208^a - .172^a) \text{ experience}^5 ] \\ = \exp[ 1.87 + .52 \text{ responsibility}^5 + .036 \text{ experience}^5 ]$$



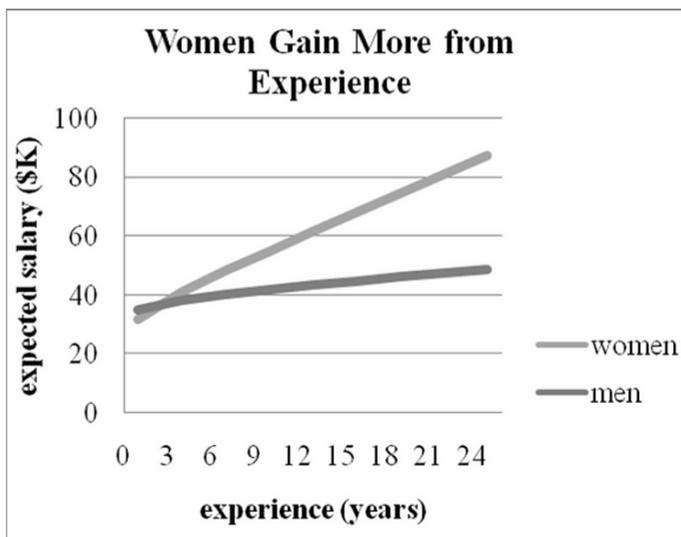
**The interaction between gender and responsibility.** In Figure 12.3, we see that among employees with median years of *experience*, seven, women (shown by the lighter gray curve) are paid more than men in positions of lower *responsibility*, though men gain more from promotion.

*Salary* response to increasing level of *responsibility* is increasing at a diminishing rate for women (but not for men).

Figure 12.3 Salaries (K\$) by Responsibility and Gender.

Women in middle management (*responsibility* level 5) can expect to be paid about \$16K more than staff (*responsibility* level 1), but women in upper management (*responsibility* level 9) can expect to be paid only about \$13K more than middle management. Men in middle management can expect to earn about \$23K more than staff, and men in upper management can also expect to be paid about \$23K more still.

Men's and women's response curves are not parallel. Men benefit more from increased *responsibility*. Men can expect to gain an average of \$5.5K (= \$5,500) from promotion to level 5 from level 4; a woman can expect to gain an average of about \$3.4K (= \$3,400) from a similar promotion.



At the median level of *responsibility*, 3, women benefit from increasing *experience*, illustrated in Figure 12.4. Women with ten years of *experience* can expect to be paid about \$23,100 more. Gains from *experience* are greater among women with less *experience*.

Figure 12.4 Salary by years *experience*; the interaction between gender and *experience*.

Experienced men with median *responsibility* are rewarded less, perhaps because they are expected to advance in rank. Men with ten years of *experience* can expect to be paid about \$6,690 more than men with five years of *experience*.

Morey was confident that Slam's Club executives would be relieved with his model results, which are summarized in the memo below:

**MEMO**

**Re: Women are Paid More than Men at Slam’s Club**  
**To:** The Board  
**From:** Morey Furless, Morey Furless Consulting Associates  
**Date:** June 2007

Analysis of a random sample of 220 Slam’s Club employee salaries reveals that women are paid more than men. Level of responsibility is stronger salary driver than gender.

**Salary Model.** Using data from 220 randomly selected employee records, a model linking salary, employee responsibility level and tenure was built.

**Model Results.** Gender, level of *responsibility*, and employee tenure account for 83% of the variation in salaries.

On average, male and female employees are paid equally, though women are paid more for greater tenure.

Women with ten years tenure earn an average of \$13,000 more than men with the same tenure.

Level of responsibility also drives salaries. Middle management workers (*responsibility* level 5) can expect to be paid an average of about \$19K more than staff (*responsibility* level 1).

Men do benefit more from promotion to higher levels of *responsibility*.

A man can expect to gain an average of \$5.5K from promotion to level 5 from level 4; a woman can expect to gain an average of about \$3.4K from a similar promotion.



$$\begin{aligned} \hat{salary}(\$K) &= \exp[ 2.74 + .27 \text{ responsibility}^5 \\ &\quad + .208 \text{ experience}^5 ] \\ &\quad \text{for women} \\ &= \exp[ 1.87 + .52 \text{ responsibility}^5 \\ &\quad + .036 \text{ experience}^5 ] \\ &\quad \text{for men} \end{aligned}$$

RSquare: .83<sup>a</sup>  
<sup>a</sup>Significant at .01.

**Conclusions.** Slam’s Club does not discriminate against women. Female employees are paid more than men for their years of loyal service.

**Limitations.** This model does not explore issues related to equal opportunities for promotion to greater responsibility levels. Responsibility is a major driver of salaries. In the case that more men hold positions with higher responsibility levels, this could be considered discriminatory against women.

*Example 12.2 Car Sales in China.* Every major car manufacturer is watching China closely. As China's GDP grows rapidly, more and more Chinese consumers are buying cars. Some of those cars are imports manufactured outside China and some are the products of joint ventures between Chinese and American partners. Some cars produced in China are exported, particularly to other Asian countries where labor costs are higher. We will build a model of car sales in China based on a Leading Indicator, past year Chinese Car Production, and political leadership in China. We will include two indicators to represent changes in car sales from the baseline years 1990 through 1996, when Deng led China and set import-export policy:

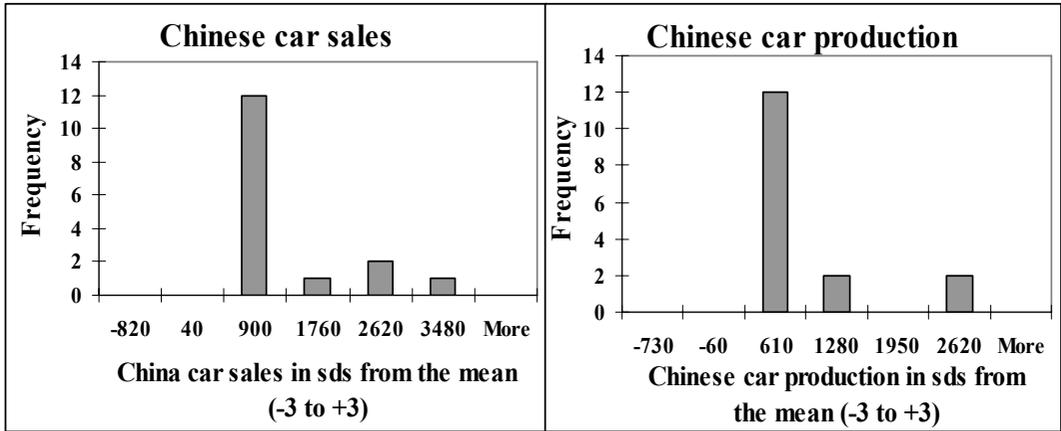
- For the period 1997 to 2002: *after Deng*, to represent Third Generation leadership following Deng's death in early 1997
- For the period 2003 through 2011: *Fourth Generation*,

We will also include an indicator of *Tiananmen Square* to assess its five-year impact on car sales.

Political leadership probably affects car sales response to car production, since imports and exports are either encouraged or discouraged by particular administrations. For this reason an indicator interaction between *after Deng* and past year Chinese car production will be included. The interaction between the *Fourth Generation* indicator and Chinese car production would be useful, though this leadership period began in the last year of the validation data, 2003, which does not provide enough information to include an indicator interaction.

Data contains time series of annual observations from 1989 through 2005 on *car sales in China*, *Chinese car production*, indicators for *Tiananmen Square*, Third Generation leadership *after Deng's* death, and *Fourth Generation* leadership.

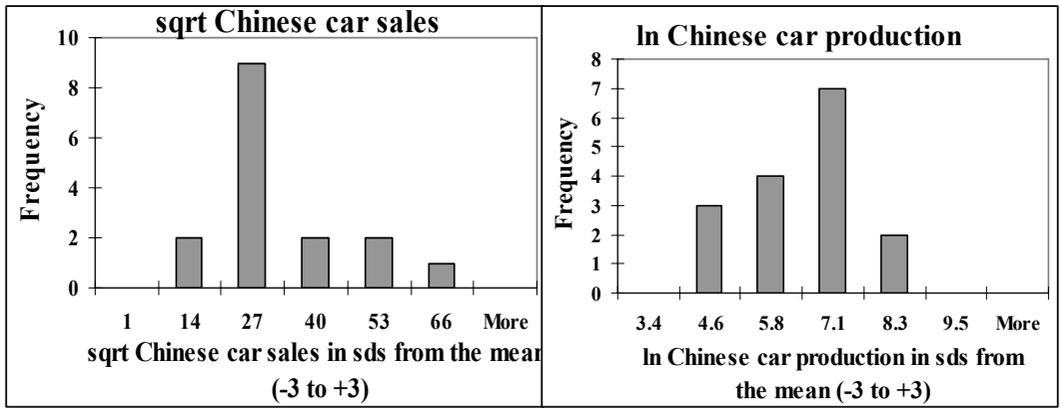
Both continuous variables, car sales and car production, shown in Figure 12.5, are positively skewed, suggesting that we shrink each by rescaling to square roots, natural logarithms, or inverses.



	<i>Chinese car sales</i>	<i>Chinese car production</i>
<i>Skewness</i>	1.50	1.64

Figure 12.5 Skewed dependent and independent variables

The square roots of *Chinese car sales* and the natural logarithms of past year *Chinese car production*, shown in Figure 12.6, reduce skewness:



	<i>sqrt Chinese car sales</i>	<i>ln Chinese car production</i>
<i>Skewness</i>	.95	-.14

Figure 12.6 Rescaled variables are less skewed

Because the dependent variable will be rescaled, the model will feature built-in synergies between predictors. The interaction terms will be products of rescaled independent variables and indicators.

The model correctly forecast car sales in China during the two most recent held out years, 2004 and 2005. Those two recent years were then included, and the model was recalibrated. Multiple regression results are in Table 12.2:

---

### SUMMARY OUTPUT

---

<i>Regression Statistics</i>						
Multiple R	0.996					
R Square	0.991					
Adjusted R Square	0.987					
Standard Error	1.48					
Observations	16					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	5	2483	497	226.4	0.0000	
Residual	10	22	2			
Total	15	2505				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-19.2	5.7	-3.4	0.007	-31.8	-6.5
<i>Tiananmen Square after Deng</i>	4.7	1.8	2.5	0.029	0.6	8.76
<i>Fourth Generation ln Chinese car production (K)<sub>t-1</sub> after Deng x ln Chinese car production (K)<sub>t-1</sub></i>	-153.1	32.6	-4.7	0.001	-225.7	-81
	15.9	2.1	7.5	0.000	11.2	20.6
	7.17	0.99	7.2	0.000	4.96	9.39
	24.4	5.2	4.7	0.001	12.9	35.9
<i>DW<sub>6,16</sub></i> :	2.16					

---

Table 12.2 Leadership and growth in car production drive growth in chinese car sales

From regression output, we can write the regression equation for the four distinct periods. In each case, we square both sides of the equation to rescale back to car sales in units:

- 1990 – 1994, following Tiananmen Square, with *Tiananmen Square* set to one, and *after Deng* and *Fourth Generation* set to zero:

$$\begin{aligned} \widehat{ChineseCarSales}(K)_t &= [-19.2 + 4.7(1) - 153.1(0) + 15.9(0) \\ &\quad + (7.2 + 24.4(0)) \ln(\text{production}(K))_{t-1}]^2 \\ &= [-14.5 + 7.3 \ln(\text{production}(K))_{t-1}]^2 \end{aligned}$$

- 1995-1996, after Tiananmen Square effects had subsided, before Deng's death, with all indicators set to zero:

$$\begin{aligned} \widehat{ChineseCarSales}(K)_t &= [-19.2 + 4.7(0) - 153.1(0) + 15.9(0) \\ &\quad + (7.2 + 24.4(0)) \ln(\text{production}(K))_{t-1}]^2 \\ &= [-19.2 + 7.3 \ln(\text{production}(K))_{t-1}]^2 \end{aligned}$$

- 1997 - 2002, Third Generation leadership after Deng's death, before Fourth Generation leadership, with the *after Deng* indicator set to one:

$$\begin{aligned} \widehat{ChineseCarSales}(K)_t &= [-19.2 + 4.7(0) - 153.1(1) + 15.9(0) \\ &\quad + (7.2 + 24.4(1)) \ln(\text{production}(K))_{t-1}]^2 \\ &= [-172.3 + 31.6 \ln(\text{production}(K))_{t-1}]^2 \end{aligned}$$

- 2003 - present, under Fourth Generation leadership, with the *Fourth Generation* indicator set to one:

$$\begin{aligned} \widehat{ChineseCarSales}(K)_t &= [-19.2 + 4.7(0) - 153.1(0) + 15.9(1) \\ &\quad + (7.2 + 24.4(0)) \ln(\text{production}(K))_{t-1}]^2 \\ &= [-4.3 + 7.2 \ln(\text{production}(K))_{t-1}]^2 \end{aligned}$$

Comparing intercepts, we see that for a given level of car production, car sales would be (and have been) highest in recent years under Fourth Generation Leadership. The impact of growth in car production is positive in all periods, but particularly strong in the period after Deng's death.

A scatterplot of the model fit in Figure 12.7 illustrates the changing patterns of car sales in China:

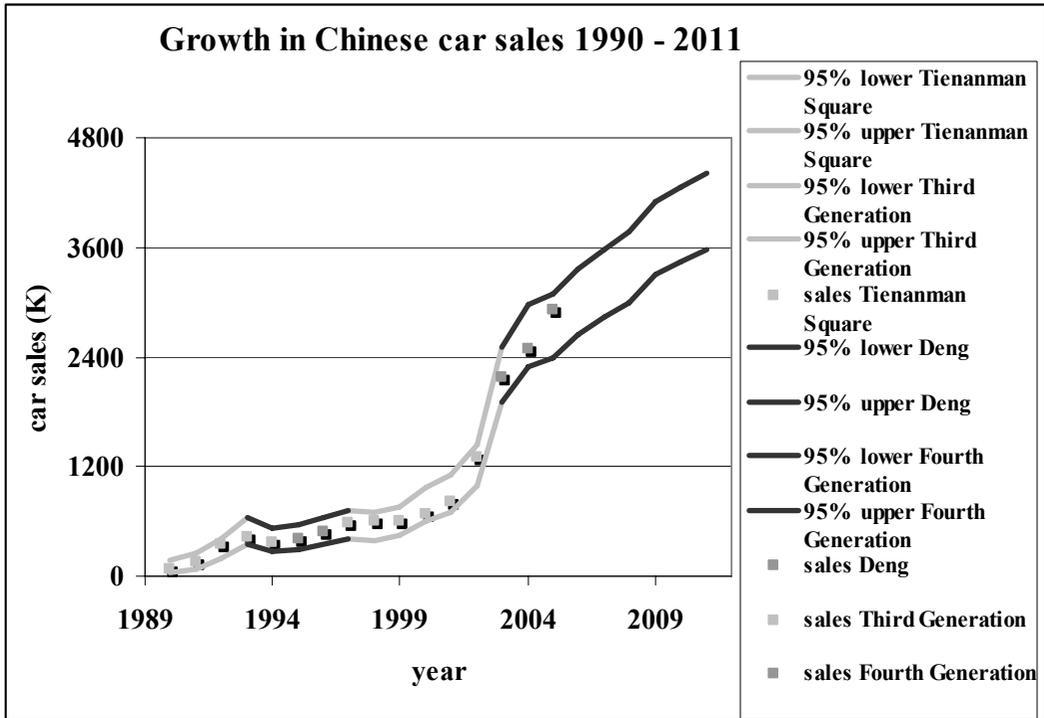
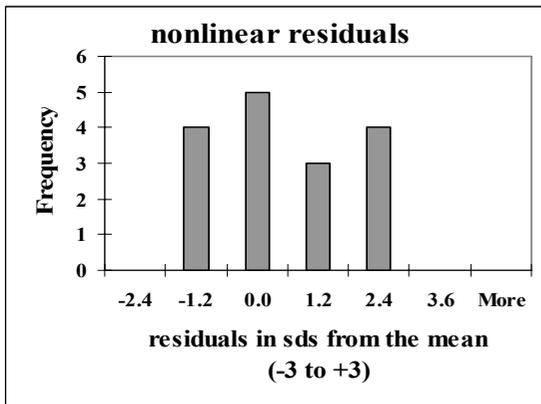


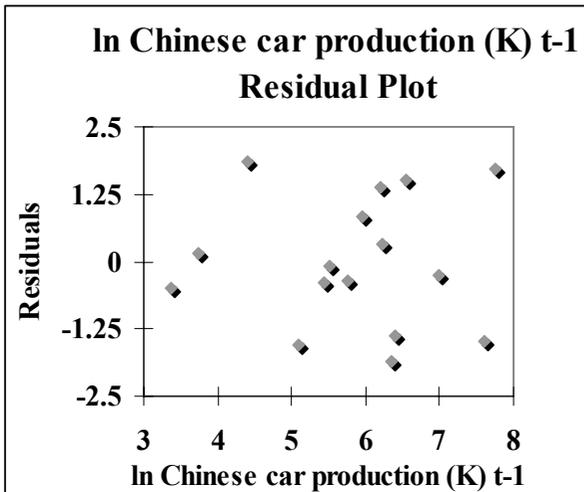
Figure 12.7 Growth in car sales in China

Residual analysis



The nonlinear model residuals, in Figure 12.8, are approximately *Normally* distributed.

Figure 12.8 Model residuals



Residuals plotted by car production in Figure 12.9, are also homoskedastic and pattern free.

Figure 12.9 Residuals by car production

### Sensitivity analysis: Fewer cars produced in China are being sold in China under Fourth Generation Leadership

To see the impact of Deng's death on Third Generation leadership and its interaction with past year *Chinese car production*, we can compare 95% mean prediction intervals for years 1998 through 2002 *had Deng's death occurred later*. Setting the *after Deng* indicator to 0 for those five years, the 95% mean prediction intervals are higher initially, but grow at a much slower pace, as Figure 12.10 and Table 12.3 illustrate. Following Deng's death, China became more market-driven, and *Chinese car sales* increased, noticeably by 2002.

We can consider a second hypothetical condition in which *Third Generation* leadership continued, rather than being replaced by the *Fourth Generation* in 2003. We accomplish this by setting the *after Deng* indicator to one and the *Fourth Generation* indicator to zero in those years. We see that *Fourth Generation* leadership led to noticeably reduced *Chinese car sales* levels, possibly because of the increased emphasis on car exports.

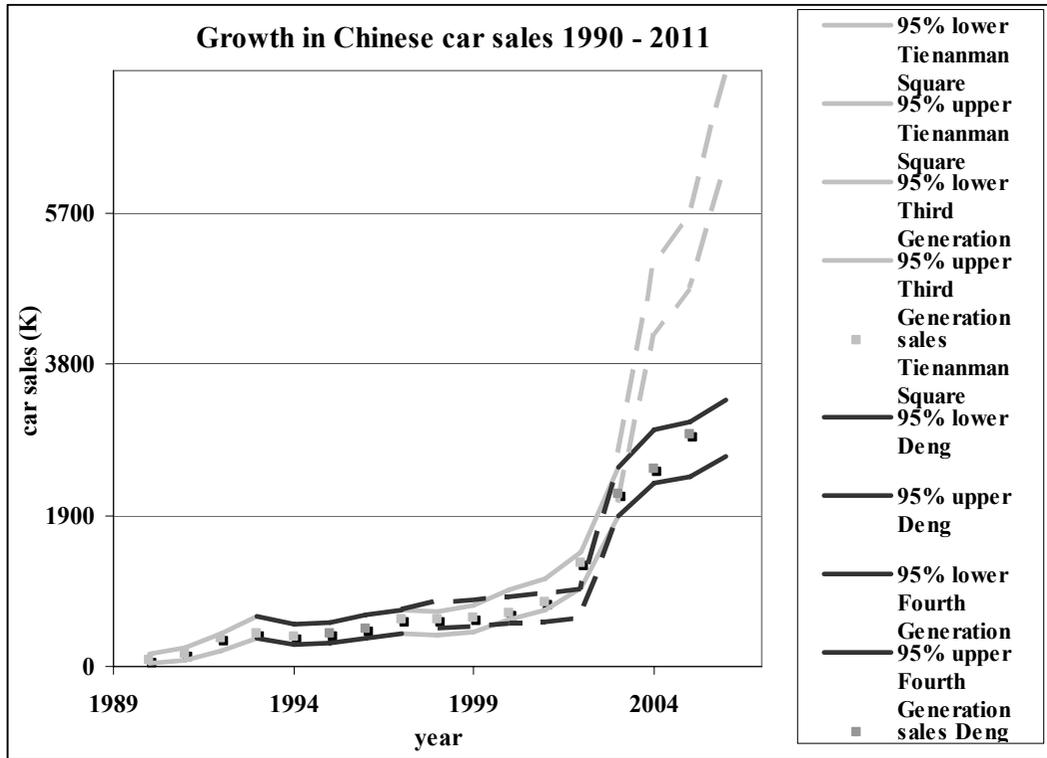


Figure 12.10 Growth in Chinese car sales under alternate leadership scenarios

<i>Predicted Car Sales in China (K) 1998 - 2006</i>				
<i>year</i>	<i>after Deng</i>	<i>Deng</i>	<i>change after Deng's death</i>	<i>% change</i>
1998	640	530	-100	-19%
1999	650	590	-60	-10%
2000	690	770	80	10%
2001	720	890	170	20%
2002	780	1200	420	35%
	<i>Fourth Generation</i>	<i>Third Generation hypothetical</i>	<i>change w Fourth Generation</i>	<i>% change</i>
2003	2200	2350	-150	-7%
2004	2620	4620	-2000	-76%
2005	2730	5230	-2500	-91%
2006	2990	6900	-3910	-131%

Table 12.3 Growth in Chinese car sales under alternate leadership

## 12.2 Indicator Interactions Capture Segment Differences or Structural Differences in Response

Segment responses can be expected to differ. Price discrimination and product differentiation strategies acknowledge this. By incorporating indicator interactions into our models, we add realism. Interactions also allow us to quantify differences in response across segments, improving the value of our results to decision makers.

In time series, structural shifts and shocks sometimes alter both the average level of response and the degree of response to changes in predictors. Adding interaction terms to models improves validity and predictive capability. Interaction terms also allow us to assess differences or changes in response to independent variables in a model. We can backcast to determine the impact of a structural change or shock, and then estimate what response would have been had the structural change or shock not occurred. We can forecast to determine the impact of similar shocks or changes in the future. Interaction terms increase the realism and value of our models.

## Excel 12.1 Add indicator interactions to capture segment differences or structural differences in response

*Car Sales in China.* We will build a model of car sales in China, including the Leading Indicator, past year Chinese car production, and indicators of the Tiananmen Square incident of 1989, Deng’s death in 1997, and the shift to Fourth Generation leadership in 2003. We will also include an interaction between the *after Deng* indicator and past year car production to allow for differences in import policies due to leadership. (The Fourth Generation shift occurred too recently to allow use in an interaction with car production.)

Data contained in **Excel 12.1 China Car Sales.xls** contain time series of annual observations from 1990 through 2005 on

- *car sales in China (K)*,
  - *Chinese car production (K) t-1* (past year),
- and indicators for
- *Tiananmen Square*,
  - *Third Generation leadership after Deng*, and
  - *Fourth Generation leadership*.

**Assess skewness to choose variable scales.** To build the most valid model, we will first rescale to reduce skewness of *Chinese car sales (K)* and *Chinese car production (K) t-1*, incorporating nonlinear, nonconstant response.

Find *skewness*, the *mean* and *standard deviation* of *Chinese car sales* and *Chinese car production*.

In **B25** enter **=SKEW(B2:B17)** [Enter].  
 In **B26** enter **=AVERAGE(B2:B17)** [Enter] and  
 In **B27** enter **=STDEV(B2:B17)** [Enter].  
 Select the new cells **B25:B27**, **Shift+→** through **C**, **Cntl+R**:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	year	Chinese car sales (K)	Chinese car production (K) t-1	Tiananmen Square	after Deng	Fourth Generation							
16	2004	2479	2019	0	0	1							
17	2005	2914	2316	0	0	1							
18	2006		3260	0	0	1							
19	2007		4180	0	0	1							
20	2008		5220	0	0	1							
21	2009		7600	0	0	1							
22	2010		9000	0	0	1							
23	2011		10500	0	0	1							
24													
25	skewness	1.49534	1.8011497										
26	mean	902	612										
27	standard d	860.4187	669.098105										

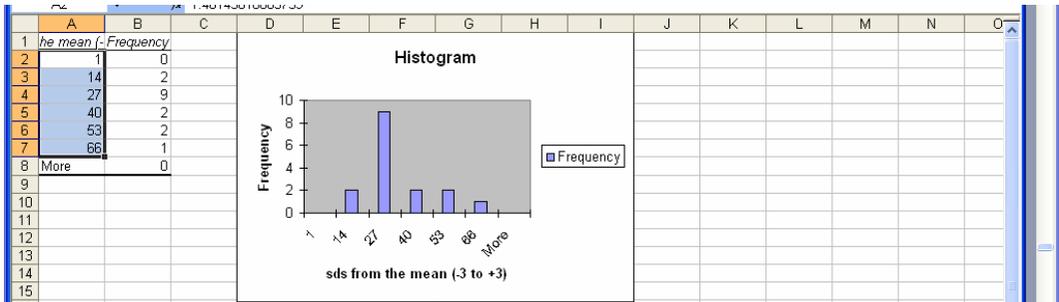
To see the distributions, make histograms of *Chinese car sales*



The square roots of *Chinese car sales* reduce skewness from 1.50 to .95, leaving some positive skew. The natural logarithms overcorrect, producing slight negative skew of -.14.

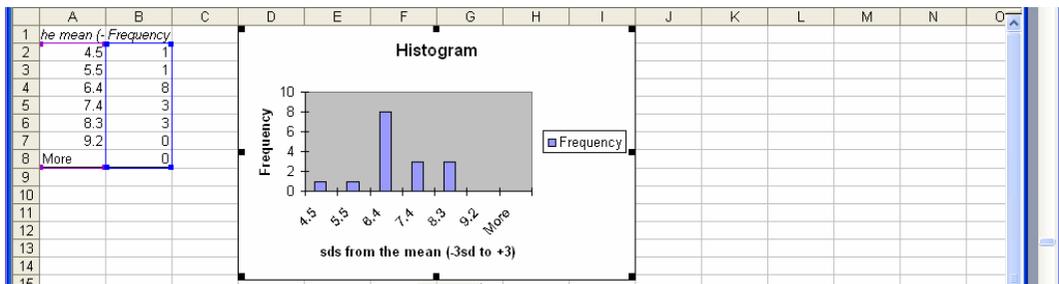
Square roots of *Chinese car production (K)* reduce skew from 1.80 to .89. Natural logarithms overcorrect, producing negative skew of -.51.

**Chinese car sales.** Both square roots and natural logarithms are acceptable options, since both produce skewness in the range -1 to 1. To compare and make a choice, make histograms of *sqrt Chinese car sales* and *ln Chinese car sales*:



The natural logarithm distribution contains a relatively large proportion of values more than two standard deviations above the mean. We will use the square roots of Chinese car sales.

**Chinese car production.** The square roots reduce skewness to .88, leaving some positive skew, while the natural logarithms overcorrect, producing negative skew, -.51.



We will use the natural logarithms with skewness closer to 0.

**Add indicator interactions.** To model varying car sales response to increasing car production by leadership regime, we will include an interaction between the indicator, *after Deng*, and *ln Chinese car production (K) t-1*, making the years before (1990 through 1997 under Deng) and after (2003 through 2011 under Fourth Generation leadership) the baseline.

Use shortcuts to add a new column **K**: Select **K**, **Alt HIC** and make the interaction *after Deng x ln Chinese car production (K) t-1*:

In **K2**, enter **=I2\*G2 [Enter]**, then double click the lower right corner of the new cell to fill in the column:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	year	Chinese car sales (K)	Chinese car production (K) t-1	sqrt Chinese car sales (K)	ln Chinese car sales	sqrt Chinese car production (K) t-1	ln Chinese car production (K) t-1	Tiananman Square	after Deng	Fourth Generation	after Deng x ln Chinese car production (K) t-1			
8	1996	478	321	22	6.2	18	5.8	0	0	0	0.0			
9	1997	589	382	24	6.4	20	5.9	0	0	0	0.0			
10	1998	600	488	24	6.4	22	6.2	0	1	0	6.2			
11	1999	607	507	25	6.4	23	6.2	0	1	0	6.2			

Rearrange columns to make the predictor columns adjacent for regression, with the continuous predictor first, followed by the two continuous predictors and the indicator interaction.

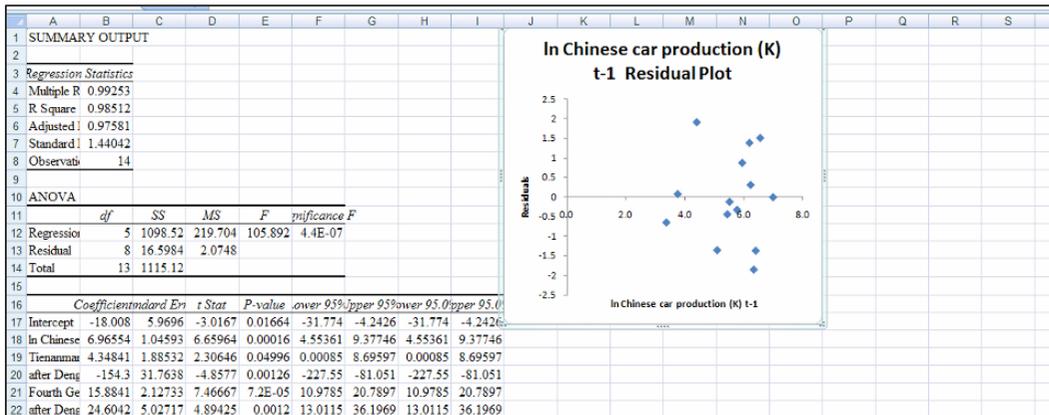
(By ordering the *ln Chinese car production (K) t-1* first, we will get the residual plot to assess heteroskedasticity.)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	year	Chinese car sales (K)	Chinese car production (K) t-1	ln Chinese car sales	sqrt Chinese car production (K) t-1	sqrt Chinese car sales (K)	ln Chinese car production (K) t-1	Tiananman Square	after Deng	Fourth Generation	after Deng x ln Chinese car production (K) t-1		
2	1990	83	29	4.4	5	9	3.4	1	0	0	0.0		
3	1991	157	42	5.1	7	13	3.7	1	0	0	0.0		
4	1992	356	81	5.9	9	19	4.4	1	0	0	0.0		

Now we have columns ready for the model regression, with

- the dependent variable, the square roots of car sales, **F**, followed by
- natural logarithms of past year car production in **G**,
- the three indicators in **H** through **J**,
- the indicator interaction with natural logarithms of past year car production in **K**.

Run the regression, excluding the two most recent years in rows **16** and **17** to later validate the model.



The model  $F$  is significant, allowing us to conclude that the shock from the Tiananmen Square incident, Chinese leadership, and growth in Chinese car production together drive car sales in China.

All coefficient estimates are significant ( $p$  values  $< .05$ ), and the sign of growth (ln) in past year *car production* is positive, as expected.

**Assess autocorrelation.** Since this is a time series model, we need to assess residual autocorrelation to see whether or not the Leading Indicator, *past year Chinese car production*, has successfully accounted for trend and cycles in *Chinese car sales*. Next to the residual column in the regression output sheet, add the Durbin Watson statistic to check for unaccounted for trend or cycles.

Durbin Watson statistic 2.04 exceeds two, allowing us to conclude that the residuals are free of autocorrelation.

**Residual assessment.** The **ln Chinese car production (K) t-1 Residual Plot** is cloud-like and no heteroskedasticity or pattern is apparent.

**Validate the model.** With the model coefficient estimates from **B18:B22** of the regression worksheet in **L1:L7**, make *predicted sqrt Chinese car sales (K)* in **O**.

With

- standard error from **B7** of the regression worksheet in **M2**, and
- $t$  for 8 error degrees of freedom in **N2**,

make the 95% lower and upper sqrt Chinese car sales (K) in **P** and **Q** to validate the model:

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	sqrt Chinese car sales (K)	Tiananman Square	after Deng	Fourth Generation	In Chinese car production (K) t-1	after Deng x In Chinese car production (K) t-1	Coefficients	se	t	predicted sqrt Chinese car sales (K)	95% lower sqrt Chinese car sales (K)	95% upper sqrt Chinese car sales (K)	95% lower Chinese car sales (K)	95% upper Chinese car sales (K)	
14	36	0	1	0	6.6	6.6				34.66772	31.3461	37.989	982.58	1443.2	
15	47	0	0	1	7.0	0.0				46.59734	43.2757	49.919	1872.8	2491.9	
16	50	0	0	1	7.6	0.0				50.8854	47.5638	54.207	2262.3	2938.4	
17	54	0	0	1	7.7	0.0				51.84256	49.521	55.164	2354.3	3043.1	
18		0	0	1	8.1	0.0				54.22319	50.9016	57.545	2591	3311.4	
19		0	0	1	8.3	0.0				55.95471	52.6331	59.276	2770.2	3513.7	
20		0	0	1	8.6	0.0				57.50236	54.1807	60.824	2935.6	3699.6	

The model correctly forecasts held-out cars sales in 2004 and 2005:

- In 2004, actual square root of car sales is 50 (K), which falls within the 95% prediction intervals of 47.6 to 54.2 (K).
- In 2005, actual square root of car sales is 54 (K), which falls within the 95% prediction interval 48.5 to 55.2 (K).

Recalibrate, including data from 2004 and 2005:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SUMMARY OUTPUT																		
2																			
3	Regression Statistics																		
4	Multiple R 0.99561																		
5	R Square 0.99124																		
6	Adjusted R Square 0.98687																		
7	Standard Error 1.48097																		
8	Observations 16																		
9																			
10	ANOVA																		
11		df	SS	MS	F	Significance F													
12	Regression	5	2483	496.601	226.419	6E-10													
13	Residual	10	21.9328	2.19328															
14	Total	15	2504.94																
15																			
16		Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%										
17	Intercept	-19.187	5.68229	-3.3766	0.00704	-31.848	-6.5257	-31.848	-6.5257										
18	ln Chinese	7.17347	0.99437	7.21411	2.9E-05	4.95789	9.38906	4.95789	9.38906										
19	Tiananman	4.66382	1.83619	2.53994	0.02937	0.57253	8.75511	0.57253	8.75511										
20	after Deng	-153.12	32.5757	-4.7004	0.00084	-225.7	-80.537	-225.7	-80.537										
21	Fourth Ge	15.8616	2.10484	7.53578	2E-05	11.1718	20.5515	11.1718	20.5515										
22	after Deng	24.3963	5.15247	4.73487	0.0008	12.9159	35.8767	12.9159	35.8767										

Together, the *Tiananmen Square* shock, Chinese leadership, and growth in past year *car production* account for 99% of the variation in *car sales in China*. The model *F* is significant: one or more of the predictors is driving *car sales*. All *p values* are significant: leadership, *production*, and their interaction drive *car sales*.

**Update forecasts.** Copy the recalibrated coefficients **B17:B22** and paste over the validation coefficients in the original worksheet to update forecasts. Change the standard error *se* to the recalibrated value 1.48. Update *t* by changing the error degrees of freedom to 10.

**Rescale to thousands of cars.** The forecasts are in square roots. To rescale back to thousands of cars, make two new columns, *95% lower* and *upper Chinese car sales (K)* in **R** and **S** by squaring the predicted square roots:

In **R2**, enter  $=P2^2$ .

In **S2**, enter  $=Q2^2$ .

Select the two new cells **R2:S2** and double click the lower right corner to fill in through row **23**:

1	year	Chinese car sales (K)	Chinese car production (K) t-1	ln Chinese car sales	sqrt Chinese car production (K) t-1	sqrt Chinese car sales (K)	ln Chinese car production (K) t-1	Tienanman Square	after Deng	Fourth Generation	after Deng x ln Chinese car production (K) t-1	coefficients	se	t	predicted sqrt Chinese car sales (K)	95% lower sqrt Chinese car sales (K)	95% upper sqrt Chinese car sales (K)	95% lower Chinese car sales (K)	95% upper Chinese car sales (K)	me
16	2003	2171	1091	7.7	33	47	7.0	0	0	1	0.0			46.9	43.6	50.2	1897	2515		
16	2004	2479	2019	7.8	45	50	7.6	0	0	1	0.0			51.3	48.0	54.6	2301	2978		
17	2005	2914	2316	8.0	48	54	7.7	0	0	1	0.0			52.3	49.0	55.6	2396	3086		
18	2006		3260		57		8.1	0	0	1	0.0			54.7	51.4	58.0	2642	3365		
19	2007		4180		65		8.3	0	0	1	0.0			56.5	53.2	59.8	2829	3575		
20	2008		5220		72		8.6	0	0	1	0.0			58.1	54.8	61.4	3001	3768		
21	2009		7600		87		8.9	0	0	1	0.0			60.8	57.5	64.1	3304	4106		
22	2010		9000		95		9.1	0	0	1	0.0			62.0	58.7	65.3	3444	4263		
23	2011		10500		102		9.3	0	0	1	0.0			63.1	59.8	66.4	3575	4408		
24																				

In 2011, we expect *Chinese car sales* of 3,580 to 4,410 (K) cars, or 3.58 to 4.41 million cars.

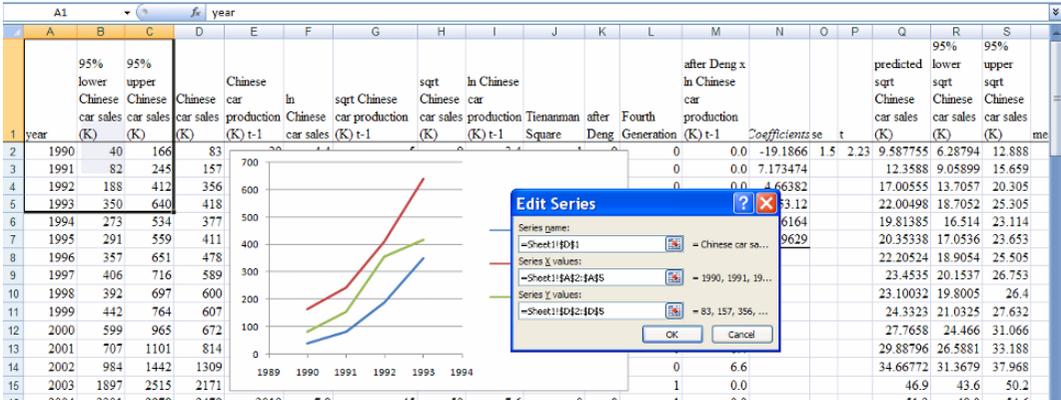
The model margin of error is half the 95% prediction interval,  $420(K) = 4410(K) - 3580(K)$  cars. We expect our forecast for 2011 to be no further than 420,000 cars from actual *Chinese car sales*.

**Illustrate the fit and forecast.** To see the fit and forecasts, make a scatterplot of actual car sales in **B** and 95% prediction intervals scaled back to the original units in **R** and **S**.

To plot the fit and forecasts with actual sales, move the prediction intervals to columns **B** and **C**.

Plot each of the distinct periods as a separate set of three series for sales, lower and upper prediction interval bounds.

First select the *Tiananmen Square* rows **A1:C5**, **Alt ND**.  
 Select the chart and right click, then **Format Data Series** to **Add Chinese car sales (K)** in **D**.



For years under *Deng's leadership*,  
**Add** the three series in rows **6** through **9**:  
*95% lower and upper Chinese car sales (K)* in **B** and **C**, and *Chinese car sales (K)* in **D**.

For the period *after Deng's death*,  
**Add** three series in rows **10** through **14**:  
*95% lower and upper Chinese car sales (K)* in **B** and **C**, and *Chinese car sales (K)* in **D**.

For the period of *Fourth Generation rule*,  
**Add** two series in rows **15** though **23**:  
*95% lower and upper Chinese car sales (K)* in **B** and **C**.  
**Add** *Chinese car sales (K)* in **D**, for years 2003 through 2005 in rows **15** through **17**.

Customize background, markers, font, and scales:



The increasing sales response to growing *car production* during the period following Deng’s death is apparent, as is the slowing of *car sales* growth in recent years under *Fourth Generation* leadership.

**Sensitivity analysis.** To estimate the impact of *Fourth Generation leadership*, relative to *Third Generation leadership after Deng’s death*, make prediction intervals for years 2003 through 2011 under the alternate scenario of continuing *Third Generation leadership*.

First, save predictions based on the actual change in leadership in 2003, removing formula references, for later comparison:

Use shortcuts to add two new columns, **B** and **C**, by selecting **B** and **C**, **Alt HIC**.

Use shortcuts to copy and paste actual predictions based on the leadership change in 2003:

Select filled cells in **D** and **E**, copy, **Cntl+C**, then paste into **B** and **C** *without formula references* by selecting **B1**, **Alt HVSU**, **ok**.

(The duplicate columns now in **B** and **C** have been copied without formula references and their values will not change when you change the indicators to reflect the alternative scenario.)

Make predicted Chinese car sales by squaring predicted square roots in **V** and save a copy without formula references in **W**:

Use shortcuts to add two empty columns **V** and **W**: Select **V:W**, **Alt HIC**.

Make *predicted Chinese car sales (K)* in **V**: in **V2**, enter the formula =S2^2 [Enter], then double click the lower right corner to fill in the column.

Save a copy of the predicted values that is free of formula references by select filled cells in **V**, **Cntl+C**, then selecting **W1**, **Alt HVSU**, **ok**:

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
	Chinese car sales (K)	Chinese car production (K) t-1	ln Chinese car sales	sqrt Chinese car production (K) t-1	sqrt Chinese car sales (K) t-1	ln Chinese car production Square	Tienanman	after Deng	Fourth Generation	after Deng x ln Chinese car production (K) t-1	Coefficients se	t	predicted Chinese car sales (K)	95% lower sqrt Chinese car sales (K)	95% upper sqrt Chinese car sales (K)	predicted Chinese car sales (K)	predicted Chinese car sales (K)	mean	
2	83	29	4.4		5	9	3.4	1	0	0	0.0	-19.1866	1.5	2.23	9.587755	6.28794	12.888	91.92505	91.925046
3	157	42	5.1		7	13	3.7	1	0	0	0.0	7.173474			12.3588	9.05899	15.659	152.74	152.74003
4	356	81	5.9		9	19	4.4	1	0	0	0.0	4.66382			17.00555	13.7057	20.305	289.1886	289.18858

Set up the hypothetical scenario of continuing Third Generation leadership *after Deng*:

In rows **15:17**, change the zeros to ones in column **L**, then change the ones to zeros in column **M**:

Read the hypothetical predictions from **V15:17** and **D15:E17**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	year	95% lower Chinese car sales (K)	95% upper Chinese car sales (K)	95% lower Chinese car sales (K)	95% upper Chinese car sales (K)	Chinese car sales (K)	ln Chinese car sales (K) t-1	sqrt Chinese car sales (K) t-1	ln Chinese car production (K) t-1	sqrt Chinese car production (K) t-1	ln Chinese car production Square	Tienanman	after Deng	Fourth Generation	after Deng x ln Chinese car production (K) t-1	Coefficients se	t	predicted Chinese car sales (K)	95% lower sqrt Chinese car sales (K)	95% upper sqrt Chinese car sales (K)	predicted Chinese car sales (K)	predicted Chinese car sales (K)	
14	2002	984	1442	984	1442	1309	704	7.2	27	36	6.6	0	1	0	6.6			34.668	31.368	37.968	1201.9	1201.9	
15	2003	1897	2515	2044	2685	2171	1091	7.7	33	47	7.0	0	1	0	7.0			48.5	45.2	51.8	2354	2195	
16	2004	2301	2978	4179	5076	2479	2019	7.8	45	50	7.6	0	1	0	7.6			67.9	64.6	71.2	4617	2628	
17	2005	2396	3086	4759	5713	2914	2316	8.0	48	54	7.7	0	1	0	7.7			72.3	69.0	75.6	5225	2730	
18	2006	2642	3365	6364	7461		3260		57		8.1	0	1	0	8.1			83.1	79.8	86.4	6902	2993	
19	2007	2829	3575	7678	8878		4180		65		8.3	0	1	0	8.3			90.9	87.6	94.2	8267	3191	
20	2008	3001	3768	8956	10249		5220		72		8.6	0	1	0	8.6			97.9	94.6	101.2	9592	3373	
21	2009	3304	4106	11342	12791		7600		87		8.9	0	1	0	8.9			109.8	106.5	113.1	12056	3694	
22	2010	3444	4263	12507	14027		9000		95		9.1	0	1	0	9.1			115.1	111.8	118.4	13256	3843	
23	2011	3575	4408	13619	15203		10500		102		9.3	0	1	0	9.3			120.0	116.7	123.3	14400	3981	

**Difference between alternative scenarios.** Find the estimated annual differences between the hypothetical leadership scenario and actual.

In **X15** through **X23**, find the difference in predictions *Third Generation instead* by comparing predictions under *Fourth generation* in **W** with predictions under the alternate scenario of *Third Generation* in **V**:

In **X15** enter =**V15-W15** [Enter], select the new cell, grab and drag through row **23**.

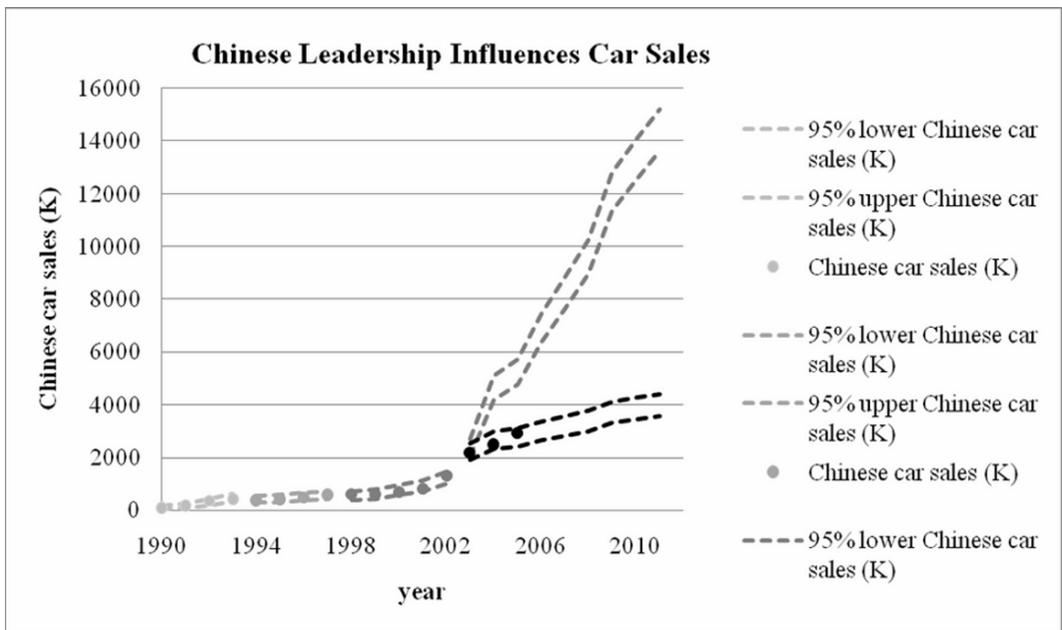
Find the % change by comparing the differences in X to forecasts under *Fourth Generation* in W:

In Y15, enter =100\*X15/W15 [Enter], select the new cell, grab and drag through row 23.

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
15	2685	2171	1091	7.7	33	47	7.0	0	1	0	7.0				48.5	45.2	51.8	2354	2195	159	7		
16	5076	2479	2019	7.8	45	50	7.6	0	1	0	7.6				67.9	64.6	71.2	4617	2628	1989	76		
17	5713	2914	2316	8.0	48	54	7.7	0	1	0	7.7				72.3	69.0	75.6	5225	2730	2495	91		
18	7461		3260		57		8.1	0	1	0	8.1				83.1	79.8	86.4	6902	2993	3909	131		
19	8878		4180		65		8.3	0	1	0	8.3				90.9	87.6	94.2	8267	3191	5076	159		
20	10249		5220		72		8.6	0	1	0	8.6				97.9	94.6	101.2	9592	3373	6218	184		
21	12791		7600		87		8.9	0	1	0	8.9				109.8	106.5	113.1	12056	3694	8362	226		
22	14027		9000		95		9.1	0	1	0	9.1				115.1	111.8	118.4	13256	3843	9413	245		
23	15203		10500		102		9.3	0	1	0	9.3				120.0	116.7	123.3	14400	3981	10419	262		

Had Third Generation leadership *after Deng* continued, sales would have been higher, by about 159 (K) cars (7%) in 2004: 2,354 (K), instead of 2,195 (K).

**Illustrate the alternative scenarios.** To see the difference that *Fourth Generation Leadership* has made, add the hypothetical prediction intervals in B and C to your forecast plot.



Had Third Generation leadership continued, growth in *car sales* would have been much greater.

## Lab Practice 12

### Car Sales in India

An American car manufacturer is considering a joint venture in India where cars would be manufactured for sale to the growing Indian population and Asian markets. Management believes that in India, the Leading Indicator, population growth, will drive car sales in the next five years.

It is also believed that structural shifts from changes in leadership affect both the demand for cars and also the proportion of cars produced which are exported, rather than sold in India. A noticeable structural shift occurred in 1991, following the death of Gandhi. The Congress Party controlled leadership after Gandhi's death until the Gandhi's BJP party again gained control in 1997. Congress took back leadership in 2004.

Follow the steps in **Excel 12.1** to build a time series model of *car sales in India*, with **Lab 12 India Car Sales.xls** including:

- an indicator of *Congress* leadership to represent the major shifts in economic policy, equal to 1 in years 1991-1996 and 2004-present
- one or more interactions between this indicator and the continuous variables in the model,
  - *past year Indian car production* and
  - *Indian population*.

**Assess skewness.** Which variable is positively skewed? \_\_\_\_\_

**Choose scales.** Which scale, square roots or natural logarithms, better *Normalizes* the positively skewed variable?

\_\_\_\_\_

**Assess autocorrelation.** Is your model is free of autocorrelation? \_\_\_\_\_  
(Assess autocorrelation. If  $DW$  is greater than  $dL$ , you do not need to add variables.)

**Validate** your model, then **recalibrate**.

**Write your model equations in the original scale** of thousands of *cars sold in India*

- For the **baseline** BJP leadership
- For leadership under *Congress*

**Forecast.** What are *Indian car sales* expected to be in 2010, with 95% confidence?

\_\_\_\_\_

**Illustrate your fit and forecast.** Make a scatterplot of *95% lower and upper predicted sales through 2010* with *actual sales through 2004* to illustrate your model fit and forecast. Plot the distinct leadership periods as separate series:

- *Leadership under BJP Party 1983 through 1990*
- *Leadership under Congress 1991 through 1997*
- *Leadership under BJP Party 1998 through 2003*
- *Leadership under Congress in 2004*
- *Continuing leadership under Congress 2005 through 2010*

**Sensitivity analysis.** Make a table to compare *Indian car sales in 2008 through 2010* under the **alternative scenario of BJP leadership from 2008**, including the percent increase or decline under BJP leadership, relative to *Congress leadership*.

Add to your scatterplot *95% lower and upper predicted sales through 2010 given BJP leadership in 2008 through 2010*.

Attach a printout of your scatterplot to your lab practice worksheet.

## *CASE 12-1 Explain and Forecast Defense Spending for Rolls-Royce*

Sales to defense contractors are critical to Rolls-Royce growth and profitability. Executives know from experience that the defense business depends critically upon government defense spending, which is influenced by political leadership, global conflict, and the Nation's productivity. Ralph Roy, Senior Assistant to the Director of Corporate Planning, has built a model of defense spending, which he must soon present to executives. He has asked you to review his model and suggest improvements.

**Indicators and drivers of defense spending.** Ralph began by interviewing executives to identify defense spending drivers. From these conversations, the list of likely influences included:

- Party leadership in the White House, *Republican White House*
- The impact of terrorism on *911*
- The Leading Indicator, past year productivity, measured by *GDP*
- The Leading Indicator, number of quarters the Nation had been engaged in military conflict in the past quarters under the current administration, *past year conflict*, since involvement in military conflict during an administration probably affected defense spending.

Past defense spending and spending habits tended to continue. Ralph included an inertia component:

- *Past year defense spending*

**Scales to reduce skewness.** *Defense spending* and *quarters ongoing conflict* were positively skewed. Ralph used natural logarithms of *defense spending* and square roots of *quarters ongoing conflict*.

Ralph included the two indicators, the two Leading Indicators, and inertia in his initial model.

Ralph was pleased that his model accounted for a high proportion of the variation in defense spending across quarters (98%), that his model was significant, and that the two indicators and three drivers were significant and had "correct" positive signs. (*Sqrt quarters ongoing conflict* was significant at a 94% level of confidence with a two tail test, but Ralph felt comfortable using a one tail test since he was convinced this influence would be positive. The one tail test *p value*, which is half the two tail *p value*, is .03, making *sqrt quarters ongoing conflict* significant at a 95% level of confidence.)

The model correctly forecast spending levels in the two most recent quarters which had been hidden to fit and validate the model.

Ralph's regression results are in the workbook **Case 12-1 defense spending.xls**.

Ralph is somewhat concerned that he may have left out one or more important variables or interactions, since the plot of his residuals (on the **residuals** worksheet) shows several patterns.

- Party control of the Senate, *Republican Senate* may influence spending and may interact with
  - *Sqrt quarters ongoing conflict*, since how aggressively Congress decided to spend on continuing conflict probably differed across the two Parties.
  - *Past year defense spending*
- There appears to be a shift in spending during Presidents' *second terms*.
- *President's tenure*, number of quarters in office, may be related to defense spending
  - And may interact with Party in control of the Senate, *Republican Senate*.

Is Ralph's model is complete? Or should additional variables be added? Document your answer with the appropriate test.

Improve Ralph's model by adding unaccounted for influences, including

- *Republican Senate leadership indicator*, and
  - Its interaction with *sqrt quarters of ongoing conflict*,
  - Its interaction with the natural *logarithm of past year defense spending*,
- a *Second Term* indicator,
- *Presidential Tenure*, and
  - Its interaction with the *Republican Senate* leadership indicator

Explain how you know whether or not you have improved Ralph's model and state your evidence.

Write the equations for your improved model in trillions of dollars for spending under four scenarios. Please use proper subscripts, superscripts, and indentations:

- i. The **first** term of a *Republican President* with a **Democratic** *Senate* in quarters **after the impact of 911 has subsided**,
- ii. The **first** term of a *Republican President* with a *Republican Senate* in quarters **after the impact of 911 has subsided**,
- iii. The **first** term of a *Democratic President* with a *Democratic Senate* in quarters **after the impact of 911 has subsided**,
- iv. The **first** term of a *Democratic President* with a *Republican Senate* in quarters **after the impact of 911 has subsided**.

Attach or embed a scatterplot of the *95% prediction intervals* and *actual defense spending* in hundred billion dollars (T\$) **through the second quarter of 2007**.

What quarterly growth in *defense spending* does your model forecast for the **second and third quarters of 2007**?

<i>Quarter</i>	<i>Forecast Defense Spending (\$T)</i>	<i>% of Forecast from previous quarter</i>
<i>I Jan 2007</i>		
<i>II Apr 2007</i>		
<i>III Jul 2007</i>		

How much lower does your model predict *defense spending* to be in the **second and third quarters of 2007** because there is a **Democratic Senate** instead of a **Republican Senate**?

<i>Quarter</i>	<i>Forecast Defense Spending (\$T)</i>		<i>% decrease relative to a Republican Senate</i>
	<i>Under Democratic Senate</i>	<i>Under Republican Senate</i>	
<i>II Apr 2007</i>			
<i>III Jul 2007</i>			

Rolls Royce revenues tend to increase with *defense spending*.

Will it be more important to contribute to the campaigns of candidates for Senate or a candidate for President—which makes a bigger difference on *defense spending*, a *Republican President* or a *Republican Senate*?

Explain how you used the model to provide evidence for your answer.

*CASE 12-2 Haier's U.S. Refrigerator Strategy\**

Use the data in **Case 12-2 Haier.xls** for analyses and preparation for class discussion.

---

\* Harvard Business School case 9705475

## 13

# Logit Regression for Bounded Responses

In this chapter we introduce *logit* regression which accommodates responses which are *limited*, or bounded above and below. For example, the likelihood of trying a new product can neither be negative nor greater than one hundred percent. Market share is similarly limited to the range between zero and one hundred percent. Indicator 0-1 responses, such as “tried the product or not” and “voted Republican” reflect probabilities, such as the probability of trying a new product, the probability of winning a game, or the probability of voting Republican. In each of these cases, we need to rescale dependent response, acknowledging these boundaries. The odds ratio rescales probabilities or shares to a corresponding unbounded measure. The *logit*, or natural logarithm of an odds ratio, rescales responses, producing an S-shaped pattern, which reflects greater response among “fence sitters” with probabilities or shares that are mid-range.

### 13.1 Rescaling Probabilities or Shares to Odds Improves Model Validity

With each response probability,  $\pi$ , there is an *odds ratio*, the chance that the response occurs relative to the chance that it does not occur.

$$odds = \pi / (1 - \pi)$$

Response shares, such as market share, also have odds ratios, which reflect percent of the market owned, relative to the percent of the market owned by competitors:

$$odds = MarketShare / (100 - MarketShare)$$

While probabilities and shares are bounded by zero, below, and one or one hundred percent, above, the corresponding odds ratio and its natural logarithm, the *logit*, are not bounded:

$$Logit = \ln(odds)$$

Rescaling to logits produces an S-shaped curve, which, for a probability at .5, or a share at 50%, has a logit of zero. Figure 13.1 illustrates this S-shaped scale.

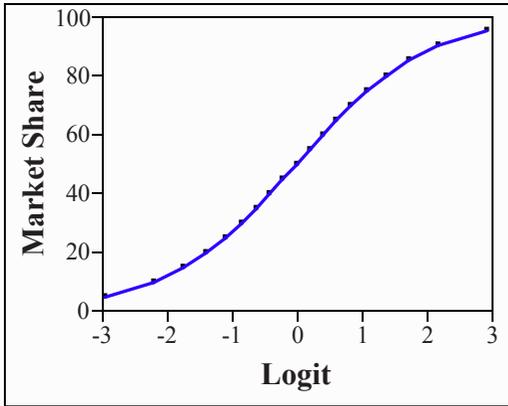


Figure 13.1 Logits of bounded shares are unbounded

*Example 13.1 The Import Challenge*<sup>1</sup>. Ford Motors executives were pondering the U.S. car market, where increasingly consumers were choosing imports. In response to Toyota’s successful launch of the hybrid Prius model, Ford had designed and begun selling hybrid Focus. American cars were known to be less fuel efficient and less reliable than imports, but also less expensive than similar cars designed abroad. What car characteristics drove U.S. car owner satisfaction? Was value enough to sustain share in the U.S. market? Ford executives asked Amanda Arnone, the Director of Quantitative Analysis to build a model of car owner satisfaction to provide answers.

Consumer Reports (consumerreports.com) routinely collects data on car owners’ satisfaction by asking the question, “Would you buy this model again?” Each model’s satisfaction rating is the percent of owners who answered “yes.” Amanda used satisfaction percents for 37 car models to build the model.

She included:

- An indicator of whether or not a car is a *hybrid*,
- An indicator of whether or not a car is an *import*
- fuel economy, *MPG*,
- an indicator interaction between *hybrid* and *MPG*,
- lack of power, *seconds* to accelerate from 0 to 60 MPH,
- *price* (K\$), to represent overall quality and luxury,
- An indicator interaction between *import* and *price*

Since the percent of owners of a car who are satisfied, *satisfaction*, is bounded below by zero and above by one hundred, Amanda used the *satisfaction logit* as the dependent variable:

<sup>1</sup> This example is a hypothetical scenario using actual data.

$$\begin{aligned} \widehat{satisfactionLogit}_i &= \ln \left[ \frac{satisfaction_i}{100 - satisfaction_i} \right] \\ &= b_0 + b_1 hybrid_i + b_2 import_i + b_3 seconds_i + b_4 MPG_i \\ &\quad + b_5 (hybrid_i x MPG_i) + b_6 price_i + b_7 (import_i x price_i) \end{aligned}$$

Regression results from the model are shown below:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.849					
R Square	0.720					
Adjusted R Square	0.657					
Standard Error	0.395					
Observations	39					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	7	12.4	1.8	11.4	0.0000	
Residual	31	4.8	0.2			
Total	38	17.2				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.33	0.82	0.4	0.6869	-1.33	2.00
<i>hybrid</i>	-2.98	1.02	-2.9	0.0065	-5.06	-0.90
<i>import</i>	1.63	0.49	3.3	0.0023	0.63	2.63
<i>seconds to go 0 to 60</i>	-0.20	0.050	-4.0	0.0004	-0.30	-0.10
<i>mpg</i>	0.029	0.017	1.7	0.0948	-0.005	0.064
<i>hybrid x mpg</i>	0.090	0.031	2.9	0.0061	0.028	0.153
<i>price (\$K)</i>	0.044	0.014	3.1	0.0042	0.015	0.073
<i>import x price</i>	-0.031	0.015	-2.0	0.0501	-0.061	0.000

Table 13.1 Regression of Satisfaction Logit by Car Characteristic

The significant and positive coefficient for the *import* indicator suggests that more import owners than domestic owners are satisfied. Quality, greater reliability and luxury featured associated with a higher price are less important to import owners.

The significant and negative coefficient for the *hybrid* indicator suggests that owners of conventional cars are more likely to be satisfied than owners of hybrids. Hybrid owners are more satisfied if fuel economy is higher.

A greater proportion of owners of all cars are satisfied if a model offers more responsive acceleration. The relative importance of each of three car characteristics is marginal and depends on a car's configuration of all three, as well as whether the car has a conventional engine or a hybrid engine, and whether or not the car is an imported or domestic model.

**Rescale equations back to satisfaction proportions.** The model equation for conventional domestic cars, setting the *hybrid* and *import* indicators to 0, is:

$$\log \hat{it}_i = .33 - .20 \text{seconds}_i + .029 \text{MPG}_i + .044 \text{price}(\$K)_i$$

The model for conventional imports, setting the *hybrid* indicator to 0 and the *import* indicator to 1, is:

$$\log \hat{it}_i = 1.96 - .20 \text{seconds}_i + .029 \text{MPG}_i + .013 \text{price}(\$K)_i$$

The model for domestic hybrids, with the *import* indicator set to zero and the *hybrid* indicator set to one, is:

$$\log \hat{it}_i = -2.65 - .20 \text{seconds}_i + .099 \text{MPG}_i + .044 \text{price}(\$K)_i$$

The model for *hybrid imports*, with both indicators set to one, is:

$$\log \hat{it}_i = 1.02 - .20 \text{seconds}_i + .099 \text{MPG}_i + .013 \text{price}(\$K)_i.$$

To see the equations in the original scale of *satisfaction proportion*, first find the *predicted satisfaction odds*, which is the exponential function of the *predicted logits*:

$$\hat{odds}_i = e^{(.33 - .20 \text{seconds}_i + .029 \text{MPG}_i + .044 \text{price}(\$K)_i)} \quad \text{for domestic conventional models,}$$

$$\hat{odds}_i = e^{(1.96 - .20 \text{seconds}_i + .029 \text{MPG}_i + .013 \text{price}(\$K)_i)} \quad \text{for imports with conventional engines,}$$

$$\hat{odds}_i = e^{(-2.65 - .20 \text{seconds}_i + .099 \text{MPG}_i + .044 \text{price}(\$K)_i)} \quad \text{for domestic hybrids,}$$

and

$$\hat{odds}_i = e^{(1.02 - .20 \text{seconds}_i + .099 \text{MPG}_i + .013 \text{price}(\$K)_i)} \quad \text{for imports with hybrid engines.}$$

Predicted proportions satisfied are then, for domestic conventional models,

$$\widehat{satisfaction}_i = 100 \frac{e^{(.33-.20 \text{ sec onds}i+.029 \text{ MPG}i+.044 \text{ price}(\$K)i)}}{1 + e^{(.33-.20 \text{ sec onds}i+.029 \text{ MPG}i+.044 \text{ price}(\$K)i)}$$

for owners of *imports* with conventional engines:

$$\widehat{satisfaction}_i = 100 \frac{e^{(1.96-.20 \text{ sec onds}i+.029 \text{ MPG}i+.013 \text{ price}(\$K)i)}}{1 + e^{(1.96-.20 \text{ sec onds}i+.029 \text{ MPG}i+.013 \text{ price}(\$K)i)}$$

for owners of domestic *hybrids*:

$$\widehat{satisfaction}_i = 100 \frac{e^{(-2.65-.20 \text{ sec onds}i+.099 \text{ MPG}i+.044 \text{ price}(\$K)i)}}{1 + e^{(-2.65-.20 \text{ sec onds}i+.099 \text{ MPG}i+.044 \text{ price}(\$K)i)}$$

and, for owners of *imports* with *hybrid* engines:

$$\widehat{satisfaction}_i = 100 \frac{e^{(1.02-.20 \text{ sec onds}i+.099 \text{ MPG}i+.013 \text{ price}(\$K)i)}}{1 + e^{(1.02-.20 \text{ sec onds}i+.099 \text{ MPG}i+.013 \text{ price}(\$K)i)}$$

Because the dependent variable has been rescaled, the logit model has built in synergies. The value of an improvement in one of the characteristics will be nonconstant, and also dependent on the levels of other characteristics. To illustrate the synergies, we will compare expected satisfaction in response to differences in one of the car characteristics, setting the remaining two at best and worst levels.

To see the difference in expected proportion of domestic owners satisfied that *price* could make, we will compare alternate *prices* for four hypothetical cars:

- least attractive (maximum *seconds* to accelerate 0 to 60 and lowest *MPG*) conventional domestic
- most attractive (minimum *seconds* to accelerate 0 to 60 and best *MPG*) conventional domestic
- least attractive domestic *hybrid*
- most attractive domestic *hybrid*

**Price/Quality/Luxury.** Increasing the *price* of car models, which implies increasing their quality, reliability or luxury, has the greatest potential impact among domestic hybrid owners, shown with solid lines in Figure 13.2. However, it does not make enough difference to compensate for lack of acceleration and poor fuel economy. Most owners of responsive, fuel efficient cars, whether domestic or imported, are satisfied, and adding quality and a more expensive price tag does not improve the already high proportion satisfied.

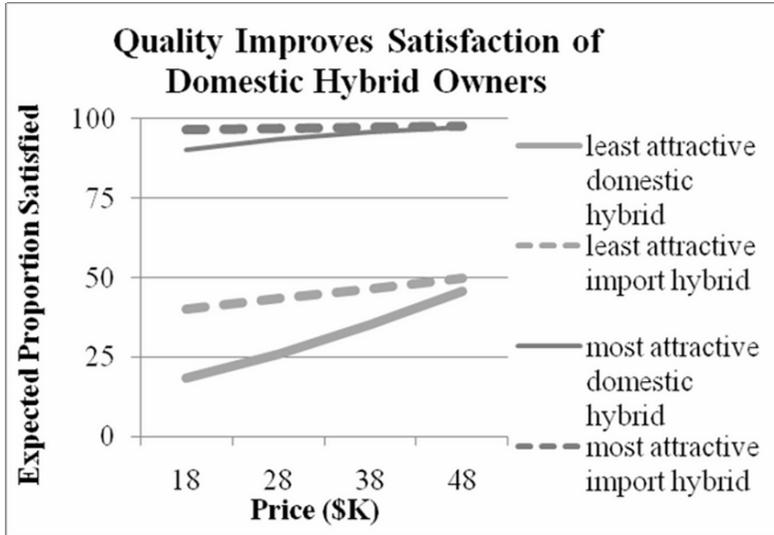


Figure 13.2 Proportion satisfied by price

**Acceleration.** Improved acceleration makes a larger difference to owners of the least desirable, least fuel efficient economy models, shown with lighter lines in Figure 13.3, whether domestic or imports. For Ford, improved response would help to satisfy more, but not enough to satisfy the majority of owners of inexpensive, less fuel efficient hybrids.

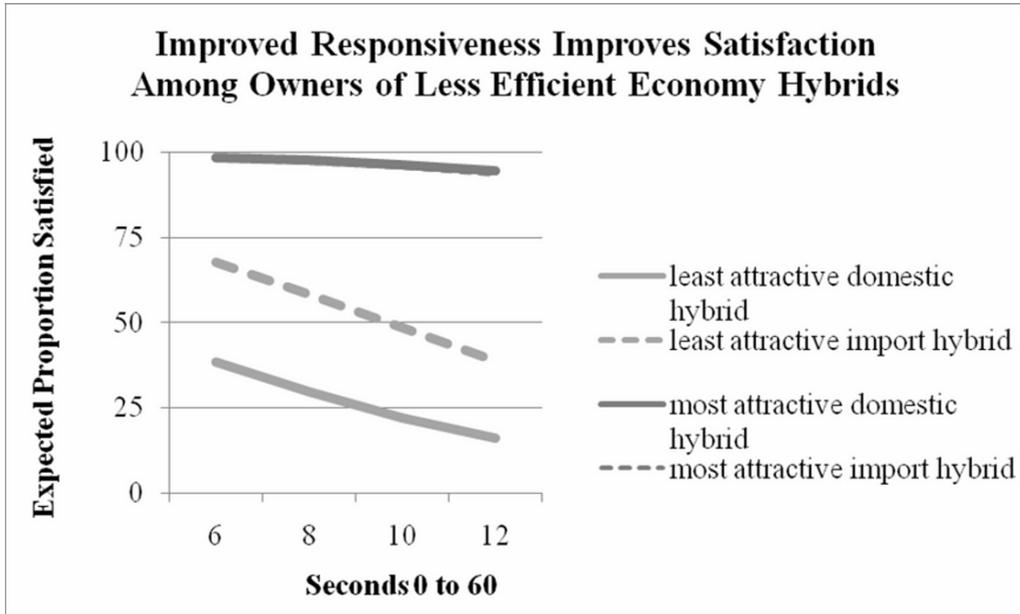
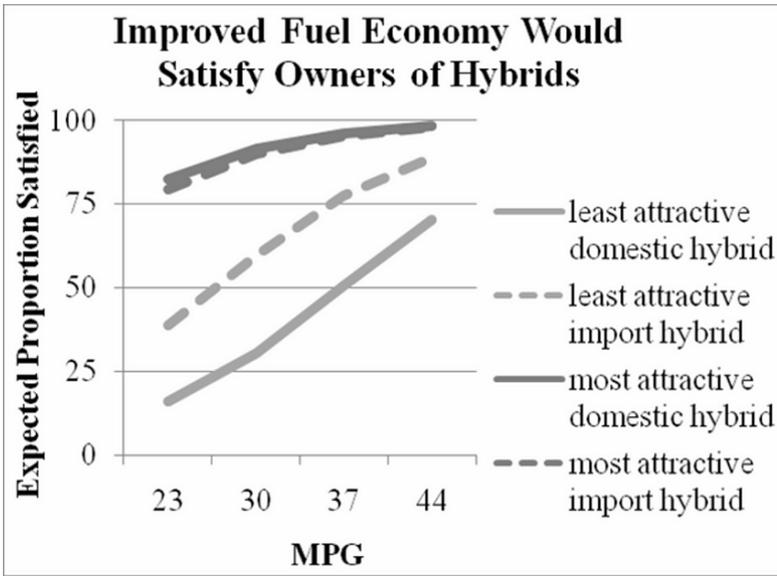


Figure 13.3 Proportions of satisfied car owners by responsiveness

**Fuel Economy.** To see the expected impact of fuel economy improvements, we compare hypothetical domestic and import hybrids with best and worst combinations of price/quality and acceleration. These are shown in Figure 13.4.



Fuel economy matters more for owners of less responsive, inexpensive models, since it compensates. Adding fuel efficiency would compensate owners of both domestic and imported hybrids and will be a key to Ford's success in hybrids.

Figure 13.4 Proportions of Satisfied Car Owners by MPG

When all but one of the characteristics are desirable, they compensate for lacking along that one characteristic. Owners of expensive, responsive luxury cars remain relatively satisfied, even with poor fuel economy. Owners of inexpensive, but responsive cars with superior fuel economy are relatively satisfied without additional luxuries. Owners of expensive, fuel efficient luxury cars are satisfied without responsiveness. However, lacking strength in any of the three important dimensions, fuel efficiency will satisfy the majority of hybrid owners.

Amanda summarized her model results for Ford executives:

**MEMO**

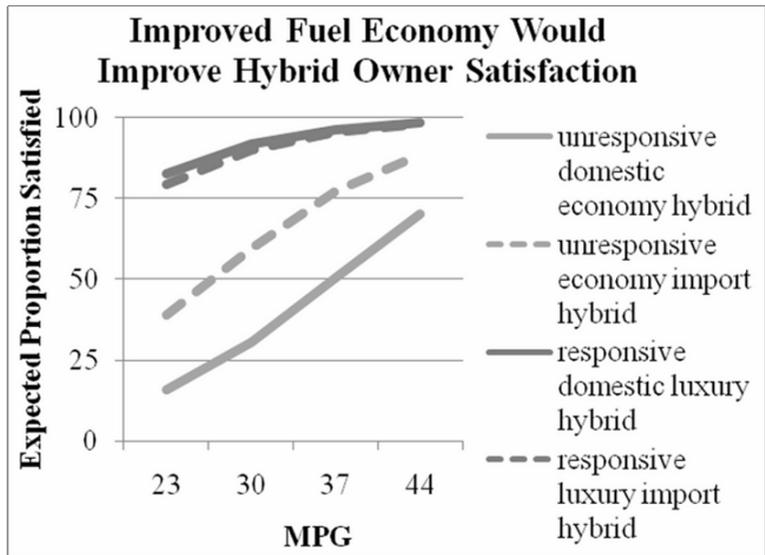
**Re: Fuel Efficiency Drives Hybrid Owner Satisfaction**  
**To: Ford Strategic Development Executives**  
**From: Amanda Arnone, Quantitative Analysis Director**  
**Date: June 2007**

A greater proportion of domestic hybrid owners would be satisfied with more fuel efficient cars. Quality and responsiveness are also important drivers of satisfaction.

A model of owner satisfaction was built from a representative sample of the proportions of owner satisfied with 40 diverse car models, both domestic and designed abroad.

**Model results.** Differences in price/quality, fuel economy and acceleration account for 72% of the variation in the proportion of car owners satisfied.

Increasing the fuel efficiency of hybrids, has the greatest potential impact to increase the proportion of domestic owners who are satisfied.



$$satisfaction_i = 100 \frac{e^{(-2.65 - .20 \text{ sec ondsi} + .099 \text{ MPG}i + .044 \text{ price}(\$K)i)}}{1 + e^{(-2.65 - .20 \text{ sec ondsi} + .099 \text{ MPG}i + .044 \text{ price}(\$K)i)}}$$

for owners of domestic hybrids

RSquare: .72<sup>a</sup>

<sup>a</sup>Significant at .01

Fuel efficiency matters more to owners of hybrids, potentially increasing the proportion of satisfied domestic owners by as much as 50%.

Price/quality and acceleration are also important satisfaction drivers which compensate to some degree for lower fuel efficiency.

**Conclusions.** Owners of hybrids would be more satisfied with more fuel efficient models, though higher priced luxury and responsiveness also drive satisfaction and partially compensate for less than ideal fuel efficiency.

*Example 13.2 Presidential Approval Proportion*<sup>2</sup>. The Republican National Committee is planning its 2008 Presidential campaign strategy, and management needs to know what drives public opinion of The President. Some believe that Presidential actions which signal defense strength rally public support, while others argue that defense references carry costs. The Committee is unsure which drives public opinion, the War on Terror and defense strength, or a healthier economy. They suspect that declining public opinion may be linked to fatalities in the ongoing war in Iraq or to slow growth in wages.

At least three shocks since re-election may have induced structural shifts in public opinion.

- In March 2006, the President signed the Patriot Act, legalizing government information gathering actions on suspected terrorists.
- In June 2006, the New York Times published an article describing illegal government information gathering actions. The White House asked for retraction, and the New York Times refused.
- In September 2006, President Bush focused a Labor Day speech on new job creation and designated September 11, 2006 as a day to remember the fifth anniversary of 911.

A structural change in political leadership probably also influenced public opinion:

- In November 2006 elections, Democrats gained control of Congress.

Public opinion polls track Americans' approval of the job The President is doing. The Roper Organization (<http://www.ropercenter.uconn.edu>) publishes results from a number of national polls. **Presidential Approval 13.3.xls** contains the *Approval Proportions* of 457 polls taken between President Bush's re-election in November 2004 and June 2007.

A consulting firm was retained to build a model of Presidential Approval which would identify and quantify drivers and provide short-term forecasts. After being briefed by Committee representatives, the consultants included

- an indicator, *Patriot*, following signing into law the Patriot Act
- an indicator, *NYT*, of the New York Times article
- an indicator *September 06* of the fifth anniversary of 911
- an indicator *Democratic Congress* in 2006 through 2007,
- cumulative military *fatalities* since re-election
- a leading indicator of past month average hourly *wage* of American workers

---

<sup>2</sup> This example is a hypothetical scenario using actual data.

The response variable which The Committee was interested in explaining and forecasting is *Proportion who Approve of The President*. This is a variable bounded below by zero and above by one hundred, so the consultants used the *Approval Logit* to estimate parameters.

Their model was:

$$\begin{aligned} \text{ApprovalLogit}_t = & b_0 + b_1 \text{Patriot}_t + b_2 \text{NYT}_t + b_3 \text{Sept06}_t + b_4 \text{DemCongress}_t + b_5 \text{fatalities}_t \\ & + b_6 \text{past month wage}_t \end{aligned}$$

The model correctly forecast the two most recent poll results and produced forecasts with a five percent margin of error. Recalibrated results are shown in Table 13.2.

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.896					
R Square	0.802					
Adjusted R Square	0.799					
Standard Error	0.112					
Observations	455					
ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	7	22.8	3.3	259.0	0.0000	
Residual	447	5.6	0.0			
Total	454	28.4				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
<i>Intercept</i>	-6.60	0.85	-7.7	0.0000	-8.28	-4.92
<i>Patriot</i>	-0.19	0.02	-10.6	0.0000	-0.22	-0.15
<i>NYT</i>	-0.30	0.08	-3.7	0.0002	-0.45	-0.14
<i>Sept 06</i>	0.20	0.03	7.1	0.0000	0.14	0.25
<i>Dem Congress elected surge</i>	-0.16	0.03	-4.7	0.0000	-0.22	-0.09
<i>wage (\$)</i>	-0.087	0.024	-3.6	0.0004	-0.13	-0.04
<i>fatalities (K) to date</i>	-0.39	0.01	-34.2	0.0000	-0.41	-0.36
<i>wage (\$)</i>	0.86	0.10	8.2	0.0000	0.66	1.07
<i>DW: 1.83</i>						

Table 13.2 Logit model of Presidential approval

The model accounts for much of the variation, 80%, in *approval logits*.

The Patriot Act, the New York Times article alleging government abuses of privacy, Democratic control of Congress and military fatalities reduce approval. The President's September 2006 focus on new jobs, followed by the memorial service commemorating the fifth year anniversary of 911, as well as growing wages, enhance public opinion.

The baseline equation, before renewal of the Patriot Act, is:

$$\widehat{ApprovalLogit}_t = -6.60 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

During the three months that followed passage of the Patriot Act, the model equation is:

$$\widehat{ApprovalLogit}_t = -6.79 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

After the New York Times publication, the equation is:

$$\widehat{ApprovalLogit}_t = -6.90 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

After the fifth 911 anniversary, the model equation is:

$$\widehat{ApprovalLogit}_t = -6.40 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

Following the 2006 election, the model equation is:

$$\widehat{ApprovalLogit}_t = -6.76 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

And following Bush's presentation of the Surge plan, the equation is:

$$\widehat{ApprovalLogit}_t = -6.69 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t$$

Re-writing the equations as expected odds:

$$\widehat{ApprovalOdds}_t = e^{(-6.60 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t)}$$

in baseline days before renewal of the Patriot Act,

$$= e^{(-6.79 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t)}$$

following renewal of the Patriot Act,

$$= e^{(-6.90 - .39 \text{ fatalities}(K) \text{ to date}_t + .86 \text{ wage}(\$) \text{ last month}_t)}$$

following the New York Times article,

$$= e^{(-6.4 - .39 \text{ fatalities}(K) \text{ to date} + .86 \text{ wage}(\$) \text{ last month})}$$

following September 2006,

$$= e^{(-6.76 - .39 \text{ fatalities}(K) \text{ to date} + .86 \text{ wage}(\$) \text{ last month})}$$

following the 2006 election, through 2007, and

$$= e^{(-6.69 - .39 \text{ fatalities}(K) \text{ to date} + .86 \text{ wage}(\$) \text{ last month})}$$

following the Surge plan speech.

Predicted *Approval Proportions*,

$$\text{Approval} \hat{P}r o p o r t i o n s_t = 100 * [ \text{Approval} \hat{O}d d s_t / (1 + \text{Approval} \hat{O}d d s_t) ]$$

are shown below by day from President Bush's re-election through June 2007.

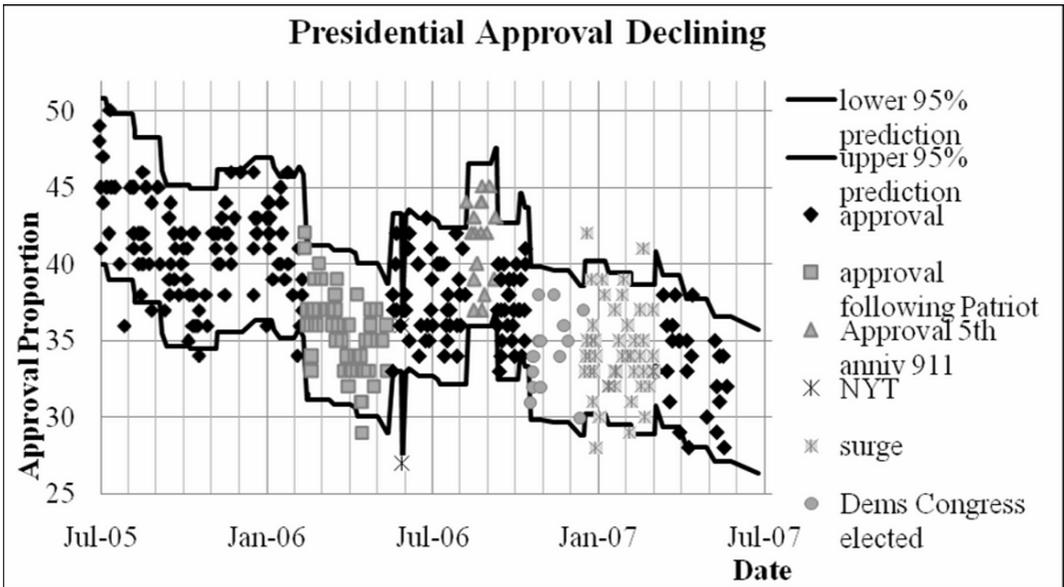


Figure 13.5 Presidential approval proportion

Predicted Presidential Approval is 51% in November 2004, following re-election. The predicted *Approval Proportion* declined gradually over the next sixteen months to 40% in March 2006. Following renewal of the Patriot Act in March 2006, a structural shift in public opinion occurred, reducing approval ratings by an estimated 4% for a three month period.

In June 2006, predicted approval is 38%, but dropped briefly to 31% following the New York Times article alleging government abuses of privacy. By September 2006, predicted approval is 37%. The President's commemoration of the fifth anniversary of 911, stimulated a brief structural shift, raising predicted *Approval Proportions* an estimated 5%.

Before the 2006 election, predicted approval is 38%. With Democratic wins insuring a Democratic Congress, a structural shift reduces approval proportions by an estimated 3%. In January 2007, after The President's presentation of the Surge plan for increased troop involvement in Iraq, predicted approval drops 2% to 35%. Increasing military fatalities and falling hourly wages bring predicted approval to a low of 31% by July of 2007. The margin of error in forecasts is five percent.

The National Committee now has evidence that the both the continuing war effort and the domestic economy, in the form of hourly wages, are driving public opinion. Democratic control of Congress is reducing approval, as well.

## 13.2 Logit Models Provide the Means to Build Valid Models of Shares And Proportions

When responses are bounded below and above, we must build this into our models to get accurate pictures of drivers and valid forecasts. Rescaling shares or proportions to odds, and then to their natural logarithms, the logits, gives us more valid models. Though both odds and logits are unbounded, the corresponding predicted proportions or shares are bounded below and above, guaranteeing believable forecasts.

## Excel 13.1 Rescale a limited dependent variable to logits

*Proportion who would try Pampers Preemies.* We will build a model of intent to try Procter & Gamble’s new preemie diapers. Procter & Gamble management believes that their new diaper may attract mothers who were choosing cloth diapers. Natural composition is a known advantage of cloth diapers. We will build a model of trial intentions to see whether the importance of natural composition and selected demographics are drivers.

**Rescale bounded dependent variables to unbounded logits.** In concept test data, **Excel 13.1 Pampers Concept Test.xls**, we have the *trial intentions* of 97 preemie mothers, measured on a 5-point scale (“Definitely Not”=.05, “Probably Not”=.25, “Maybe”=.5, “Probably”=.75, “Definitely”=.95).

From *trial intent* in **A**, which is bounded between zero and one, make

- *trial odds*, the chance of trying to the chance of not trying, and
- *trial logit*, the natural logarithms of the *trial odds*.

Insert two new columns: Select **B** and **C**, **Alt HIC**, and add labels *trial odds* and *trial logit*.

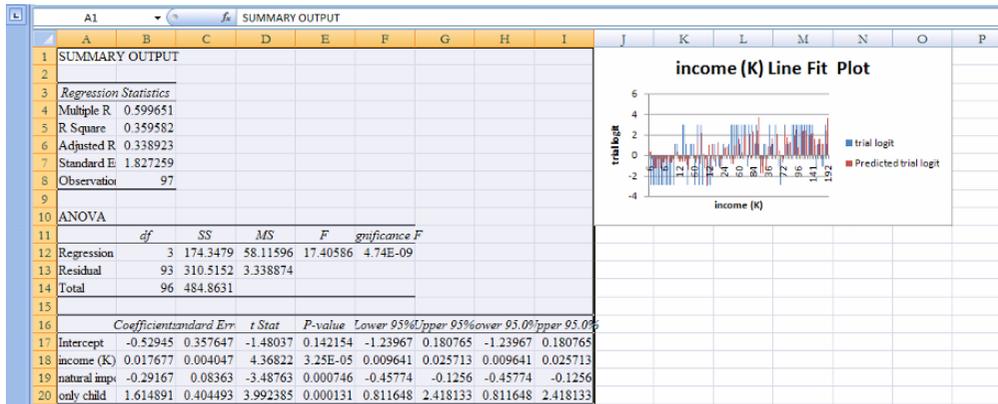
In **B2**, enter the formula for *trial odds* = $A2/(1-A2)$  [Enter].

In **C2**, enter the formula for *trial logit* = $LN(B2)$  [Enter], then select **B2:C2** and double click the lower right corner to fill in the columns:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	trial intention at premium price	trial odds	trial logit	income (K)	natural importance	only child										
2	0.05	0.053	-2.94	6	3	1										
3	0.05	0.053	-2.94	6	3	1										
4	0.05	0.053	-2.94	6	3	0										

The concept test measures include the importance rating of natural composition, *natural Importance*, and household demographics, *income (\$K)* and an indicator of absence of other children in the households, *only child*. We will include the importance of natural composition and these demographics in the model.

Run regression of *trial logit* in **C** with *income*, *natural importance* and *only child*:



All coefficient signs are “correct.”

- Mothers from higher income households with no other children are more likely to try.
- Mothers who rate natural composition of diapers as more important are less likely to try.

The model equation is:

$$\text{logit}(\hat{tr}ial_i) = -.53 + 1.61 \text{ only child}_i - .29 \text{ natural importance}_i + .018 \text{ income}(\$K)_i$$

**Sensitivity analysis.** To quantify the influence of each driver, find predicted trial intentions for hypothetical combinations of the three predictors.

To find the sample ranges for each, find the

- *minimum*, using the Excel function **MIN(array)**,
- *median*, using the Excel function **MEDIAN(array)**, and
- *maximum*, using the Excel function **MAX (array)**

of *Income(\$K)* and *Natural Importance* in **D** and **E**.

In **A101:A103** type in those labels.

In **D101**, enter **=MIN(D2:D98)** [Enter].

In **D102** enter **=MEDIAN(D2:D98)** [Enter].

In **D103** enter **=MAX(D2:D98)** [Enter], then select **D101:D103**, grab and drag through **E**:

	A	B	C	D	E	F
trial intention						
at premium	price	odds	logit	income (\$K)	natural importance	only child
96	0.25	0.333	-1.1	146	9	1
97	0.95	19	2.944	156	5	1
98	0.75	3	1.099	192	3	1
99						
100						
101	min			6		1
102	median			48		3
103	max			199		9

Compare the marginal impact of each driver when the other three drivers are at most favorable and unfavorable levels.

**Natural composition.** First add twelve hypothetical mothers to the bottom of the dataset: Select rows **99:110**, **Alt HIR**.

Enter hypothetical premie mom characteristics in columns **A, D, E** and **F** for

- six mothers with lowest *income (\$K)* (6)
  - three with no other children (*only child* is 0) and
  - three with, other children (*only child* is 1),
- six with highest *income (\$K)* (199)
  - three with no other children (*only child* is 0) and
  - three with, other children (*only child* is 1),

Within each set of three demographically identical moms, let

- one rate natural composition unimportant (*natural importance* is 1),
- one rate natural composition of median importance (*natural importance* is 3),
- one rate natural composition of greatest importance (*natural importance* is 9):

	A	B	C	D	E	F
trial intention						
at premium	price	odds	logit	income (\$K)	natural importance	only child
96	0.25	0.333	-1.1	146	9	1
97	0.95	19	2.944	156	5	1
98	0.75	3	1.099	192	3	1
99	low income, no other kids			6	1	0
100	low income, no other kids			6	3	0
101	low income, no other kids			6	9	0
102	low income, other kids			6	1	1
103	low income, other kids			6	3	1
104	low income, other kids			6	9	1
105	high income, no other kids			199	1	0
106	high income, no other kids			199	3	0
107	high income, no other kids			199	9	0
108	high income, other kids			199	1	1
109	high income, other kids			199	3	1
110	high income, other kids			199	9	1

**Predicted Trial Logits.** Use the coefficient estimates from your regression output sheet to make *predicted trial logits* in **H**, using the regression equation.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	trial intention at premium price	trial odds	trial logit	income (K)	natural importance	only child	predicted trial logits									
93		0.75	3	1.099	144	7	1	1.59								
94		0.75	3	1.099	144	7	1	1.59								
95		0.25	0.333	-1.1	146	9	1	1.04								
96		0.25	0.333	-1.1	146	9	1	1.04								
97		0.95	19	2.944	156	5	1	2.38								
98		0.75	3	1.099	192	3	1	3.60								
99	low income, no other kids				6	1	0	-0.72								
100	low income, no other kids				6	3	0	-1.30								
101	low income, no other kids				6	9	0	-3.05								

**Rescale to Find Predicted Trial Intentions.** Rescale *predicted trial logit* to *predicted odds* in **I** and *predicted trial intention* in **J**:

In **I2** enter =EXP(H2) [Enter].

In **J2** enter =I2/(1+I2) [Enter], then select **I2:J2** and double click the lower right corner to fill in the columns:

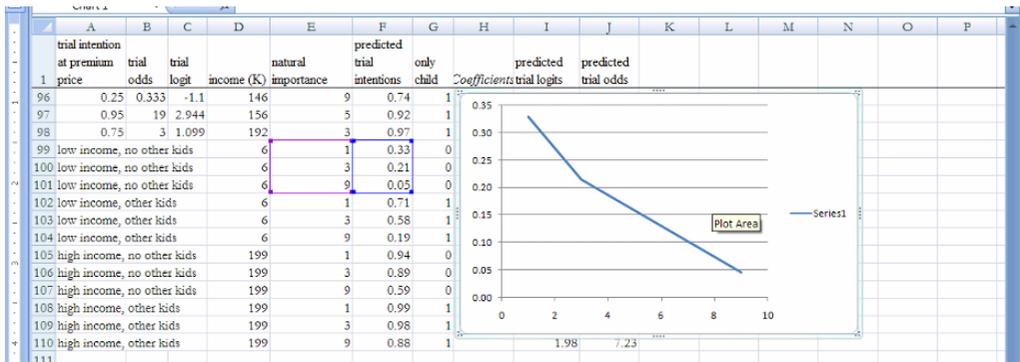
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	trial intention at premium price	trial odds	trial logit	income (K)	natural importance	only child	predicted trial logits	predicted trial odds	predicted trial intentions							
96		0.25	0.333	-1.1	146	9	1	1.04	2.83	0.74						
97		0.95	19	2.944	156	5	1	2.38	10.86	0.92						
98		0.75	3	1.099	192	3	1	3.60	36.76	0.97						
99	low income, no other kids				6	1	0	-0.72	0.49	0.33						
100	low income, no other kids				6	3	0	-1.30	0.27	0.21						
101	low income, no other kids				6	9	0	-3.05	0.05	0.05						
102	low income, other kids				6	1	1	0.90	2.46	0.71						
103	low income, other kids				6	3	1	0.32	1.37	0.58						
104	low income, other kids				6	9	1	-1.43	0.24	0.19						
105	high income, no other kids			199	1	0	2.70	14.83	0.94							
106	high income, no other kids			199	3	0	2.11	8.28	0.89							
107	high income, no other kids			199	9	0	0.36	1.44	0.59							
108	high income, other kids			199	1	1	4.31	74.55	0.99							
109	high income, other kids			199	3	1	3.73	41.60	0.98							
110	high income, other kids			199	9	1	1.98	7.23	0.88							

**Illustrate synergies between predictors.** To see the synergies between the importance of natural composition, income, and absence of other children, use shortcuts to move *predicted trial intentions* next to *natural importance*:

Select **J**, **Cntl+X**, then select **F**, **Alt HIE**.

Plot *predicted trial intentions* by *natural importance*, making each set of three demographically identical moms a separate series.

Select **E99:F101**, **Alt ND**:



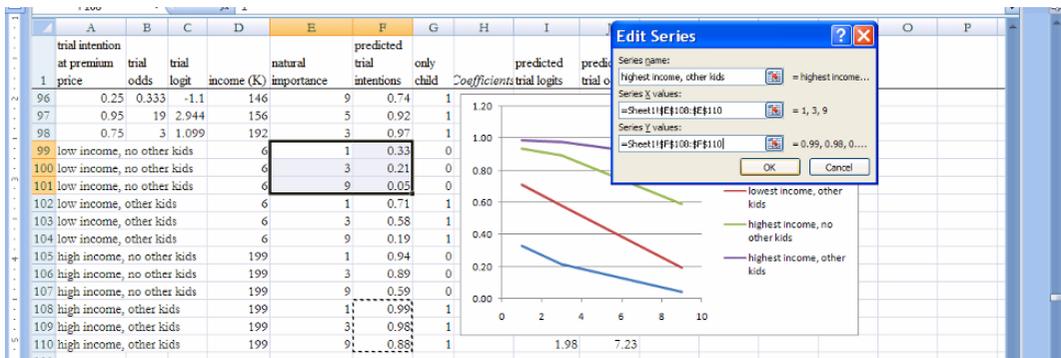
Right click inside the chart and **Select Data**.

**Edit Series 1** and enter **Name lowest income other kids**.

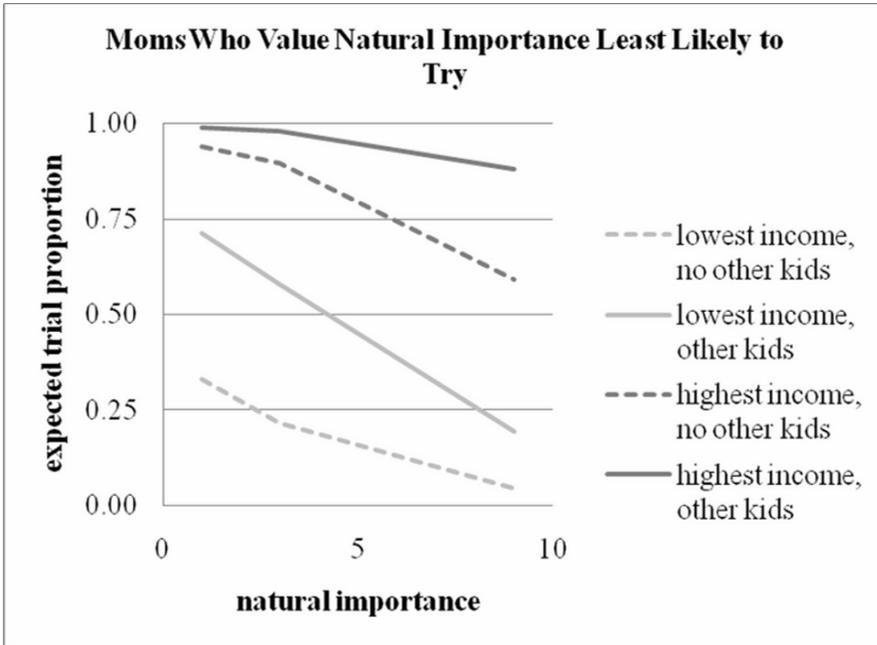
Add, with **Name, highest income other kids**, **X Values, E102:E104**, **Y Values, F102:F104**,

Add, **Name, lowest income no other kids**, **X Values, E105:E107**, **Y Values, F105:F107**,

Add, **Name, highest income no other kids**, **X Values, E108:E110**, **Y Values, F108:F110**.



Add title and axes titles, **Finish**:



**Find the marginal difference that natural composition makes given alternate demographics.** To quantify the marginal difference that the importance of natural composition makes in expected trial intention, add column **K** with label *marginal difference in expected trial intention*.

In **K99**, enter  $=F99-F101$  [Enter],  
 in **K102** enter  $=F102-F104$  [Enter], and  
 in **K105** enter  $=F105-F107$  [Enter]:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	price	trial odds	trial logit	income (K)	natural importance	predicted trial intentions	only child	Coefficients	predicted trial logits	predicted trial odds	marginal difference in trial intentions					
99	low income, no other kids			6	1	0.33	0		-0.72	0.49						
100	low income, no other kids			6	3	0.21	0		-1.30	0.27						
101	low income, no other kids			6	9	0.05	0		-3.05	0.05	0.28					
102	low income, other kids			6	1	0.71	1		0.90	2.46						
103	low income, other kids			6	3	0.58	1		0.32	1.37						
104	low income, other kids			6	9	0.19	1		-1.43	0.24	0.52					
105	high income, no other kids			199	1	0.94	0		2.70	14.83						
106	high income, no other kids			199	3	0.89	0		2.11	8.28						
107	high income, no other kids			199	9	0.59	0		0.36	1.44	0.35					
108	high income, other kids			199	1	0.99	1		4.31	74.55						
109	high income, other kids			199	3	0.98	1		3.73	41.60						
110	high income, other kids			199	9	0.88	1		1.98	7.23	0.11					





## *Assignment 13-1 Big Drug Co Scripts*

The leading manufacturer of a popular anti-allergy drug would like to know how reformulations affect their share of prescriptions dispensed. Big Drug's major competition comes from generic copycat brands. When the generic competition begins to gain share, Big Drug introduces a reformulation, which sends the generics back to the lab to reformulate their copies. Reformulation is expensive, since it includes research and development, as well as repackaging and reformulating promotional materials.

Semi annual data in **Assignment 13-1 Big Drug Co.xls** include time series of a semi annual counter of time periods, the share of prescriptions dispensed of Big Drug Co's anti-allergy drug, and indicators for a major and a minor reformulation.

*Build a logit trend model to estimate the impact of reformulations on Big Drug Co's share and to forecast Big Drug Co's share in the next five years.*

*Write a one-page memo to Big Drug Co management concerning the impact of reformulations on share and share forecasts for the next five years. Embed one figure to illustrate your results. Include in your memo:*

- *Share estimates had the drug not been reformulated*
- *Suggested date for Big Drug Cos introduction of Reformulation 3, and recommendations for either a major or a minor reformulation*

## *CASE 13-1 Alltel's Plans to Capture Share in the Cell Phone Service Market\**

Alltel offers competitive cell phone network service in a limited geographic area. Buoyed by their success against the big competitors, Verizon, Sprint, t-mobile and Cingular, Alltel has plans to expand into more areas and to increase their share in existing markets. In twenty cities, samples of 1,000 cell phone customers were drawn and surveyed. Survey measures included *service provider*, *satisfaction*, *service coverage rating*, *dropped calls rating*, and *static rating*.

Ratings were on a five point scale, where a higher number indicated better service.

In the data file, **Case 13-1 Alltel.xls**, are

- *City*
- *Service provider*
- *proportions of customers satisfied*
- *coverage rating*
- *dropped calls rating*, and
- *static rating*
- *cingular*
- *sprint*
- *t-mobile*
- *Verizon*.

Alltel is the baseline.

Build a model of customer satisfaction for the Alltel executives which quantifies the importance of *service provider*, *coverage*, *dropped calls*, and *static*.

*Proportion satisfied* is a limited dependent variable with values between 0 and 100. Rescale to acknowledge these limits.

**PivotCharts and indicator interactions.** Executives are counting on their hunch that Sprint customers are increasingly dissatisfied with lack of network *coverage*. Few of Sprint's new phones have analog capability, limiting coverage in rural areas. This is an opportunity for Alltel, if it can be confirmed that coverage influences customer satisfaction.

Make a PivotChart to compare average *coverage* ratings by service provider. Do *Sprint* customers rate coverage lower than other networks' customers?

Executives believe that *Verizon* has achieved a competitive advantage with a low percentage of *dropped calls*. This is could be an opportunity for Alltel to attract *Verizon*

---

\*The case is a hypothetical scenario using actual data.

customers, if it can be confirmed that *dropped calls* lead to customer dissatisfaction, and if Alltel can achieve a superior *dropped calls rating*.

Make a PivotChart to compare average *dropped call* ratings by service provider. Do *Verizon* customers rate dropped calls higher than customers of other networks?

According to research reports, *t-mobile* customers are dissatisfied with *static* in the network. This is an opportunity for Alltel, since Alltel service is crystal clear.

Make a PivotChart to compare *static* ratings by service provider. Do *t-mobile* customers rate *static* ratings lower than other network customers?

To incorporate executive judgment, include in your model, interactions between

- *Sprint* and *coverage*
- *Verizon* and *dropped calls*
- *t-mobile* and *static*.

Fit your model, first removing insignificant indicator interactions, and then removing insignificant variables and indicators.

- If an indicator interaction is significant, but either one of the main effects involved in the interaction are not, keep the main effects in the model to support the interaction.
- Since the indicator interactions are based on executive judgment, use one tail *t-tests* of the coefficient estimates by dividing the two tail *p values* by 2

Use your coefficient estimates to make *predicted logits*, and then rescale to make *predicted proportion satisfied*.

Write your equations for the *predicted satisfaction odds*.

Please use proper subscripts, superscripts, and indentations.

- For Alltel customers,
- For *Sprint* customers,
- For *Verizon* customers, and
- For *t-mobile* customers.

Alltel management believes that it is possible to improve one service aspect—*coverage*, *dropped calls*, OR *static*—to achieve ratings that are **one point higher** within the next year.

- Which service aspect improvement would make the greatest difference in the *expected proportion* of Alltel customers satisfied?

- How much would the *expected proportion* of customers satisfied change with this improvement of **one rating scale point** in a single service aspect?

Alltel managers are aware that competitors will also focus on service improvements.

- If *Sprint*, *t-mobile* and *Verizon* managements decided to improve their weakest service aspect to achieve ratings that were higher by **one rating scale point**, what aspect would each choose?
- How much difference in the *expected proportion* of customers satisfied would improvement by *one rating scale point* in the weakest service dimension make for each?

Add hypothetical services to the data file, comparing *predicted customer satisfaction proportions* across the competing service providers, Alltel, *Sprint*, *t-mobile* and *Verizon* given **current average** service aspect ratings and hypothetical improvements in each of the three service aspects.

(If a service provider, such as Alltel, has a current average rating of 3.3 along a service aspect, such as static, consider hypothetical services with static ratings of 4 and 5, adding three hypothetical Alltel rows.)

Make three scatterplots showing expected response in the *proportion of customers satisfied* following these hypothetical improvements in *coverage*, *dropped calls*, and *static*.

- If *Sprint*, *t-mobile* and *Verizon* managements used statistics to achieve competitive advantage, which service aspect would they each work to improve first?
  - How much difference in the *expected proportion of customers satisfaction* would improvement by **one rating scale point** in this single aspect make for each?
- Which competitor(s) pose the greatest threat to Alltel: Which competitor(s) could achieve a greater *proportion of customers satisfied* than Alltel?
  - What service aspect(s) would the most threatening competitor(s) need to improve to satisfy more customers than Alltel?

*Case 13-2 Pilgrim Bank (A): Customer Profitability\* and Pilgrim Bank (B): Customer Retention\*\**

Use the file **Case 13-2 Pilgrim.xls** for data analysis and preparation for class discussion.

---

\*Harvard Business School Case 9602095

\*\*Harvard Business School Case 9602103

# Index

## A

approximate 95% Confidence Intervals, 43  
attribute importance, 282, 295–299  
autocorrelation, 242, 253–257

## B

bounded dependent variable, 378–398  
built in synergies, 315, 334, 346, 353, 381, 394

## C

categorical, 11–12, 15–16  
Central Tendency, 11–12  
column chart, 15–16, 27–28, 61–65, 71–72  
confidence interval, 41–43, 49–58, 60–61, 70–1, 74–77  
    alternate scenarios, pairs, 54–58, 74–77  
    conservative, 55  
    margin of error, 45  
    one sample, 41–44, 60–61  
    proportion, 50–54  
    two sample, two segment, 49–50, 70–71  
conjoint analysis, 278–283, 295–299  
    attribute importance, 282, 295  
    hypotheticals, 279–280, 295  
    orthogonal array, 280–281  
    part worth utilities, 279–283, 295–296  
contingency analysis, 171–192  
    *chi square*, 174–177, 187–190  
    *chi square*, sparse cells, 175–177  
    conditional probability, 171–174  
    crosstabulation, 171–172  
    joint probability, 171–172  
    Simpsons Paradox, 177–182  
    sparse cells, 175–177  
continuous, 11–13  
correlation, 105–113  
    and regression, 109–113

correlation, cont.  
    to choose lags, 249  
cross sectional  
    difference between cross sectional and time series, 243  
Crystal Ball, 44–47, 65–69  
    90% confidence interval, 45  
    assumptions, 44–47, 65–68  
cumulative distribution, 7, 23

## D

descriptive statistics, 5–30  
dispersion, 11  
dummy variables, 275–305  
Durbin Watson, 242–246, 253–257

## E

Empirical Rule, 13–14  
equations, 91, 103–104, 202, 224, 275, 277, 279, 288–289, 292–293, 301–303, 319–320, 333–334, 343–344, 347–348, 354, 377, 379–381, 387–389, 392  
    in logits, 377, 379–381, 387–389, 392  
    interactions, 343–344, 347–348, 354  
    natural logarithms, 347–348  
    rescaling from logits, 380–381, 388–389  
    square roots, 320, 334, 354  
    standard format, 103–104  
    with indicator variables, 275, 277, 279, 288–289, 292–293, 302–303  
Excel  
    autocorrelation, assess, 253–257  
    chi square, PivotTable, 187–190  
    column chart, 27–28, 61–65, 71–72  
    confidence interval, 60–63, 70–71, 76–77  
    alternate scenarios, pairs, 76–77  
    one sample, 60–63  
    two segments, 71–72  
    conjoint analysis, 295–299  
    contingency analysis, 185–194

- Excel, contingency analysis, cont.
- chi square, 187–190
  - summary data, 190–192
- correlation, 124–125
- crosstabulation, PivotTable, 185–187
- Crystal Ball, 65–69
- Durbin Watson, 253–257
- fit and forecast, 260–263
- forecasting, 258–271
- Durbin Watson, 262
  - illustrate fit and forecast, 260–263, 365–367
  - impact of drivers, 263–264, 334–337, 367–369
  - lag, choice of, 250–253
  - prediction intervals, 258–260, 301–302
  - predictions from model equation, 257–260, 301–303, 333–334, 336–337, 363, 367–368, 392–394
  - recalibrate, 259–260, 302–303, 364–365
  - validation, 257–259, 301–302, 363–364
- histogram, 20
- hypothesis test, 59–60, 69, 74–76
- alterante scenarios, pairs, 74–76
  - one sample, 59–60
  - two sample, 69
- indicator variables, 295–305
- interactions, 326–337
- adding, 361–362
  - illustrate fit and forecast, 365–367
  - sensitivity analysis, 367–369
- lag, choice of, 250–253
- logit regression, 386–398
- equations, 393
  - marginal impact, 392–398
  - rescale, 391–398
    - bounded dependent variable to logits, 391
    - bounded dependent variable to odds, 391
    - from logits, 394
    - from odds, 394
    - odds to logits, 391
- Excel, logit regression, cont.
- sensitivity analysis, 392–398
  - synergies, 394–398
- marginal impact of drivers, 221–227, 263–264, 334–336, 367–369, 393–396
- model building, 224–35
- autocorrelation, assess, 253–257
  - Durbin Watson, 253–257
  - forecasting, 250–265
  - illustrate fit and forecast, 260–263, 365–367
  - impact of drivers, 263–264, 334–337, 367–369
  - lag, choice of, 250–253
  - multicollinearity symptoms, 216
  - partial F test*, 217–220
  - prediction intervals, 258–260, 301–302
  - predictions from model equation, 257–260, 301–303, 333–334, 336–337, 363, 367–368, 392–394
  - sensitivity analysis, 221–226, 263–265, 297–299, 303–305, 334–336, 367–369, 393–394
  - time series, 250–265
- model validation, 257–259, 301–302, 363–364
- monte carlo simulation, 67–71
- multicollinearity symptoms, 216
- multiple regression, 216–227
- partial F test*, 217–220
  - sensitivity analysis, 221–226
- nonlinear regression, 326–337
- assess skewness, 326–327
  - equation, square roots, 334
  - marginal impact, 334–337
  - marginal response, 334–337
  - rescale, 327–328, 334, 336
    - back from square roots, 334
    - inverses, 328
    - natural logarithms, 327–328
    - square roots, 327–328
  - sensitivity analysis, hypotheticals, 336
  - synergies, 335–336

## Excel, cont.

- partial F test, 217–220
- pie chart, 74–75
- PivotChart, PivotTable, 26
- portfolio analysis, 170–175
  - beta, 172
  - Efficient Frontier, 172–175
  - expected rate of return, beta, 170–171
- prediction intervals, 258–260, 301–302
- predictions from model equation,
  - 257–260, 301–303, 333–334,
  - 336–337, 363, 367–368, 392–394
- recalibrate, 259–260, 302–303, 364–365
- regression, 114–127
- rescale, 326–328
- sensitivity analysis, multiple
  - regression, 221–226
- shortcuts, 29–30, 78–79, 126–127, 193–194
- t* test, 59–60, 69, 74–76
  - one sample, 59–60
  - paired, alternative scenarios, 74–76
  - two segments, two samples, 69
- time series, 253–264, 301–303, 333–337, 363–369
  - autocorrelation, assess, 253–257
  - Durbin Watson, 253–257
  - illustrate fit and forecast, 260–263, 365–367
  - impact of drivers, 263–264, 334–337, 367–369
  - lag, choice of, 250–253
  - prediction intervals, 258–260, 301–302
  - predictions from model equation, 257–260, 301–303, 333–334, 336–337, 363, 367–368, 392–394
  - recalibrate, 259–260, 302–303, 364–365
  - validation, 257–259, 301–302, 363–364
- validation, 257–259, 301–302, 363–364

**F**

- forecasting, 235–265
  - autocorrelation, 242, 254–257

## forecasting, cont.

- correlation to choose lags, 241, 244, 252–253, 256
- Durbin Watson, 242–246, 253–257
- hold out observations, 241
- inertia, 238–239
- interactions, 343–344
- lag, choice of, 239–241, 244, 250–253, 256
- Leading Indicator, 238
- recalibration, 246, 259–260
- residual analysis to identify
  - unaccounted for trend or cycles, 242–244, 253–256
- validation, 235, 241, 246, 257–259
- variable selection, time series, 237–239

**G**

- gains from nonlinear regression, 324

**H**

- histogram, 5–6, 17–19
- hold out observations, 249
- hypothesis, 38–40, 48–49, 54–57, 59–60, 69, 74–76
  - alternate scenarios, pairs, 54–57, 74–76
  - alternative, 38
  - null, 38
  - one sample, 38–40, 59–60
  - paired, alternate scenarios, 54–57, 74–76
  - two segment, two sample, 48–49, 69
- hypotheticals, 222–223, 279–280, 295, 334–336, 356–357, 368, 381–384, 392–393

**I**

- indicator variables, 275–305
  - conjoint analysis, 278–283, 295–299
    - hypotheticals, 279–280, 295
    - part worth utilities, 279–283, 295
  - equations, 275–277, 279, 286, 288–289
  - modify intercept, 275–276
  - seasonality, 283–290
  - segment differences, 276–278
  - structural shift, 291–293, 299–305

indicator variables, cont.  
 value of product attributes, 278–283,  
 295–299  
 inertia, 238–239, 255  
 inference, 35–77  
 interactions, 343–369  
 baseline, 343–344, 347, 351, 361  
 built in synergies, 346, 348–349,  
 353–355  
 equations, 343–344, 347–348, 354  
 main effect not significant, 347  
 modify slope, 343–344, 348–349  
 segment response differences, 343–350  
 sensitivity analysis, 356–357, 367–369  
 structural shifts, 351–69  
 time series, 359–69

## J

jointly significant, 209

## L

lag, choice of, 239–241, 244, 250–253,  
 256  
 Leading Indicator, 238  
 limited, dependent variable, 377–398  
 logit regression, 377–398  
 built in synergies, 381–384, 394–396  
 equations, 377, 379–381, 387–389  
 limited or bounded dependent variable,  
 377  
 logits, 377, 379–380, 387–388,  
 391–392  
 odds, 377, 380, 388  
 rescaling, 377, 379, 380, 387–388, 391,  
 394  
 back from logits, 380, 388, 394  
 to logits, 377, 379, 387, 391  
 to odds, 380, 388, 394  
 s shaped response, 377

## M

margin of error, 43–44, 60–62, 70–71, 73,  
 76–77  
 memos, 147–148  
 model building, 201–227, 235–265,  
 275–305

model building, cont.  
 autocorrelation, 242, 253–257  
 correlation to choose lags, 241, 244,  
 252–253, 256  
 cross sectional versus time series, 243  
 equation, 202, 206, 209, 224  
*F test*, multiple regression, 204  
 forecasting, 239–244, 246, 253–257,  
 259–260  
 autocorrelation, 242, 253–257  
 lag, choice of, 239–241, 250–253  
 recalibration, 246, 259–260  
 residual analysis to identify  
 unaccounted for trend or cycles,  
 242–244, 253–256  
 goals, 201, 235  
 indicator variables, 275–305  
 inertia, 238–239  
 joint significance, 209  
 Leading Indicator, 238  
 marginal response, multiple regression,  
 202  
 multicollinearity, 203–209, 217–220  
 joint significance, 209  
*partial F test*, 207–209, 217–220  
 remedies, 206–207  
 symptoms, 205,  
 multiple regression, 201–227  
 equation, 202, 224, 275, 277, 279,  
 288–289, 292–293, 301–303,  
 319–320, 333–334, 343–344,  
 347–348, 354, 377, 379–381,  
 387–389, 392  
*F test*, 204  
 joint significance, 209  
 marginal response, 202  
 multicollinearity, 203–209, 217–220  
 partial F test, 207–209, 217–220  
 remedies, 206–207  
 symptoms, 205  
 RSquare, 212  
 sensitivity analysis, 211–213,  
 221–227, 320–322, 334–337,  
 356–357, 367–369  
*partial F test*, 207–209  
*RSquare*, multiple regression, 212

model building, cont.  
 sensitivity analysis, 211–213, 221–227,  
 320–322, 334–337, 356–357,  
 367–369  
 time series, 235–246, 250–259  
 autocorrelation, 242, 253–257  
 hold out observations, 241  
 lag, choice of, 239–241, 244,  
 250–253, 256  
 recalibration, 246, 253–257  
 residual analysis to identify  
 unaccounted for trend or cycles,  
 242–244, 253–256  
 validation, 235, 241, 246, 257–259  
 validation, 235, 241, 246, 257–259  
 variable selection, logic, 201–202  
 variable selection, time series, 237–246  
 model building process, 201–227,  
 235–246  
 monte carlo simulation, 44–47, 65–69

## N

nominal, 12  
 nonlinear regression, 331–337  
 built in synergies, 315, 334–338  
 equation, square roots, 320, 334  
 nonconstant response, 313  
 Normalize positively skew, 314–315,  
 327  
 relative strength of drivers, 320–322,  
 334–337  
 rescaling, 314–315, 317, 320, 324,  
 327–328, 334, 348  
 back from square roots, 320, 334  
 from natural logarithms, 348  
 gains, 324  
 negative values, inverses, 314–315  
 square roots, natural logarithms,  
 317, 327–328  
 sensitivity analysis, 320–322, 334–337  
 square roots, natural logarithms, 317,  
 320, 327–328, 334  
 Tukey's Ladder of Powers, 313–315,  
 327  
 Normalize positively skewed, 314–315,  
 327

Normally distributed, 12–14

## O

one tail test, 39–41  
 orthogonal array, 279–280  
 outliers, 7–10, 20–22

## P

*p value*, 39, 59–60, 69, 74  
 part worth utilities, 279–283, 295–299  
*partial F test*, 207–209, 217–220  
 pie chart, 54, 72–73  
 PivotChart, PivotTable, 24–28, 172–173,  
 185–187, 190–192  
 portfolio analysis, 149–168  
 beta, 152–160, 165–166  
 Efficient Frontier, 161, 166–168  
 expected rate of return, 149–151, 158,  
 164–165  
 PowerPoints, 145–147  
 predicted performance,  $\hat{y}$ , 91  
 prediction intervals, 99–102, 118–123

## Q

quantitative, 11–12

## R

recalibration, 246, 259–260  
 regression, 91–127  
 ANOVA, 95  
 conditional mean prediction intervals,  
 101–102, 122–123  
 equation, 92–93, 114–115  
 equation, standard format, 114–115  
*F test*, 93–96  
 heteroskedasticity, 98, 116  
 mean square error, MSE, 94  
 prediction intervals, 99–100, 118–123  
 regression sum of squares, SSR, 94–95  
 residuals, 93–94, 98–99, 116–117  
 plot, 98, 114, 116  
 Normal, 99, 117  
*RSquare*, 95, 107  
 sensitivity analysis, 101  
 slope, 96–98, 109–112

regression, cont.  
 standard error, 94–95, 99–100, 116  
 sum of squared errors, SSE, 94  
 relative strength of drivers, 320–322,  
 334–337  
 rescaling, 318, 320, 324, 334, 348,  
 377–379, 387, 391–392  
 from bounded dependent variable to  
 logits, 377, 379, 387, 391  
 from limited dependent variable to  
 logits, 377, 379, 387, 391  
 from natural logarithms, 348  
 from square roots, 320, 334  
 gains, 324  
 negative values, inverses, 318  
 s shaped response, 377–378  
 to logits, 377, 379, 387, 391  
 to odds, 392  
 to square roots, natural logarithms, 317,  
 327–328  
 residual analysis to identify unaccounted  
 for trend or cycles, 242–244, 253–256  
 round, 10

## S

scale, 11–12  
 seasonality, 283–289  
 sensitivity analysis, 219–222, 328–331  
 significance level, 39, 69

skewness, 313–319, 326, 328  
 assess, 315–316, 326–327  
 correction, 317–318, 327–328  
 Normalize positively skew, 317,  
 327–328  
 rescaling negative values, inverses,  
 318, 328  
 Tukey's Ladder of Powers, 313–315,  
 327  
 standard error, 36–38, 51, 53, 57, 59, 70  
 structural shift, 291–293, 299–305  
*Student t*, 36–38

## T

time series  
 autocorrelation, 242, 254–257  
 correlation to choose lags, 241, 244,  
 252–253, 256  
 difference from cross sectional, 243  
 Durbin Watson, 242–246, 253–257  
 interactions, 351–377  
 residual analysis to identify  
 unaccounted for trend or cycles,  
 242–244, 253–256  
 variable selection, 237–239  
 Tukey's Ladder of Powers, 313–315, 327

## V

validation, 235, 241, 246, 249, 257–259