

Marketing Analytics

Data-Driven Techniques with Microsoft® Excel®

Wayne L. Winston

WILEY

Marketing Analytics: Data-Driven Techniques with Microsoft® Excel®

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Copyright © 2014 by Wayne L. Winston

Published by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-1-118-37343-9
ISBN: 978-1-118-43935-7 (ebk)
ISBN: 978-1-118-41730-0 (ebk)

Manufactured in the United States of America
10987654321

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com

Library of Congress Control Number: 2013954089

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. Microsoft and Excel are registered trademarks of Microsoft Corporation. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

To my wonderful family: Gregory, Jennifer, and Vivian

Credits

Executive Editor

Robert Elliott

Project Editor

Victoria Swider

Technical Editor

Lopo Rego

Production Editor

Daniel Scribner

Copy Editor

San Dee Phillips

Editorial Manager

Mary Beth Wakefield

Freelancer Editorial

Manager Rosemarie Graham

Associate Director of

Marketing David Mayhew

Marketing Manager

Ashley Zurcher

Business Manager

Amy Knies

Vice President and Executive Group

Publisher

Richard Swadley

Associate Publisher

Jim Minatel

Project Coordinator, Cover

Katie Crocker

Proofreaders

Josh Chase, Word One

Louise Watson, Word One

Indexer

Ron Strauss

Cover Image

Wiley

Cover Designer

Ryan Sneed

About the Author



Wayne Winston is Professor Emeritus at the Indiana University Kelley School of Business and is currently a Visiting Professor at the University of Houston Bauer College of Business. Wayne has won more than 45 teaching awards at Indiana University. These awards include six school-wide MBA teaching awards. He has authored 25 reference journal articles and a dozen books including, *Operations Research : Applications and Algorithms*

(Cengage, 1987), *Practical Management Science* (Cengage, 2011), *Data Analysis and Decision Making* (Cengage, 2013), *Simulation Modeling with @RISK* (Cengage, 2004), *Mathletics* (Princeton, 2009), and *Excel 2013 Data Analysis and Business Modeling* (O'Reilly, 2014). Wayne has also developed two online courses for Harvard Business School: Spreadsheet Modeling, and Mathematics for Management. He has taught Excel modeling and consulted for many organizations including the U.S. Army, the U.S. Navy, Broadcom, Cisco, Intel, Pfizer, Eli Lilly, Ford, GM, PWC, Microsoft, IAC, Deloitte Consulting, Booz Allen Hamilton, QAS, eBay, the Dallas Mavericks, and the New York Knicks. Lastly, Wayne is a two-time *Jeopardy!* champion.

About the Technical Editor



Lopo Rego joined the Kelley School of Business at Indiana University in 2011 as an Associate Professor of Marketing. Trained in Economics, he “converted to the dark side” during his MBA and has since been interested in understanding the association between marketing strategy and firm performance. This proved to be a life-long quest, leading him to Ann Arbor where he eventually earned his Ph.D. in Marketing at

the University of Michigan's Ross School of Business. Not surprisingly, his research interests focus primarily in understanding how marketing decisions, strategies, and investments translate into firm performance, be it at the product-marketplace level, financial-accounting level or shareholder wealth level. Additionally, Lopo is interested in marketing analytics, namely in developing and analyzing marketing metrics that drive firm performance. His research has been published in such outlets as the *Journal of Marketing* *Marketing Science* *European Journal of Marketing* *Journal of Empirical Generalisations in Marketing* *Harvard Business Review* *Journal of Research in Marketing*, and *Marketing Science Institute Working Paper Series*

Acknowledgments

Of all my books, this one was probably the hardest to write. Thanks to my wonderful wife Vivian who was so nice to me when I got frustrated during the authoring process. Wiley acquisitions editor Robert Elliott was always encouraging and his input was a great help in shaping the final product. Wiley project editor Victoria Swider did a great job in pushing me to become a better writer.

Lastly, I must give a special note of thanks to my technical editor, Associate Professor of Marketing at the Kelly School of Business, Lopo Rego. Lopo did an amazing job of suggesting alternative wording and catching errors. He went above and beyond his role as technical editor, and I am truly indebted to him for his Herculean efforts.

Contents

Introduction	xxiii
I Using Excel to Summarize Marketing Data	1
1 Slicing and Dicing Marketing Data with PivotTables	3
2 Using Excel Charts to Summarize Marketing Data	29
3 Using Excel Functions to Summarize Marketing Data	59
II Pricing	83
4 Estimating Demand Curves and Using Solver to Optimize Price	85
5 Price Bundling	107
6 Nonlinear Pricing	123
7 Price Skimming and Sales	135
8 Revenue Management	143
III Forecasting	159
9 Simple Linear Regression and Correlation	161
10 Using Multiple Regression to Forecast Sales	177
11 Forecasting in the Presence of Special Events	213
12 Modeling Trend and Seasonality	225
13 Ratio to Moving Average Forecasting Method	235
14 Winter's Method	241
15 Using Neural Networks to Forecast Sales	249

IV	What do Customers Want?	261
16	Conjoint Analysis	263
17	Logistic Regression	285
18	Discrete Choice Analysis	303
V	Customer Value	325
19	Calculating Lifetime Customer Value	327
20	Using Customer Value to Value a Business	339
21	Customer Value, Monte Carlo Simulation, and Marketing Decision Making	347
	Allocating Marketing Resources between Customer Acquisition and Retention	365
VI	Market Segmentation	375
23	Cluster Analysis	377
24	Collaborative Filtering	393
25	Using Classification Trees for Segmentation	403
VII	Forecasting New Product Sales	413
26	Using S Curves to Forecast Sales of a New Product	415
27	The Bass Diffusion Model	427
28	Using the Copernican Principle to Predict Duration of Future Sales	439
VIII	Retailing	443
29	Market Basket Analysis and Lift	445

30	RFM Analysis and Optimizing Direct Mail Campaigns	459
31	Using the SCAN*PRO Model and Its Variants	471
32	Allocating Retail Space and Sales Resources	483
33	Forecasting Sales from Few Data Points	495
IX	Advertising	503
34	Measuring the Effectiveness of Advertising	505
35	Media Selection Models	517
36	Pay per Click (PPC) Online Advertising	529
X	Marketing Research Tools	539
37	Principal Components Analysis (PCA)	541
38	Multidimensional Scaling (MDS)	559
	Classification Algorithms: Naive Bayes Classifier and Discriminant Analysis	577
40	Analysis of Variance: One-way ANOVA	595
41	Analysis of Variance: Two-way ANOVA	607
XI	Internet and Social Marketing	619
42	Networks	621
43	The Mathematics Behind <i>The Tipping Point</i>	641
44	Viral Marketing	653
45	Text Mining	663
	Index	673



Contents

Introduction	xxiii
1 Using Excel to Summarize Marketing Data	1
Slicing and Dicing Marketing Data with PivotTables	3
Analyzing Sales at True Colors Hardware	3
Analyzing Sales at La Petit Bakery	14
Analyzing How Demographics Affect Sales.....	21
Pulling Data from a PivotTable with the GETPIVOTDATA Function	25
Summary	27
Exercises.....	27
2 Using Excel Charts to Summarize Marketing Data	29
Combination Charts	29
Using a PivotChart to Summarize Market Research Surveys	36
Ensuring Charts Update Automatically When New Data is Added	39
Making Chart Labels Dynamic	40
Summarizing Monthly Sales-Force Rankings	43
Using Check Boxes to Control Data in a Chart	45
Using Sparklines to Summarize Multiple Data Series	48
Using GETPIVOTDATA to Create the End-of-Week Sales Report	52
Summary	55
Exercises	55
3 Using Excel Functions to Summarize Marketing Data	59
Summarizing Data with a Histogram	59
Using Statistical Functions to Summarize Marketing Data	64
Summary	79
Exercises	80

II Pricing	83
Estimating Demand Curves and Using Solver to Optimize Price	85
Estimating Linear and Power Demand Curves	85
Using the Excel Solver to Optimize Price	90
Pricing Using Subjectively Estimated Demand Curves	96
Using SolverTable to Price Multiple Products	99
Summary	103
Exercises	104
5 Price Bundling	107
Why Bundle?	107
Using Evolutionary Solver to Find Optimal Bundle Prices	111
Summary	119
Exercises	119
6 Nonlinear Pricing	123
Demand Curves and Willingness to Pay	124
Profit Maximizing with Nonlinear Pricing Strategies	125
Summary	131
Exercises	132
7 Price Skimming and Sales	135
Dropping Prices Over Time	135
Why Have Sales?	138
Summary	142
Exercises	142
8 Revenue Management	143
Estimating Demand for the Bates Motel and Segmenting Customers	144
Handling Uncertainty	150
Markdown Pricing	153

Summary	156
Exercises..... 15	6
III Forecasting	159
9 Simple Linear Regression and Correlation	161
Simple Linear Regression	161
Using Correlations to Summarize Linear Relationships .	170
Summary	174
Exercises	175
10 Using Multiple Regression to Forecast Sales	177
Introducing Multiple Linear Regression	178
Running a Regression with the Data Analysis Add-In	179
Interpreting the Regression Output	182
Using Qualitative Independent Variables in Regression .	186
Modeling Interactions and Nonlinearities	192
Testing Validity of Regression Assumptions.....	195
Multicollinearity	204
Validation of a Regression	207
Summary	209
Exercises.....	210
11 Forecasting in the Presence of Special Events	213
Building the Basic Model	213
Summary	222
Exercises.....	222
12 Modeling Trend and Seasonality	225
Using Moving Averages to Smooth Data and Eliminate Seasonality	225
An Additive Model with Trends and Seasonality	228
A Multiplicative Model with Trend and Seasonality	231
Summary	234
Exercises	234

13	Ratio to Moving Average Forecasting Method	235
	Using the Ratio to Moving Average Method	235
	Applying the Ratio to Moving Average Method to Monthly Data	238
	Summary	238
	Exercises	239
14	Winter's Method	241
	Parameter Definitions for Winter's Method	241
	Initializing Winter's Method	243
	Estimating the Smoothing Constants	244
	Forecasting Future Months	246
	Mean Absolute Percentage Error (MAPE)	247
	Summary	248
	Exercises	248
15	Using Neural Networks to Forecast Sales	249
	Regression and Neural Nets	249
	Using Neural Networks	250
	Using NeuralTools to Predict Sales	253
	Using NeuralTools to Forecast Airline Miles	258
	Summary	259
	Exercises	259
IV	What do Customers Want?	261
16	Conjoint Analysis	263
	Products, Attributes, and Levels	263
	Full Profile Conjoint Analysis	265
	Using Evolutionary Solver to Generate Product Profiles	272
	Developing a Conjoint Simulator	277
	Examining Other Forms of Conjoint Analysis	279
	Summary	281
	Exercises	281

17	Logistic Regression	285
	Why Logistic Regression Is Necessary	286
	Logistic Regression Model	289
	Maximum Likelihood Estimate of Logistic Regression Model	290
	Using StatTools to Estimate and Test Logistic Regression Hypotheses	293
	Performing a Logistic Regression with Count Data	298
	Summary	300
	Exercises	300
18	Discrete Choice Analysis	303
	Random Utility Theory	303
	Discrete Choice Analysis of Chocolate Preferences	305
	Incorporating Price and Brand Equity into Discrete Choice Analysis	309
	Dynamic Discrete Choice	315
	Independence of Irrelevant Alternatives (IIA) Assumption	316
	Discrete Choice and Price Elasticity.....	317
	Summary	318
	Exercises.....	319
V	Customer Value	325
19	Calculating Lifetime Customer Value	327
	Basic Customer Value Template	328
	Measuring Sensitivity Analysis with Two-way Tables	330
	An Explicit Formula for the Multiplier	331
	Varying Margins	331
	DIRECTV, Customer Value, and <i>Friday Night Lights</i> (FNL)	333
	Estimating the Chance a Customer Is Still Active	334
	Going Beyond the Basic Customer Lifetime Value Model	335
	Summary	336
	Exercises.....	336

20	Using Customer Value to Value a Business	339
	A Primer on Valuation	339
	Using Customer Value to Value a Business	340
	Measuring Sensitivity Analysis with a One-way Table	343
	Using Customer Value to Estimate a Firm's Market Value	344
	Summary	344
	Exercises	345
21	Customer Value, Monte Carlo Simulation, and Marketing Decision Making	347
	A Markov Chain Model of Customer Value	347
	Using Monte Carlo Simulation to Predict Success of a Marketing Initiative	353
	Summary	359
	Exercises	360
22	Allocating Marketing Resources between Customer Acquisition and Retention	347
	Modeling the Relationship between Spending and Customer Acquisition and Retention	365
	Basic Model for Optimizing Retention and Acquisition Spending . .	368
	An Improvement in the Basic Model.....	371
	Summary	373
	Exercises.....	374
VI	Market Segmentation	375
23	Cluster Analysis	377
	Clustering U.S. Cities	378
	Using Conjoint Analysis to Segment a Market	386
	Summary	391
	Exercises.....	391
24	Collaborative Filtering	393
	User-Based Collaborative Filtering	393
	Item-Based Filtering	398

Comparing Item- and User-Based Collaborative Filtering	400
The Netflix Competition	40 1
Summary	401
Exercises.....	402
25 Using Classification Trees for Segmentation	403
Introducing Decision Trees	403
Constructing a Decision Tree	404
Pruning Trees and CART	409
Summary	410
Exercises.....	4 10
VII Forecasting New Product Sales	413
26 Using S Curves to Forecast Sales of a New Product	415
Examining S Curves	415
Fitting the Pearl or Logistic Curve	418
Fitting an S Curve with Seasonality	420
Fitting the Gompertz Curve	422
Pearl Curve versus Gompertz Curve	425
Summary	425
Exercises	425
27 The Bass Diffusion Model	427
Introducing the Bass Model	427
Estimating the Bass Model	428
Using the Bass Model to Forecast New Product Sales	431
Deflating Intentions Data	434
Using the Bass Model to Simulate Sales of a New Product	435
Modifications of the Bass Model	437
Summary	438
Exercises	438

Using the Copernican Principle to Predict Duration of Future Sales	439
Using the Copernican Principle	439
Simulating Remaining Life of Product	440
Summary	441
Exercises	441

VIII Retailing 443 29

Market Basket Analysis and Lift	445
Computing Lift for Two Products	445
Computing Three-Way Lifts	449
A Data Mining Legend Debunked!	453
Using Lift to Optimize Store Layout	454
Summary	456
Exercises	456

30 RFM Analysis and Optimizing Direct Mail Campaigns 459

RFM Analysis	459
An RFM Success Story	465
Using the Evolutionary Solver to Optimize a Direct Mail Campaign	465
Summary	468
Exercises.....	4 68

31 Using the SCAN*PRO Model and Its Variants 471

Introducing the SCAN*PRO Model	471
Modeling Sales of Snickers Bars	472
Forecasting Software Sales	475
Summary	480
Exercises.....	48 0

32 Allocating Retail Space and Sales Resources 483

Identifying the Sales to Marketing Effort Relationship.....	48 3
Modeling the Marketing Response to Sales Force Effort	484

Optimizing Allocation of Sales Effort	489
Using the Gompertz Curve to Allocate Supermarket Shelf Space	492
Summary	492
Exercises.....	49 3

33 Forecasting Sales from Few Data Points	495
Predicting Movie Revenues	495
Modifying the Model to Improve Forecast Accuracy	498
Using 3 Weeks of Revenue to Forecast Movie Revenues	499
Summary	501
Exercises.....	501

IX Advertising 503

34 Measuring the Effectiveness of Advertising	505
The Adstock Model	505
Another Model for Estimating Ad Effectiveness	509
Optimizing Advertising: Pulsing versus Continuous Spending. . . .	511
Summary	514
Exercises	515

35 Media Selection Models	517
A Linear Media Allocation Model	517
Quantity Discounts	520
A Monte Carlo Media Allocation Simulation	522
Summary	527
Exercises.....	527

36 Pay per Click (PPC) Online Advertising	529
Defining Pay per Click Advertising	529
Profitability Model for PPC Advertising	531
Google AdWords Auction	533
Using Bid Simulator to Optimize Your Bid	536
Summary	537
Exercises	537

X Marketing Research Tools	539
37 Principal Components Analysis (PCA)	541
Defining PCA	541
Linear Combinations, Variances, and Covariances	542
Diving into Principal Components Analysis	548
Other Applications of PCA	556
Summary	557
Exercises	558
38 Multidimensional Scaling (MDS)	559
Similarity Data	559
MDS Analysis of U.S. City Distances	560
MDS Analysis of Breakfast Foods	566
Finding a Consumer's Ideal Point	570
Summary	574
Exercises	574
Classification Algorithms: Naive Bayes Classifier and Discriminant Analysis	577
Conditional Probability	578
Bayes' Theorem	579
Naive Bayes Classifier	581
Linear Discriminant Analysis	586
Model Validation	591
The Surprising Virtues of Naive Bayes	592
Summary	592
Exercises	593
40 Analysis of Variance: One-way ANOVA	595
Testing Whether Group Means Are Different	595
Example of One-way ANOVA	596
The Role of Variance in ANOVA	598
Forecasting with One-way ANOVA	599

Contrasts	601
Summary	603
Exercises.....	604
41 Analysis of Variance: Two-way ANOVA	607
Introducing Two-way ANOVA	607
Two-way ANOVA without Replication 60	8
Two-way ANOVA with Replication	611
Summary	616
Exercises	617
XI Internet and Social Marketing	619
42 Networks	621
Measuring the Importance of a Node	621
Measuring the Importance of a Link	626
Summarizing Network Structure	628
Random and Regular Networks	631
The Rich Get Richer	634
Klout Score	636
Summary	637
Exercises.....	638
43 The Mathematics Behind <i>The Tipping Point</i>	641
Network Contagion	641
A Bass Version of the Tipping Point	646
Summary	650
Exercises	650
44 Viral Marketing	653
Watts' Model	654
A More Complex Viral Marketing Model	655
Summary	660
Exercises	661

45	Text Mining	663
	Text Mining Defi nitions	664
	Giving Structure to Unstructured Text	664
	Applying Text Mining in Real Life Scenarios	668
	Summary	671
	Exercises	671
	Index	673

Introduction

In the last 20 years, the use of analytic techniques in marketing has greatly increased. In April 2013, *Forbes* magazine reported a 67-percent growth in marketing-related analytics hires during the previous year and an amazing 136-percent growth during the previous 3 years.

Given this growth of interest in marketing analytics and my love of Excel modeling, I decided in 2004 to create a 7-week MBA elective in marketing analytics (K509) at the Indiana University Kelley School of Business. Although there are several excellent advanced marketing analytics books. (I am partial to *Database Marketing* by Robert Blattberg, Byung-Do Kim, and Scott Neslin (Springer, 2008).)

I could not find an Excel-based book that provided a how-to-do-it approach suitable for an MBA elective or an advanced undergraduate course. With no suitable book in hand, I wrote up course notes that I used in classes for 10 years. The course has been wildly successful with nearly 65 percent of all MBA's at the Kelley School taking the class. In May 2013, I was honored to receive the Eli Lilly MBA teaching award as the best teacher in the MBA program, primarily for teaching K509. In November 2011, Robert Elliott of Wiley Publishing approached me about turning my notes into a book, and this book is the result. In addition to being utilized in K509, portions of the book have been used to teach marketing analytics to senior managers at Deloitte consulting, Booz Allen Hamilton consulting, and 3M marketing analysts.

How This Book Is Organized

Since I started using Excel in classes in 1992, I have become a total convert to teaching by example. This book is no exception. Virtually every chapter's primary focus is to teach you the concepts through how-to examples. Each example has the following components:

- Step-by-step instructions
- A downloadable Excel file containing data and solutions
- Screenshots of various steps and sections of the Excel file for clarity

The downloadable Excel files provide complete solutions to the examples, but the instructions encourage you to follow along and work through them on your own. If you follow along using the provided Excel files, you can work in empty

cells alongside the completed solution and compare your result with the provided solution to ensure your success.

The book has been organized around 11 topical areas.

Part I: Using Excel to Summarize Marketing Data

This part of the book introduces the marketing analyst to many Excel tools that can be used to analyze marketing problems: PivotTables (Chapter 1), charting (Chapter 2), and Excel statistical functions (Chapter 3), including the incredibly useful COUNTIF, COUNTIFS, SUMIF, SUMIFS, AVERAGEIF, and AVERAGEIFS functions.

Part II: Pricing

The determination of a profit maximizing pricing strategy is always difficult. In this section you learn how to quickly estimate demand curves and use the Excel Solver (Chapter 4) to determine profit maximizing prices. The Excel Solver is then used to optimize price bundling (Chapter 5), nonlinear pricing strategies (Chapter 6), and price-skimming strategies (Chapter 7). A brief introduction to revenue management, also known as yield management (Chapter 8), is also included.

Part III: Forecasting

Businesses need accurate forecasts of future sales. Sales forecasts drive decisions involving production schedules, inventory management, manpower planning, and many other parts of the business. In this section you first learn about two of the most used forecasting tools: simple linear (Chapter 9) and simple multiple regression (Chapters 10 and 11). Then you learn how to estimate the trend and seasonal aspects of sales (Chapter 12) and generate forecasts using two common extrapolation forecasting methods: the Ratio to Moving Average method (Chapter 13), and Winter's Method for exponential smoothing (Chapter 14) with trend and seasonality. Then you learn about neural networks (Chapter 15), a form of artificial intelligence whose role in marketing forecasting is rapidly growing.

Part IV: What Do Customers Want?

Every brand manager wants to know how various product attributes drive the sales of a product. For example, what is most important in a consumer's choice of car: price, brand, engine horsepower, styling, or fuel economy? In this section you learn how conjoint analysis (Chapter 16) and discrete choice (Chapter 18) can be used to rank the importance of product attributes and also rank levels of product attributes. For example, what type of styling on an SUV is most preferred? You also

learn about the widely used tool of logistic regression (Chapter 17), which is used to estimate probabilities involving situations in which two, or binary, outcomes must be forecasted. For example, how a person's demographic information can be used to predict the chance that he will subscribe to a magazine.

Part V: Customer Value

Companies cannot make intelligent decisions on how to spend money acquiring customers unless they understand the value of their customers. After all, spending \$400 to acquire a customer who will generate \$300 in long-term profits is a sure recipe for going out of business. In this section you learn how to measure customer value (Chapter 19), value companies based on the customer value concept (Chapter 20), incorporate uncertainty in customer value models (Chapter 21), and use your understanding of customer value to optimally allocate resources (Chapter 22) between acquisition and retention of customers.

Part VI: Market Segmentation

No matter what product you sell, your market consists of different market segments. For example, in Chapter 23 you will use cluster analysis to show that every U.S. city can be classified into one of four demographic segments. In Chapter 25 you will learn how classification trees can be used to segment a market. You are also introduced to the exciting concepts behind collaborative filtering (Chapter 24), which is the basis for Amazon.com and Netflix recommendations.

Part VII: Forecasting New Product Sales

With little or no history about sales of a product, it is difficult to predict future product sales. Given a few data points, S curves (Chapter 26) can be used to predict future product sales. The famous Bass diffusion model (Chapter 27) explains how sales of products evolve over time and can be used to predict product sales even before a product comes to the market. The little-known Copernican Principle (Chapter 28) enables you to predict the remaining time for which a product will be sold.

Part VIII: Retailing

Analytic techniques can help retailers deal with many important issues. The concepts of market basket analysis and lift (Chapter 29) help retailers derive a store layout that maximizes sales from complementary products. Recency, frequency, and monetary value analysis (Chapter 30) helps direct mailers maximize profit from their mailings. The widely known SCAN*PRO (Chapter 31) model helps retailers

determine how factors such as seasonality, price, and promotions influence product sales. In Chapter 32 you learn how to use analytic techniques to determine optimal allocation of store space between products and also optimize the use of a corporate sales force. Finally in Chapter 33 you learn how to forecast total sales of a product from a few data points.

Part IX: Advertising

Department store owner John Wanamaker said, “Half the money I spend on advertising is wasted; the trouble is I don’t know which half.” In Chapter 34 you learn how John Wanamaker could have used the ADSTOCK model to measure the effectiveness of his advertising expenditures. In Chapter 35 you learn how to allocate ads between the available media outlets to maximize the effectiveness of ads. Chapter 36 deals with the math behind online ad auctions.

Part X: Marketing Research Tools

Often the marketing analyst must deal with data sets involving many variables. Principal components (Chapter 37) and Multidimensional Scaling (Chapter 38) enable the marketing analysts to reduce data sets involving many variables to a few easily understood variables. Often the marketing analyst must classify objects into one of several groups. Naïve Bayes and discriminant analysis (Chapter 39) are great tools for developing classification rules. When the marketing analyst wants to determine if a single factor or a pair of factors has a significant effect on product sales, ANOVA (Chapter 40 and Chapter 41) is a useful tool.

Part XI: The Internet and Social Marketing

In the last 20 years, the Internet has turned our world upside down, and marketing is no exception. Social media such as Facebook and Twitter create many interesting opportunities for the marketer, which require careful analysis. In Chapter 42 you learn how the theory of networks sheds light on how you can identify people who are the key to spreading the word about your product. Chapter 43 discusses the math behind Malcom Gladwell’s bestselling book *The Tipping Point* (Back Bay Books, 2002). Chapter 44 discusses the math behind videos (such as the notorious “Gangnam Style”) going viral. Finally, in Chapter 45 you learn how text mining can be used to glean useful insight from Twitter, blogs, and Facebook posts.

Who Should Read This Book

There **is plenty of material in this book for a one-semester** course on marketing analytics at the advanced undergraduate or MBA level. I also believe the book can be useful to any corporate marketing analyst. With regard to prerequisites for the book, I assume you understand the Copy command in Excel. That is, you know when and where to put dollars signs in a formula. If you work hard, that’s about all the prior knowledge needed to get a lot out of the book.

I always try to write my books in a modular fashion, so you can skip around and read about what interests you. If you don’t want to read the book from start to finish, the following table should help you navigate the book.

Chapters	Chapter Prerequisites
Chapter 1: Slicing and Dicing Marketing Data with PivotTables	None
Chapter 2: Using Excel Charts to Summarize Marketing Data	1
Chapter 3: Using Excel Functions to Summarize Marketing Data	2
Chapter 4: Estimating Demand Curves and Using Solver to Optimize Price	None
Chapter 5: Price Bundling	4
Chapter 6: Nonlinear Pricing	5
Chapter 7: Price Skimming and Sales	5
Chapter 8: Revenue Management	4
Chapter 9: Simple Linear Regression and Correlation	3
Chapter 10: Using Multiple Regression to Forecast Sales	9
Chapter 11: Forecasting in the Presence of Special Events	10
Chapter 12: Modeling Trend and Seasonality	5 and 11
Chapter 13: Ratio to Moving Average Forecasting Method	3 and 12
Chapter 14: Winter’s Method	12
Chapter 15: Using Neural Networks to Forecast Sales	10
Chapter 16: Conjoint Analysis	10

Continues

(continued)

Chapters	Chapter Prerequisites
Chapter 17: Logistic Regression	16
Chapter 18: Discrete Choice Analysis	17
Chapter 19: Calculating Lifetime Customer Value	3
Chapter 20: Using Customer Value to Value a Business	19
Chapter 21: Customer Value, Monte Carlo Simulation, and Marketing Decision Making	19
Chapter 22: Allocating Marketing Resources between Customer Acquisition and Retention	4 and 19
Chapter 23: Cluster Analysis	5
Chapter 24: Collaborative Filtering	23
Chapter 25: Using Classification Trees for Segmentation	24
Chapter 26: Using S Curves to Forecast Sales of a New Product	5 and 12
Chapter 27: The Bass Diffusion Model	26
Chapter 28: Using the Copernican Principle to Predict Duration of Future Sales	None
Chapter 29: Market Basket Analysis and Lift	19
Chapter 30: RFM Analysis and Optimizing Direct Mail Campaigns	29
Chapter 31: Using the SCAN*PRO Model and Its Variants	12
Chapter 32: Allocating Retail Space and Sales Resources	5
Chapter 33: Forecasting Sales from Few Data Points	31
Chapter 34: Measuring the Effectiveness of Advertising	31
Chapter 35: Media Selection Models	4, 21, and 34
Chapter 36: Pay Per Click (PPC) Online Advertising	None
Chapter 37: Principal Component Analysis (PCA)	10 and 23
Chapter 38: Multidimensional Scaling (MDS)	37
Chapter 39: Classification Algorithms: Naive Bayes Classifier and Discriminant Analysis	37 and 38

Chapters	Chapter Prerequisites
Chapter 40: Analysis of Variance: One-way ANOVA	10
Chapter 41: Analysis of Variance: Two-way ANOVA	40
Chapter 42: Networks	None
Chapter 43: The Mathematics Behind <i>The Tipping Point</i>	42
Chapter 44: Viral Marketing	10, 15, and 39
Chapter 45: Text Mining	3

For example, before reading Chapter 5 you need to have read Chapter 4; before reading Chapter 34 you need to have read Chapter 31, and so on.

Tools You Need

To work through the vast majority of the book, all you need is Excel 2007, 2010, or 2013. Chapters 15, 21, and 35 require use of the Palisade.com Decision Tools Suite. You can download a 15-day trial version of the suite from www.Palisade.com.

What's on the Website

From the book website (www.wiley.com/go/marketinganalytics) you can download all Excel files used in the book as well as answers to all of the Exercises at the end of each chapter.

Errata

We make every effort to ensure that there are no errors in the text or in the code. However, no one is perfect, and mistakes do occur. If you find an error in this book, like a spelling mistake or a calculation error, we would be very grateful for your feedback. By sending in errata you may save another reader hours of frustration and at the same time you will be helping us provide even higher quality information.

To submit errata for this book go to <http://support.wiley.com> and complete the form on the Ask a Question tab there to send us the error you have found. We'll check

Marketing Analytics

the information and, if appropriate, post a message to the book's errata page and fix the problem in subsequent editions of the book.

Summary

A **famous Chinese** proverb (popularized by the late management guru Stephen Covey) states, "If you give a man a fish you feed him for a day. If you teach a man to fish you feed him for a lifetime." Hopefully this book will teach you enough about marketing analytics so you will be well equipped to develop your own quantitative marketing models for most problems that come your way. Happy modeling!

Using Excel to Summarize Marketing Data

Chapter 1: Slicing and Dicing Marketing Data with
PivotTables

Chapter 2: Using Excel Charts to Summarize
Marketing Data

Chapter 3: Using Excel Functions to
Summarize Marketing Data



1

Slicing and Dicing Marketing Data with PivotTables

In many marketing situations you need to analyze, or “slice and dice,” your data to gain important marketing insights. Excel PivotTables enable you to quickly summarize and describe your data in many different ways. In this chapter you learn how to use PivotTables to perform the following:

- Examine sales volume and percentage by store, month and product type.

- Analyze the influence of weekday, seasonality, and the overall trend on sales at your favorite bakery.

- Investigate the effect of marketing promotions on sales at your favorite bakery.

- Determine the influence that demographics such as age, income, gender and geographic location have on the likelihood that a person will subscribe to *ESPN: The Magazine*

Analyzing Sales at True Colors Hardware

To start analyzing sales you first need some data to work with. The data worksheet from the `PARETO.xlsx` file (available for download on the companion website) contains sales data from two local hardware stores (uptown store owned by Billy Joel and downtown store owned by Petula Clark). Each store sells 10 types of tape, 10 types of adhesive, and 10 types of safety equipment. Figure 1-1 shows a sample of this data.

Throughout this section you will learn to analyze this data using Excel PivotTables to answer the following questions:

- What percentage of sales occurs at each store?

- What percentage of sales occurs during each month?

- How much revenue does each product generate?

- Which products generate 80 percent of the revenue?

Part I: Using Excel to Summarize Marketing Data

	Y	Z	AA	AB
7	Product	Month	Store	Price
8	Tape 10	April	downtown	\$2.50
9	Safety 8	August	uptown	\$10.00
10	Safety 2	February	uptown	\$10.00
11	Safety 8	November	uptown	\$10.00
12	Tape 10	October	uptown	\$2.50
13	Safety 8	January	uptown	\$10.00
14	Safety 8	December	downtown	\$10.00
15	Safety 1	September	downtown	\$12.00
16	Safety 2	May	uptown	\$10.00
17	Adhesive	July	uptown	\$7.00
18	Adhesive	March	uptown	\$7.00
19	Safety 8	August	downtown	\$10.00
20	Safety 8	October	downtown	\$10.00
21	Tape 10	July	downtown	\$2.50
22	Safety 2	February	downtown	\$10.00

Figure 1-1: Hardware store data

Calculating the Percentage of Sales at Each Store

The first step in creating a PivotTable is ensuring you have headings in the first row of your data. Notice that Row 7 of the example data in the data worksheet has the headings Product, Month, Store, and Price. Because these are in place, you can begin creating your PivotTable. To do so, perform the following steps:

Place your cursor anywhere in the data cells on the data worksheet, and then click PivotTable in the Tables group on the Insert tab. Excel opens the Create PivotTable dialog box, as shown in Figure 1-2, and correctly guesses that the data is included in the range Y7:AB1333.

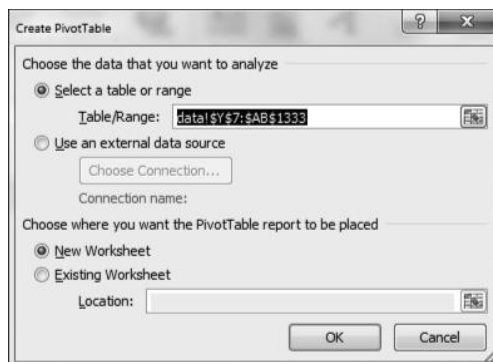


Figure 1-2: PivotTable Dialog Box

NOTE If you select Use an External Data Source here, you could also refer to a database as a source for a PivotTable. In Exercise 14 at the end of the chapter you can practice creating PivotTables from data in different worksheets or even different workbooks.

Click OK and you see the PivotTable Field List, as shown in Figure 1-3.

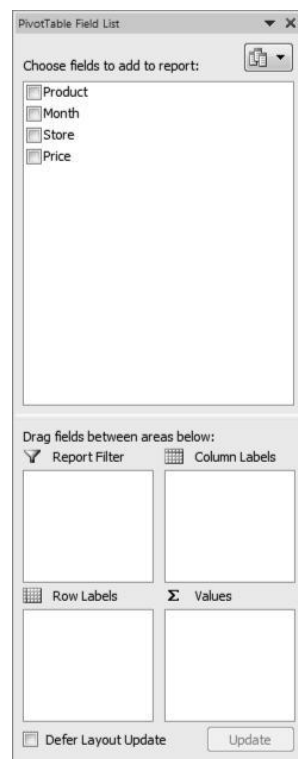


Figure 1-3: PivotTable Field List

Fill in the PivotTable Field List by dragging the PivotTable headings or *fields* into the boxes or *zones*. You can choose from the following four zones:

Row Labels: Fields dragged here are listed on the left side of the table in the order in which they are added to the box. In the current example, the Store field should be dragged to the Row Labels box so that data can be summarized by store.

Part I: Using Excel to Summarize Marketing Data

Column Labels: Fields dragged here have their values listed across the top row of the PivotTable. In the current example no fields exist in the Column Labels zone.

Values: Fields dragged here are summarized mathematically in the PivotTable. The Price field should be dragged to this zone. Excel tries to guess the type of calculation you want to perform on a field. In this example Excel guesses that you want all Prices to be summed. Because you want to compute total revenue, this is correct. If you want to change the method of calculation for a data field to an average, a count, or something else, simply double-click the data field or choose Value Field Settings. You learn how to use the Value Fields Setting command later in this section.

Report Filter: Beginning in Excel 2007, *Report Filter* is the new name for the Page Field area. For fields dragged to the Report Filter zone, you can easily pick any subset of the field values so that the PivotTable shows calculations based only on that subset. In Excel 2010 or Excel 2013 you can use the exciting *Slicers* to select the subset of fields used in PivotTable calculations. The use of the Report Filter and Slicers is shown in the “Report Filter and Slicers” section of this chapter.

NOTE To see the field list, you need to be in a field in the PivotTable. If you do not see the field list, right-click any cell in the PivotTable, and select Show Field List.

Figure 1-4 shows the completed PivotTable Field List and the resulting PivotTable is shown in Figure 1-5 as well as on the FirstStorePT worksheet.

Figure 1-5 shows the downtown store sold \$4,985.50 worth of goods, and the uptown store sold \$4,606.50 of goods. The total sales are \$9592.

If you want a percentage breakdown of the sales by store, you need to change the way Excel displays data in the Values zone. To do this, perform these steps:

- %4 Right-click in the summarized data in the FirstStorePT worksheet and select Value Field Settings.
- %4 Select Show Values As and click the drop-down arrow on the right side of the dialog box.
- %4 Select the % of Column Total option, as shown in Figure 1-6.

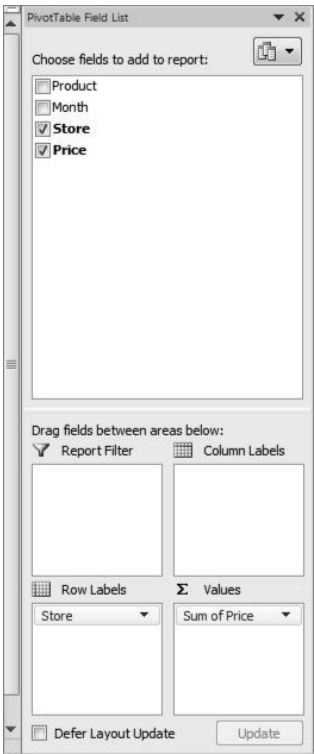


Figure 1-4: Completed PivotTable Field List

	A	B
3	Row Labels	Sum of Price
4	downtown	4985.5
5	uptown	4606.5
6	Grand Total	9592

Figure 1-5: Completed PivotTable

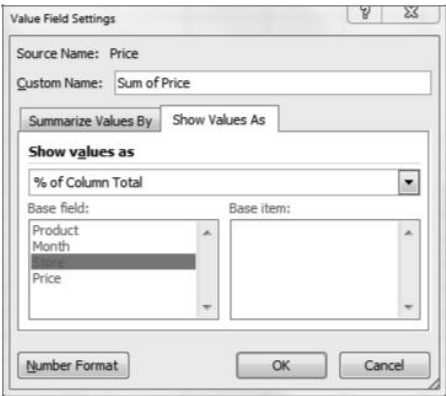


Figure 1-6: Obtaining percentage breakdown by Store

%4 Part I: Using Excel to Summarize Marketing Data

Figure 1-7 shows the resulting PivotTable with the new percentage break-down by Store with 52 percent of the sales in the downtown store and

%4 percent in the uptown store. You can also see this in the revenue by store worksheet of the PARETO.xlsx file.

NOTE If you want a PivotTable to incorporate a different set of data, then under Options, you can select Change Data Source and select the new source data. To have a PivotTable incorporate changes in the original source data, simply right-click and select Refresh. If you are going to add new data below the original data and you want the PivotTable to include the new data when you select Refresh, you should use the Excel Table feature discussed in Chapter 2, “Using Excel Charts to Summarize Marketing Data.”

	A	B
1		
2		
3	Row Labels	Sum of Price
4	downtown	51.98%
5	uptown	48.02%
6	Grand Total	100.00%

Figure 1-7: Percentage breakdown by Store

Summarizing Revenue by Month

You can also use a PivotTable to break down the total revenue by month and calculate the percentage of sales that occur during each month. To accomplish this, perform the following steps:

- %4 Return to the data worksheet and bring up the PivotTable Field List by choosing Insert PivotTable.
- %4 Drag the Month field to the Row Labels zone and the Price field to the Values zone. This gives the total sales by month. Because you also want a percentage breakdown of sales by month, drag the Price field again to the Values zone.
- %4 As shown in Figure 1-8, right-click on the first column in the Values zone and choose Value Field Settings; then choose the % of Column Total option. You now see the percentage monthly breakdown of revenue.
- %4 Double-click the Column headings and change them to Percentage of Sales by Month and Total Revenue.

%4 Finally, double-click again the Total Revenue Column; select Number Format, and choose the Currency option so the revenue is formatted in dollars.

	A	B	C
1			
2			
3	Row Labels	Percentage of Sales by Month	Total Revenue
4	January	8.81%	\$845.00
5	February	8.55%	\$820.00
6	March	6.58%	\$631.00
7	April	8.10%	\$776.50
8	May	8.48%	\$813.00
9	June	8.87%	\$850.50
10	July	8.77%	\$841.00
11	August	7.13%	\$684.00
12	September	8.71%	\$835.50
13	October	7.85%	\$753.00
14	November	9.11%	\$873.50
15	December	9.06%	\$869.00
16	Grand Total	100.00%	\$9,592.00

Figure 1-8: Monthly percentage breakdown of Revenue

You can see that \$845 worth of goods was sold in January and 8.81 percent of the sales were in January. Because the percentage of sales in each month is approximately 1/12 (8.33 percent), the stores exhibit little seasonality. Part III, “Forecasting Sales of Existing Products,” includes an extensive discussion of how to estimate seasonality and the importance of seasonality in marketing analytics.

Calculating Revenue for Each Product

Another important part of analyzing data includes determining the revenue generated by each product. To determine this for the example data, perform the following steps:

- %4 Return to the data worksheet and drag the Product field to the Row Labels zone and the Price field to the Values zone.
- %4 Double-click on the Price column, change the name of the Price column to Revenue, and then reformat the Revenue Column as Currency.
- %4 Click the drop-down arrow in cell A3 and select Sort A to Z so you can alpha-betize the product list and obtain the PivotTable in the products worksheet, as shown in Figure 1-9.

%4 Part I: Using Excel to Summarize Marketing Data

	A	B
3	Row Labels	Revenue
4	Adhesive 1	\$24.00
5	Adhesive 10	\$70.00
6	Adhesive 2	\$63.00
7	Adhesive 3	\$49.00
8	Adhesive 4	\$238.00
9	Adhesive 5	\$168.00
10	Adhesive 6	\$42.00
11	Adhesive 7	\$70.00
12	Adhesive 8	\$42.00
13	Adhesive 9	\$91.00
14	Safety 1	\$732.00
15	Safety 10	\$90.00
16	Safety 2	\$740.00
17	Safety 3	\$140.00
18	Safety 4	\$90.00
19	Safety 5	\$50.00
20	Safety 6	\$220.00
21	Safety 7	\$420.00
22	Safety 8	\$5,050.00
23	Safety 9	\$40.00
24	Tape 1	\$93.00
25	Tape 10	\$490.00
26	Tape 2	\$25.00
27	Tape 3	\$15.00
28	Tape 4	\$35.00
29	Tape 5	\$47.50
30	Tape 6	\$67.50
31	Tape 7	\$340.00
32	Tape 8	\$20.00
33	Tape 9	\$30.00
34	Grand Total	\$9,592.00

Figure 1-9: Sales by Product

You can now see the revenue that each product generated individually. For example, Adhesive 1 generated \$24 worth of revenue.

The Pareto 80–20 Principle

When slicing and dicing data you may encounter a situation in which you want to find which set of products generates a certain percentage of total sales. The well-known *Pareto 80/20 Principle* states that for most companies 20 percent of their products generate around 80 percent of their sales. Other examples of the Pareto Principle include the following:

- %4 Twenty percent of the population has 80 percent of income.
- %4 Of all possible problems customers can have, 20 percent of the problems cause 80 percent of all complaints.

To determine a percentage breakdown of sales by product, perform the following steps:

- %4 Begin with the PivotTable in the products worksheet and click the drop-down arrow in cell A3.

- %4 Select Value Filters; then choose Top 10...
- %4 Change the settings, as shown in Figure 1-10, to choose the products generating 80 percent of the revenue.

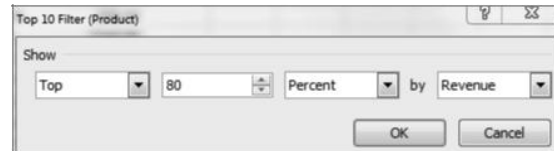


Figure 1-10: Using Value Filters to select products generating 80% of sales

The resulting PivotTable appears in the Top 80% worksheet (see Figure 1-11) and shows that the six products displayed in Figure 1-11 are the smallest set of products generating at least 80 percent of the revenue. Therefore only 20 percent of the products (6 out of 30) are needed to generate 80 percent of the sales.

NOTE By clicking the funnel you may clear our filters, if desired.

	A	B
3	Row Labels	Revenue
4	Safety 8	\$5,050.00
5	Safety 2	\$740.00
6	Safety 1	\$732.00
7	Tape 10	\$490.00
8	Safety 7	\$420.00
9	Tape 7	\$340.00
10	Grand Total	\$7,772.00

Figure 1-11: 6 Products Generate 80% of Revenue

The Report Filter and Slicers

One helpful tool for analyzing data is the Report Filter and the exciting Excel 2010 and 2013 Slicers Feature. Suppose you want to break down sales from the example data by month and store, but you feel showing the list of products in the Row or Column Labels zones would clutter the PivotTable. Instead, you can drag the Month field to the Row Labels zone, the Store field to the Column Labels zone, the Price field to the Value zone, and the Product field to the Report Filter zone. This yields the PivotTable in worksheet Report filter unfiltered, as shown in Figure 1-12.

%4 Part I: Using Excel to Summarize Marketing Data

	A	B	C	D
1	Product	(All)		
2				
3	Sum of Price	Column Labels		
4	Row Labels	downtown	uptown	Grand Total
5	January	\$482.00	\$363.00	\$845.00
6	February	\$363.00	\$457.00	\$820.00
7	March	\$299.00	\$332.00	\$631.00
8	April	\$407.00	\$369.50	\$776.50
9	May	\$408.50	\$404.50	\$813.00
10	June	\$400.50	\$450.00	\$850.50
11	July	\$446.00	\$395.00	\$841.00
12	August	\$372.50	\$311.50	\$684.00
13	September	\$446.50	\$389.00	\$835.50
14	October	\$394.00	\$359.00	\$753.00
15	November	\$503.50	\$370.00	\$873.50
16	December	\$463.00	\$406.00	\$869.00
17	Grand Total	4985.5	4606.5	9592

Figure 1-12: PivotTable used to illustrate Slicers

By clicking the drop-down arrow in the Report Filter, you can display the total revenue by Store and Month for any subset of products. For example, if you select products Safety 1, Safety 7, and Adhesive 8, you can obtain the PivotTable in the Filtered with a slicer worksheet, as shown in Figure 1-13. You see here that during May, sales of these products downtown are \$10.00 and uptown are \$34.00.

	A	B	C	D
1	Product	(Multiple Items)		
2				
3	Sum of Price	Column Labels		
4	Row Labels	downtown	uptown	Grand Total
5	January	\$46.00	\$22.00	\$68.00
6	February	\$44.00	\$68.00	\$112.00
7	March	\$51.00	\$63.00	\$114.00
8	April	\$44.00	\$44.00	\$88.00
9	May	\$10.00	\$34.00	\$44.00
10	June	\$89.00	\$56.00	\$145.00
11	July	\$49.00		\$49.00
12	August	\$10.00	\$36.00	\$46.00
13	September	\$88.00	\$44.00	\$132.00
14	October	\$102.00	\$87.00	\$189.00
15	November	\$56.00	\$70.00	\$126.00
16	December	\$32.00	\$49.00	\$81.00
17	Grand Total	621	573	1194

Figure 1-13: PivotTable showing sales for Safety 1, Safety 7 and Adhesive 8

As you can see from Figure 1-13, it is difficult to know which products were used in the PivotTable calculations. The new Slicer feature in Excel 2010 and 2013 (see the Filtered with a slicer worksheet) remedies this problem. To use this tool perform the following steps:

- %4 Put your cursor in the PivotTable in the Filtered with a Slicer worksheet and select Slicer from the Insert tab.
- %4 Select Products from the dialog box that appears and you see a Slicer that enables you to select any subset of products (select a product and then hold down the Control Key to select another product) from a single column.
- %4 Click inside the Slicer and you will see Slicer Tools on the ribbon. After selecting the Buttons section from Slicer Tools change Columns to 5. Now the products show up in five columns (see Figure 1-14).

Product				
Adhesive 1	Adhesive 10	Adhesive 2	Adhesive 3	Adhesive 4
Adhesive 5	Adhesive 6	Adhesive 7	Adhesive 8	Adhesive 9
Safety 1	Safety 10	Safety 2	Safety 3	Safety 4
Safety 5	Safety 6	Safety 7	Safety 8	Safety 9
Tape 1	Tape 10	Tape 2	Tape 3	Tape 4
Tape 5	Tape 6	Tape 7	Tape 8	Tape 9

Figure 1-14: Slicer selection for sales of Safety 1, Safety 7 and Adhesive 8.

A Slicer provides sort of a “dashboard” to filter on subsets of items drawn from a PivotTable field(s). The Slicer in the Filtered with a slicer worksheet makes it obvious that the calculations refer to Safety 1, Safety 7, and Adhesive 8. If you hold down the Control key, you can easily resize a Slicer.

NOTE If you double-click in a cell in a PivotTable, Excel drills down to the source data used for that cell’s calculations and places the source data in a separate sheet. For example, if in the Report filtered unfiltered worksheet you double-click in the January downtown cell, you can obtain the source data in the worksheet January downtown, as shown in Figure 1-15.

%4 Part I: Using Excel to Summarize Marketing Data

You have learned how PivotTables can be used to slice and dice sales data. Judicious use of the Value Fields Settings capability is often the key to performing the needed calculations.

	A	B	C	D
1	Product	Month	Store	Price
2	Safety 7	January	downtown	10
3	Safety 1	January	downtown	12
4	Safety 1	January	downtown	12
5	Safety 1	January	downtown	12

Figure 1-15: Drilling down on January Downtown Sales

Analyzing Sales at La Petit Bakery

La Petit Bakery sells cakes, pies, cookies, coffee, and smoothies. The owner has hired you to analyze factors affecting sales of these products. With a PivotTable and all the analysis skills you now have, you can quickly describe the important factors affecting sales. This example paves the way for a more detailed analysis in Part III of this book.

The file BakeryData.xlsx contains the data for this example and the file LaPetitBakery.xlsx contains all the completed PivotTables. In the Bakerydata.xlsx workbook you can see the underlying daily sales data recorded for the years 2013–2015. Figure 1-16 shows a subset of this data.

	A	B	C	D	E	F	G	H	I	J	K	L	M
4					daywk	weekday	Date	Cakes	Pies	Cookies	Smoothies	Coffee	promotion
5					Tuesday	2	1/1/2013	79	46	518		60	233 none
6		1 Monday			Wednesday	3	1/2/2013	91	50	539		161	427 none
7		2 Tuesday			Thursday	4	1/3/2013	47	60	222		166	347 none
8		3 Wednesday			Friday	5	1/4/2013	89	64	734		153	358 none
9		4 Thursday			Saturday	6	1/5/2013	112	73	764		240	392 none
10		5 Friday			Sunday	7	1/6/2013	89	57	922		259	510 none
11		6 Saturday			Monday	1	1/7/2013	70	50	476		120	334 none
12		7 Sunday			Tuesday	2	1/8/2013	70	48	496		222	316 none
13					Wednesday	3	1/9/2013	59	37	587		181	156 none
14					Thursday	4	1/10/2013	71	36	488		178	298 none
15					Friday	5	1/11/2013	74	50	645		100	490 none
16					Saturday	6	1/12/2013	119	71	438		162	416 none
17					Sunday	7	1/13/2013	90	51	568		137	434 none
18					Monday	1	1/14/2013	96	48	585		194	573 promotion

Figure 1-16: Data for La Petit Bakery

%4 1 in the weekday column indicates the day was Monday, a 2 indicates Tuesday, and so on. You can obtain these days of the week by entering the formula = WEEKDAY(G5,2) in cell F5 and copying this formula from F5 to the range

F6:F1099. The second argument of 2 in the formula ensures that a Monday is recorded as 1, a Tuesday as 2, and so on. In cell E5 you can enter the formula `=VLOOKUP(F5,lookday,2)` to transform the 1 in the weekday column to the actual word Monday, the 2 to Tuesday, and so on. The second argument lookday in the formula refers to the cell range A6:B12.

NOTE To name this range lookday simply select the range and type lookday in the Name box (the box directly to the left of the Function Wizard) and press Enter. Naming a range ensures that Excel knows to use the range lookday in any function or formula containing lookday

The VLOOKUP function finds the value in cell F5 (2) in the first column of the lookday range and replaces it with the value in the same row and second column of the lookday range (Tuesday.) Copying the formula `=VLOOKUP(F5,lookday,2)` from E5 to E6:E1099 gives you the day of the week for each observation. For example, on Friday, January 11, 2013, there was no promotion and 74 cakes, 50 pies, 645 cookies, 100 smoothies, and 490 cups of coffee were sold.

Now you will learn how to use PivotTables to summarize how La Petit Bakery's sales are affected by the following:

- %4 Day of the week
- %4 Month of the year
- %4 An upward (or downward!) trend over time in sales
- %4 Promotions such as price cuts

Summarizing the Effect of the Day of the Week on Bakery Sales

La Petit Bakery wants to know how sales of their products vary with the day of the week. This will help them better plan production of their products.

In the day of week worksheet you can create a PivotTable that summarizes the average daily number of each product sold on each day of the week (see Figure 1-17). To create this PivotTable, perform the following steps:

- %4 Drag the daywk field to the Row Labels zone and drag each product to the Values zone.
- %4 Double-click each product, and change the summary measure to Average. You'll see, for example, that an average of 96.5 cakes was sold on Sunday.

%4 Part I: Using Excel to Summarize Marketing Data

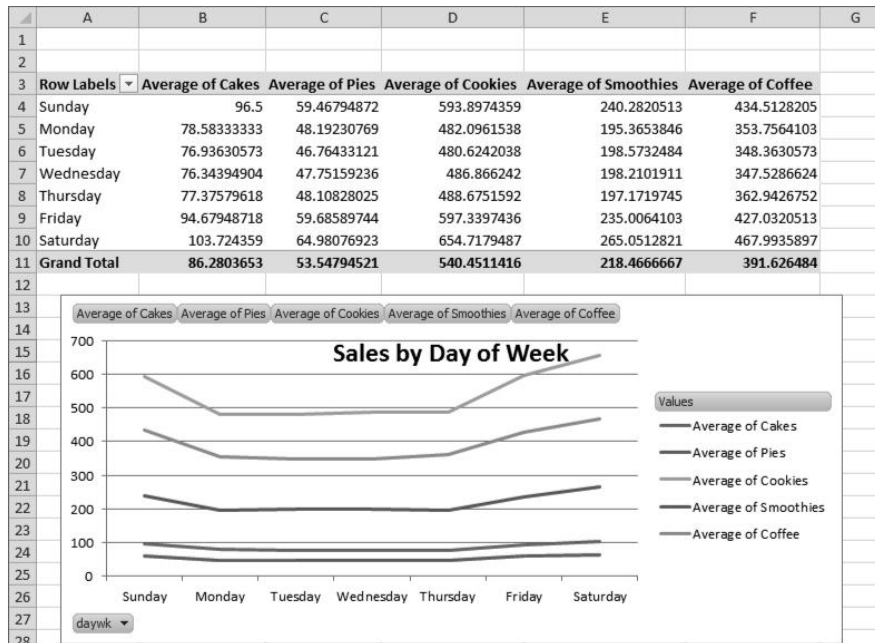


Figure 1-17: Daily breakdown of Product Sales

As the saying (originally attributed to Confucius) goes, “A picture is worth a thousand words.” If you click in a PivotTable and go up to the Options tab and select PivotChart, you can choose any of Excel’s chart options to summarize the data (Chapter 2, “Using Excel Charts to Summarize Marketing Data,” discusses Excel charting further). Figure 1-17 (see the Daily Breakdown worksheet) shows the first Line option chart type. To change this, right-click any series in a chart. The example chart here shows that all products sell more on the weekend than during the week. In the lower left corner of the chart, you can filter to show data for any subset of weekdays you want.

Analyzing Product Seasonality

If product sales are approximately the same during each month, they do not exhibit seasonality. If, however, product sales are noticeably higher (or lower) than average during certain quarters, the product exhibits seasonality. From a marketing standpoint, you must determine the presence and magnitude of seasonality to more efficiently plan advertising, promotions, and manufacturing decisions and investments. Some real-life illustrations of seasonality include the following:

- %4 Amazon's fourth quarter revenues are approximately 33 percent higher than an average quarter. This is because of a spike in sales during Christmas.
- %4 Tech companies such as Microsoft and Cisco invariably have higher sales during the last month of each quarter and reach maximum sales during the last month of the fiscal year. This is because the sales force doesn't get its bonuses unless it meets quarterly or end of year quotas.

To determine if La Petit Bakery products exhibit seasonality, you can perform the following steps:

- %4 Begin with your cursor anywhere in the data in the BakeryData.xlsx workbook. From the Insert tab select PivotTable and the PivotTable Field List will appear. Drag the Date field to the Row Labels zone and as before, drag each product to the Values zone and again change the entries in the Values zone to average sales for each product.
- %4 At first you see sales for every day, but you actually want to group sales by month. To do this, put the cursor on any date, right-click, and choose Group.
- %4 To group the daily sales into monthly buckets, choose Months from the dialog box, as shown in Figure 1-18.

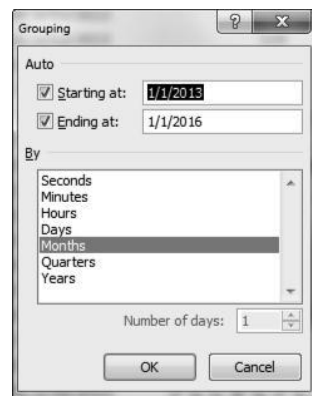


Figure 1-18: Grouping data by Month

- %4 Now select PivotChart from the Options tab. After selecting the first Line chart option you obtain the PivotChart and PivotTable, as shown in the monthly breakdown worksheet of the LaPetitBakery.xlsx file (see Figure 1-19).

%4 Part I: Using Excel to Summarize Marketing Data

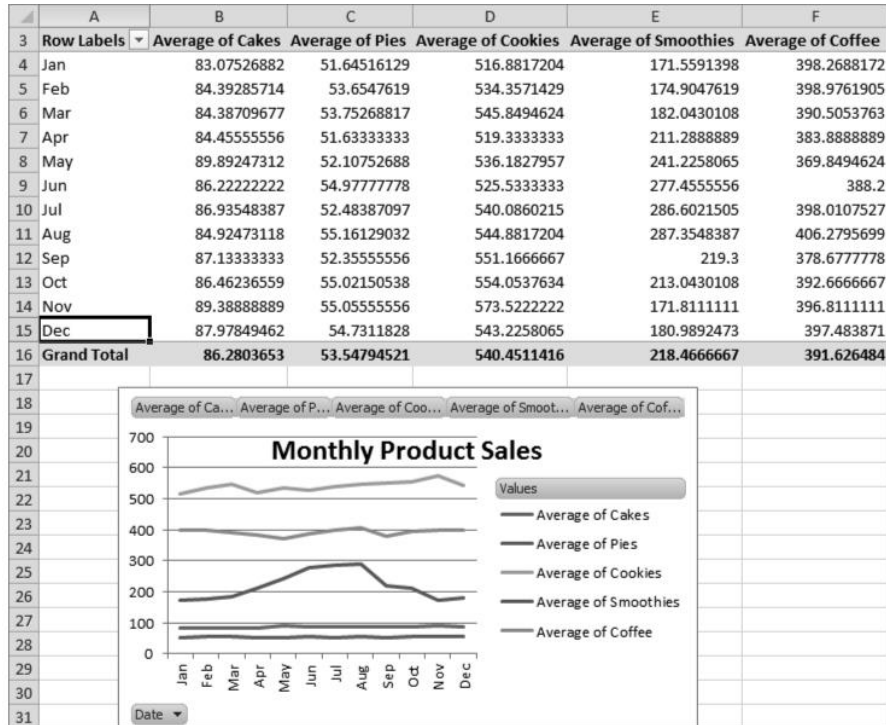


Figure 1-19: Monthly breakdown of Bakery Sales

This chart makes it clear that smoothie sales spike upward in the summer, but sales of other products exhibit little seasonality. In Part III of this book you can find an extensive discussion of how to estimate seasonality.

Given the strong seasonality and corresponding uptick in smoothie sales, the bakery can probably “trim” advertising and promotions expenditures for smoothies between April and August. On the other hand, to match the increased demand for smoothies, the bakery may want to guarantee the availability and delivery of the ingredients needed for making its smoothies. Similarly, if the increased demand places stress on the bakery’s capability to serve its customers, it may consider hiring extra workers during the summer months.

NOTE If you right-click any month and select Ungroup, you can undo the grouping and return to a display of the daily sales data.

Analyzing the Trend in Bakery Sales

The owners of La Petit Bakery want to know if sales are improving. Looking at a graph of each product's sales by month will not answer this question if seasonality is present. For example, Amazon.com has lower sales every January than the previous month due to Christmas. A better way to analyze this type of trend is to compute and chart average daily sales for each year. To perform this analysis, complete the following steps:

- %4 Put your cursor inside the data in the Data worksheet of the BakeryData.xlsx file, create a PivotTable, and drag each product to the Values zone and again change the method of summary from Sum to Average.
- %4 Then drag the Date field to the Row Labels zone, place the cursor on any date, and right-click Group and choose Years. You see a monthly summary of average daily sales for each Month and Year.
- %4 Drag the Date field away from the Row Labels zone and you are left with the summary of product sales by year, as shown in Figure 1-20 (see the work in worksheet by Year).
- %4 As you did before, create a Line PivotChart. The chart shows sales for each product are trending upward, which is good news for the client.

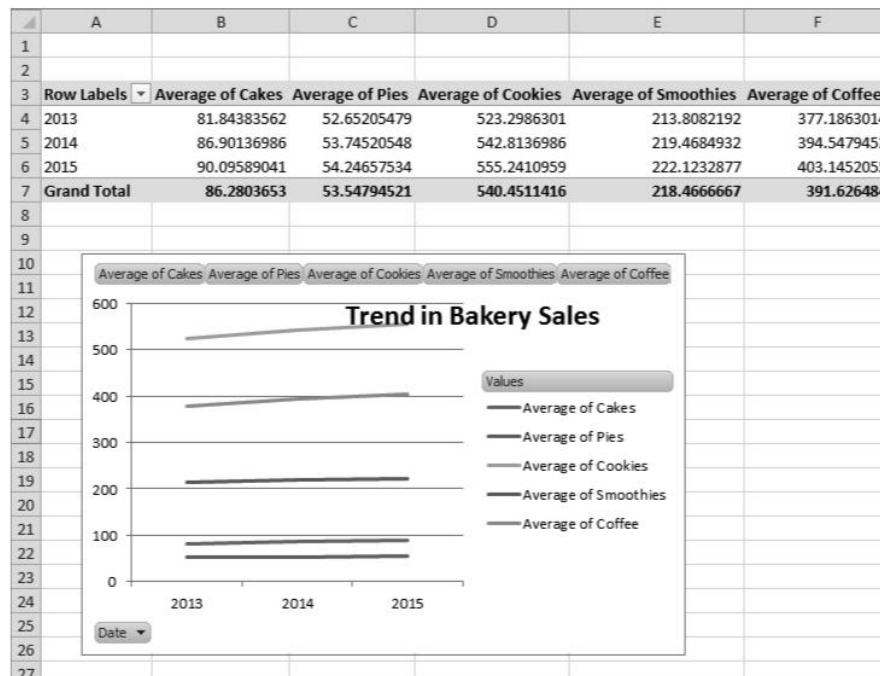


Figure 1-20: Summary of Sales by Year

%4 Part I: Using Excel to Summarize Marketing Data

The annual growth rates for products vary between 1.5 percent and 4.9 percent. Cake sales have grown at the fastest rate, but represent a small part of overall sales. Cookies and coffee sales have grown more slowly, but represent a much larger percentage of revenues.

Analyzing the Effect of Promotions on Sales

To get a quick idea of how promotions affect sales, you can determine average sales for each product on the days you have promotions and compare the average on days with promotions to the days without promotions. To perform these computations keep the same fields in the Value zone as before, and drag the Promotion field to the Row Labels zone. After creating a line PivotChart, you can obtain the results,

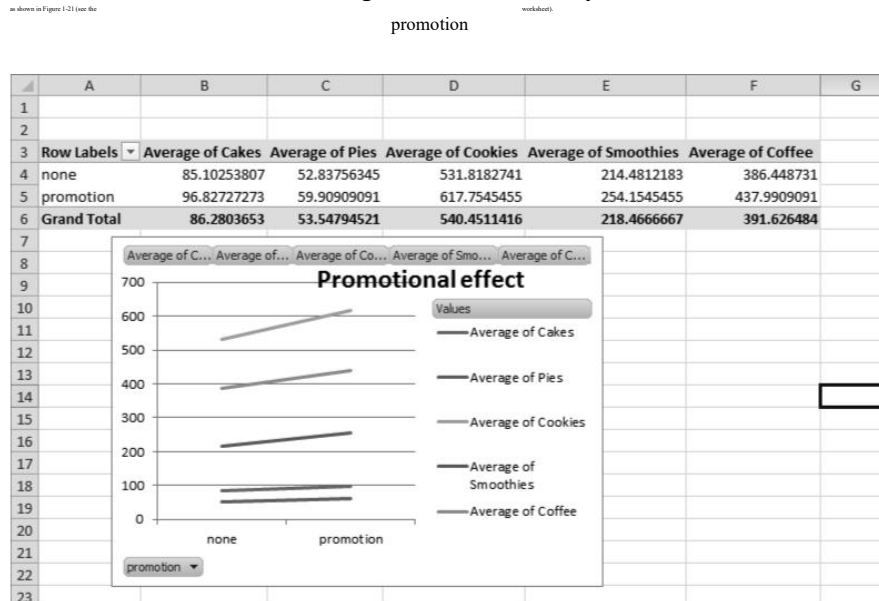


Figure 1-21: Effect of Promotion on Sales

The chart makes it clear that sales are higher when you have a promotion. Before concluding, however, that promotions increase sales, you must “adjust” for other factors that might affect sales. For example, if all smoothie promotions occurred during summer days, seasonality would make the average sales of smoothies on promotions higher than days without promotions, even if promotions had no real effect on sales. Additional considerations for using promotions include costs; if the costs of the promotions outweigh the benefits, then the promotion should not be undertaken. The

marketing analyst must be careful in computing the benefits of a promotion. If the promotion yields new customers, then the long-run value of the new customers (see Part V) should be included in the benefit of the promotion. In Parts VIII and IX of this book you can learn how to perform more rigorous analysis to determine how marketing decisions such as promotions, price changes, and advertising influence product sales.

Analyzing How Demographics Affect Sales

Before the marketing analyst recommends where to advertise a product (see Part IX), she needs to understand the type of person who is likely to purchase the product. For example, a heavy metal fan magazine is unlikely to appeal to retirees, so advertising this product on a television show that appeals to retirees (such as *Golden Girls*) would be an inefficient use of the advertising budget. In this section you will learn how to use PivotTables to describe the demographic of people who purchase a product.

Take a look at the data worksheet in the `espn.xlsx` file. This has demographic information on a randomly chosen sample of 1,024 subscribers to *ESPN: The Magazine*. Figure 1-22 shows a sample of this data. For example, the first listed subscriber is a 72-year-old male living in a rural location with an annual family income of \$72,000.

Analyzing the Age of Subscribers

One of the most useful pieces of demographic information is age. To describe the age of subscribers, you can perform the following steps:

- %4 Create a PivotTable by dragging the Age field to the Row Labels zone and the Age field to the Values zone.
- %4 Unfortunately, Excel assumes that you want to calculate the Sum of Ages. Double-click the Sum of Ages heading, and change this to Count of Age.
- %4 Use Value Field settings with the % of Column Total setting to show a per-centage breakdown by age.
- %4 Finally, right-click on any listed age in the PivotTable and select Group. This enables you to group ages in 10-year increments. You can also use the PivotChart feature using the first Column chart option to create a column

%4 Part I: Using Excel to Summarize Marketing Data

chart (see Figure 1-23 and the age worksheet) to summarize the age distribution of subscribers.

	G	H	I	J
3	Age	Gender	Income	Location
4	72	m	72	rural
5	29	m	68	suburban
6	33	m	57	suburban
7	25	m	62	suburban
8	38	m	164	urban
9	33	m	44	urban
10	18	f	62	rural
11	17	m	68	urban
12	32	f	53	urban
13	24	f	92	urban
14	26	f	54	suburban
15	40	m	88	urban
16	26	f	46	rural
17	21	m	83	suburban
18	29	m	144	rural
19	51	f	30	suburban
20	17	m	47	urban
21	29	m	61	urban
22	92	m	123	urban
23	20	m	76	suburban
24	20	f	57	rural
25	65	m	104	urban
26	26	m	46	urban
27	74	m	113	urban
28	37	f	109	rural

Figure 1-22: Demographic data for *ESPN: The Magazine*

You find that most of the magazine's subscribers are in the 18–37 age group. This knowledge can help ESPN find TV shows to advertise that target the right age demographic.

Analyzing the Gender of Subscribers

You can also analyze the gender demographics of *ESPN: The Magazine* subscribers. This will help the analyst to efficiently spend available ad dollars. After all, if all subscribers are male, you probably do not want to place an ad on *Project Runway*

%4 In the data worksheet, drag the Gender field to the Column Labels zone and the Gender field to the Values zone.

%4 Right-click the data and use Value Field Settings to change the calculations to Show Value As % of Row Total; this enables you to obtain the PivotTable shown in the gender worksheet (see Figure 1-24.)

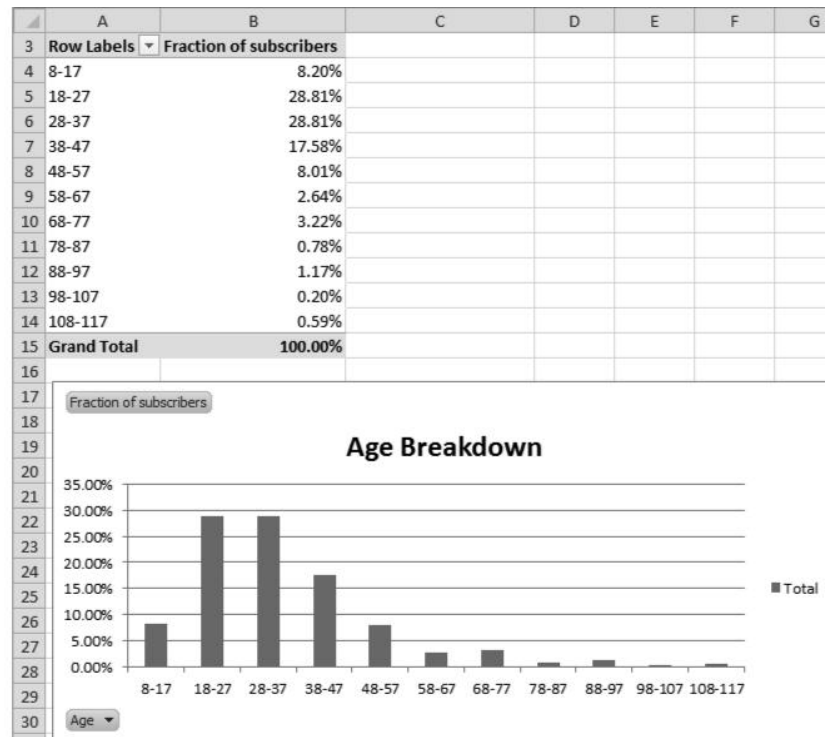


Figure 1-23: Age distribution of subscribers

	A	B	C	D
1				
2				
3	Column Labels			
4		f	m	Grand Total
5	Count of Gender	20.31%	79.69%	100.00%

Figure 1-24: Gender breakdown of subscribers

You find that approximately 80 percent of subscribers are men, so ESPN may not want to advertise on *The View* !

Describing the Income Distribution of Subscribers

In the Income worksheet (see Figure 1-25) you see a breakdown of the percentage of subscribers in each income group. This can be determined by performing the following steps:

- 4 In the data worksheet drag the Income field to the Row Labels zone and the Income field to the Values zone.

%4 Part I: Using Excel to Summarize Marketing Data

Change the Income field in the Values zone to Count of Income (if it isn't already there) and group incomes in \$25,000 increments.

Finally, use Value Field Settings > Show Value As > % of Column Total to display the percentage of subscribers in each income bracket.

You see that a majority of subscribers are in the \$54,000–\$103,000 range. In addition, more than 85 percent of *ESPN: The Magazine* subscribers have income levels well above the national median household income, making them an attractive audience for ESPN's additional marketing efforts.

	A	B
1		
2		
3	Row Labels	Count of Income
4	29-53	13.18%
5	54-78	35.06%
6	79-103	29.20%
7	104-128	14.55%
8	129-153	5.66%
9	154-178	1.56%
10	179-203	0.59%
11	229-253	0.10%
12	254-278	0.10%
13	Grand Total	100.00%

Figure 1-25: Income distribution of subscribers

Describing Subscriber Location

Next you will determine the breakdown of *ESPN: The Magazine* subscribers between suburbs, urban, and rural areas. This will help the analyst recommend the TV stations where ads should be placed.

%4 Put your cursor inside the data from the data worksheet and drag the Location field to the Column Labels zone and Value zone.

%4 Apply Value Field Settings and choose Show Value As > % of Row Total to obtain the PivotTable, as shown in the Location worksheet and Figure 1-26. You find that 46 percent of subscribers live in the suburbs; 40 percent live in urban areas and 15 percent live in rural areas.

	A	B	C	D	E
1					
2					
3		Column Labels			
4		rural	suburban	urban	Grand Total
5	Count of Location	14.55%	45.70%	39.75%	100.00%

Figure 1-26: Breakdown of Subscriber Location

Constructing a Crosstabs Analysis of Age and Income

Often marketers break down customer demographics simultaneously on two attributes. Such an analysis is called a *crosstabs* analysis. In the data worksheet you can perform a crosstabs analysis on Age and Income. To do so, perform the following steps:

- %4 Begin by dragging the Age field to the Column Labels zone and the Income field to the Row Labels and Value Labels zones.
- %4 Next group ages in 10-year increments and income in \$25,000 increments.
- %4 Finally, use Value Field Settings to change the method of calculation to Show Value As $\geq$ % of Row Total. The Income and Age worksheet (shown in Figure 1-27) shows the resulting PivotTable, which indicates that 28.13 percent of subscribers in the \$54,000 to \$78,000 bracket are in the 28–37 age group.

Crosstabs analyses enable companies to identify specific combinations of customer demographics that can be used for a more precise allocation of their marketing investments, such as advertising and promotions expenditures. Crosstabs analyses are also useful to determine where firms *should not* make investments. For instance, there are hardly any subscribers to *ESPN: The Magazine* that are 78 or older, or with household incomes above \$229,000, so placing ads on TV shows that are heavily watched by wealthy retiree is not recommended.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2													
3	Count of Income	Column Labels											
Row Labels	8-17	18-27	28-37	38-47	48-57	58-67	68-77	78-87	88-97	98-107	108-117	Grand Total	
29-53	11.11%	31.85%	25.19%	15.56%	8.15%	3.70%	0.74%	0.74%	0.74%	0.74%	1.48%	100.00%	
54-78	8.08%	27.86%	28.13%	17.27%	9.47%	2.51%	3.62%	1.11%	1.39%	0.28%	0.28%	100.00%	
79-103	7.69%	28.43%	29.43%	18.73%	7.02%	3.34%	3.01%	1.00%	1.00%	0.00%	0.33%	100.00%	
104-128	9.40%	25.50%	34.90%	15.44%	8.72%	1.34%	2.68%	0.00%	1.34%	0.00%	0.67%	100.00%	
129-153	1.72%	37.93%	29.31%	18.97%	3.45%	1.72%	6.90%	0.00%	0.00%	0.00%	0.00%	100.00%	
154-178	12.50%	37.50%	12.50%	25.00%	0.00%	0.00%	6.25%	0.00%	6.25%	0.00%	0.00%	100.00%	
179-203	0.00%	0.00%	0.00%	50.00%	16.67%	0.00%	16.67%	0.00%	0.00%	0.00%	16.67%	100.00%	
229-253	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	
254-278	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	
Grand Total	8.20%	28.81%	28.81%	17.58%	8.01%	2.64%	3.22%	0.78%	1.17%	0.20%	0.59%	100.00%	

Figure 1-27: Crosstabs Analysis of subscribers

Pulling Data from a PivotTable with the GETPIVOTDATA Function

Often a marketing analyst wants to pull data from a PivotTable to use as source information for a chart or other analyses. You can achieve this by using the GETPIVOTDATA function. To illustrate the use of the GETPIVOTDATA function, take a second look at

%4 Part I: Using Excel to Summarize Marketing Data

the True Colors hardware store data in the products worksheet from the PARETO.xlsx file.

%4 With the cursor in any blank cell, type an = sign and point to the cell (B12) containing Adhesive 8 sales.

%4 You will now see in the formerly blank cell the formula =GETPIVOTDATA("Price",\$A\$3,"Product","Adhesive 8"). Check your result against cell E10.

This pulls the sales of Adhesive 8 (\$42.00) from the PivotTable into cell E10.

This formula always picks out Adhesive 8 sales from the Price field in the PivotTable whose upper left corner is cell A3. Even if the set of products sold changes, this function still pulls Adhesive 8 sales.

In Excel 2010 or 2013 if you want to be able to click in a PivotTable and return a cell reference rather than a GETPIVOTTABLE function, simply choose File > Options, and from the Formulas dialog box uncheck the Use GetPivotData functions for the PivotTable References option (see Figure 1-28).

	A	B	C	D	E	F	G	H	I
1									
2									
3	Row Labels	Revenue							
4	Adhesive 1	\$24.00							
5	Adhesive 10	\$70.00							
6	Adhesive 2	\$63.00							
7	Adhesive 3	\$49.00							
8	Adhesive 4	\$238.00							
9	Adhesive 5	\$168.00							
10	Adhesive 6	\$42.00			42				
11	Adhesive 7	\$70.00			GETPIVOTDATA("Price",\$A\$3,"Product","Adhesive 8")				
12	Adhesive 8	\$42.00							
13	Adhesive 9	\$91.00							
14	Safety 1	\$732.00							
15	Safety 10	\$90.00							
16	Safety 2	\$740.00							
17	Safety 3	\$140.00							
18	Safety 4	\$90.00							
19	Safety 5	\$50.00							
20	Safety 6	\$220.00							
21	Safety 7	\$420.00							
22	Safety 8	\$5,050.00							
23	Safety 9	\$40.00							
24	Tape 1	\$93.00							
25	Tape 10	\$490.00							
26	Tape 2	\$25.00							
27	Tape 3	\$15.00							
28	Tape 4	\$35.00							
29	Tape 5	\$47.50							
30	Tape 6	\$67.50							
31	Tape 7	\$340.00							
32	Tape 8	\$20.00							
33	Tape 9	\$30.00							
34	Grand Total	\$9,592.00							

Figure 1-28: Example of GETPIVOTDATA function

This function is widely used in the corporate world and people who do not know it are at a severe disadvantage in making best use of PivotTables. Chapter 2 covers this topic in greater detail.

Summary

In this chapter you learned the following:

- %4 Sketch out in your mind how you want the PivotTable to look before you fill in the Field List.
- %4 Use Value Field Settings to change the way the data is displayed or the type of calculation (Sum, Average, Count, etc.) used for a Value Field.
- %4 A PivotChart can often clarify the meaning of a PivotTable.
- %4 Double-click in a cell to drill down to the source data that was used in the cell's calculation.
- %4 The GETPIVOTDATA function can be used to pull data from a PivotTable.

Exercises

%4 The `Makeup2007.xlsx` file (available for download on the companion website) gives sales data for a small makeup company. Each row lists the salesperson, product sold, location of the sale, units sold, and revenue generated. Use this file to perform the following exercises:

Summarize the total revenue and units sold by each person of each product.

Summarize the percentage of each person's sales that came from each location. Create a PivotChart to summarize this information.

Summarize each girl's sales by location and use the Report Filter to change the calculations to include any subset of products.

%4 The `Station.xlsx` file contains data for each family including the family size (large or small), income (high or low), and whether the family bought a station wagon. Use this file to perform the following exercises:

Does it appear that family size or income is a more important determinant of station wagon purchases?

Compute the percentage of station wagon purchasers that are high or low income.

%4 Part I: Using Excel to Summarize Marketing Data

Compute the fraction of station wagon purchasers that come from each of the following four categories: High Income Large Family, High Income Small Family, Low Income Large Family, and Low Income Small Family.

The `cranberrydata.xlsx` file contains data for each quarter in the years 2006–2011 detailing the pounds of cranberries sold by a small grocery store. You also see the store's price and the price charged by the major competitor. Use this file to perform the following exercises:

Ignoring price, create a chart that displays the seasonality of cranberry sales.

Ignoring price, create a chart that shows whether there is an upward trend in sales.

Determine average sales per quarter, breaking it down based on whether your price was higher or lower than the competitor's price.

The `tapedata.xlsx` file contains data for weeks during 2009–2011 for the unit sales of 3M tape, price charged, whether an ad campaign was run that week (1 = ad campaign), and whether the product was displayed on the end of the aisle (1 = end cap). Use this file to perform the following exercises:

Does there appear to be an upward trend in sales?

Analyze the nature of the monthly seasonality of tape sales.

Does an ad campaign appear to increase sales?

Does placing the tape in an end-cap display appear to increase sales?

The files `EAST.xlsx` and `WEST.xlsx` contain information on product sales (products A–H) that you sell in January, February, and March. You want to use a PivotTable to summarize this sales data. The Field List discussed in this chapter does not enable you to create PivotTables on data from multiple ranges. If you hold down the ALT key and let go and hold down the D key and let go of the D key and hold down the P key, you can see the Classic PivotTable Wizard that enables you to select multiple ranges of data to key a PivotTable. Let Excel create a single page field for you and create a PivotTable to summarize total sales in the East and West during each month. Use the Filters so that the PivotTable shows only January and March sales of products A, C, and E.

Using Excel Charts

2 to Summarize Marketing Data

An important component of using analytics in business is a push towards “visu-alization.” Marketing analysts often have to sift through mounds of data to draw important conclusions. Many times these conclusions are best represented in a chart or graph. As Confucius said, “a picture is worth a thousand words.” This chapter focuses on developing Excel charting skills to enhance your ability to effectively summarize and present marketing data. This chapter covers the following skills:

- %4 Using a combination chart and a secondary axis
- %4 Adding a product picture to your column graphs
- %4 Adding labels and data tables to your graphs
- %4 Using the Table feature to ensure your graphs automatically update when new data is added
- %4 Using PivotCharts to summarize a marketing survey
- %4 Making chart labels dynamic
- %4 Using Sparklines to summarize sales at different stores
- %4 Using Custom Icon Sets to summarize trends in sales force performance
- %4 Using a check box to control the data series that show in a graph
- %4 Using the Table feature and GETPIVOTDATA function to automate creation of end-of-month sales reports

Combination Charts

Companies often graph actual monthly sales and targeted monthly sales. If the marketing analyst plots both these series with straight lines it is difficult to differentiate between actual and target sales. For this reason analysts often summarize actual sales versus targeted sales via a combination chart in which the two series are displayed using different formats (such as a line and column graph.) This section explains how to create a similar combination chart.

%4 Part I: Using Excel to Summarizing Marketing Data

All work for this chapter is located in the file Chapter2charts.xlsx. In the work-sheet Combinations you see actual and target sales for the months January through July. To begin charting each month's actual and target sales, select the range F5:H12 and choose a line chart from the Insert tab. This yields a chart with both actual and target sales displayed as lines. This is the second chart shown in Figure 2-1; however, it is difficult to see the difference between the two lines here. You can avoid this by changing the format of one of the lines to a column graph. To do so, perform the following steps:

- %4 Select the Target sales series line in the line graph by moving the cursor to any point in this series.
- %4 Then right-click, and choose Change Series Chart Type...
- %4 Select the first Column option and you obtain the first chart shown in Figure 2-1.

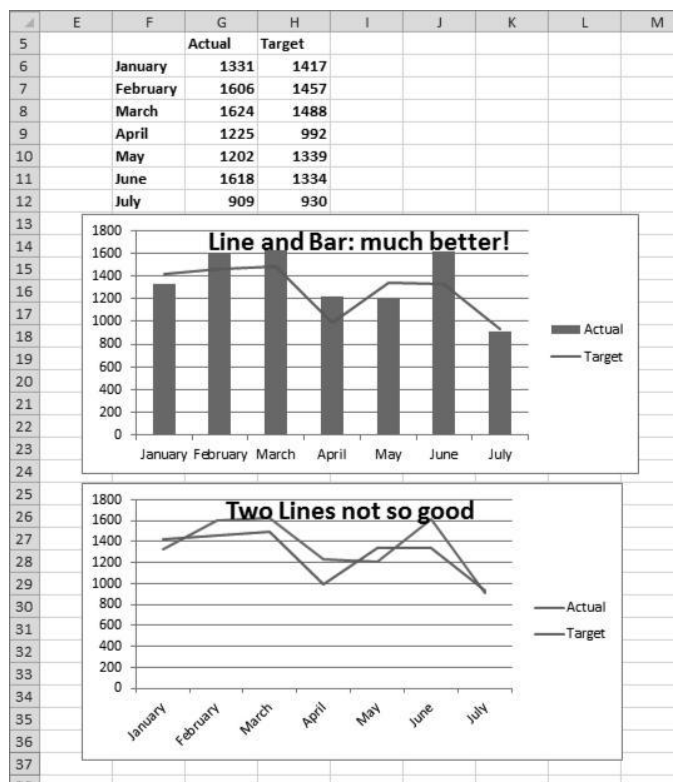


Figure 2-1: Combination chart

With the first chart, it is now easy to distinguish between the Actual and Target sales for each month. A chart like this with multiple types of graphs is called a *combination chart*.

You'll likely come across a situation in which you want to graph two series, but two series differ greatly in magnitude, so using the same vertical axis for each series results in one series being barely visible. In this case you need to include a secondary axis for one of the series. When you choose a secondary axis for a series Excel tries to scale the values of the axis to be consistent with the data values for that series. To illustrate the creation of a secondary axis, use the data in the Secondary Axis worksheet. This worksheet gives you monthly revenue and units sold for a company that sells expensive diamonds.

Because diamonds are expensive, naturally the monthly revenue is much larger than the units sold. This makes it difficult to see both revenue and units sold with only a single vertical axis. You can use a secondary axis here to more clearly summarize the monthly units sold and revenue earned. To do so, perform the following steps:

- %4 In the Secondary Axis worksheet select D7:F16, click the Insert tab, click the Line menu in the Charts section, and select the first Line chart option. You obtain the chart shown in Figure 2-2.

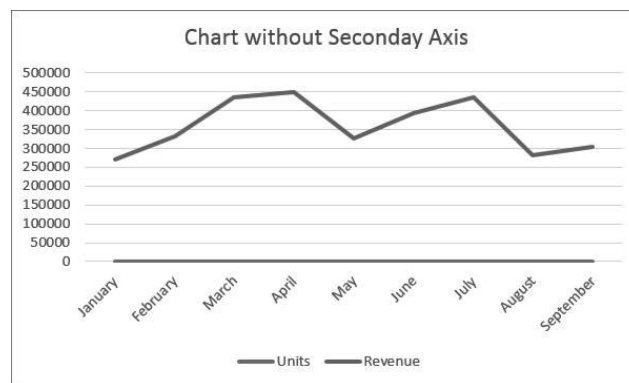


Figure 2-2: Line graph shows need for Secondary Axis

- %4 You cannot see the units' data though, because it is so small relative to the revenue. To remedy this problem, select the Revenue series line in the graph and right-click.

%4 Part I: Using Excel to Summarizing Marketing Data

Choose Format Data Series, and select Secondary Axis.

Then select any point in the Revenue series, right-click, and change the chart type for the Revenue series to a column graph.

The resulting chart is shown in Figure 2-3, and you can now easily see how closely units and revenue move together. This occurs because $\text{Revenue} = \text{Average price} \times \text{units sold}$, and if the average price during each month is constant, the revenue and units sold series will move in tandem. Your chart indeed indicates that average price is consistent for the charted data.

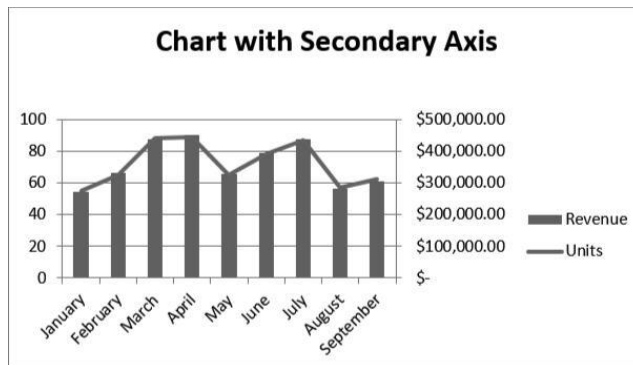


Figure 2-3: Chart with a secondary axis

Adding Bling to a Column Graph with a Picture of Your Product

While column graphs are quite useful for analyzing data, they tend to grow dull day after day. Every once in a while, perhaps in a big presentation to a potential client, you might want to spice up a column graph here and there. For instance, in a column graph of your company's sales you could add a picture of your company's product to represent sales. Therefore if you sell Ferraris, you could use a .bmp image of a Ferrari in place of a bar to represent sales. To do this, perform the following steps:

- %4 Create a Column graph using the data in the Picture worksheet. This data shows monthly sales of cars at an L.A. Ferrari dealer.

- %4 After creating a Column graph, right-click on any column and choose Format Data Series... followed by Fill.
- %4 Click Picture or texture fill, as shown in Figure 2-4.
- %4 Next click File below the Insert from query and choose the Ferrari .bmp file (available for download from the companion site).
- %4 Choose Stack and Scale with **200** units per picture. This ensures each car represents sales of 200 cars.

The resulting chart is shown in Figure 2-5, as well as in the Picture worksheet.

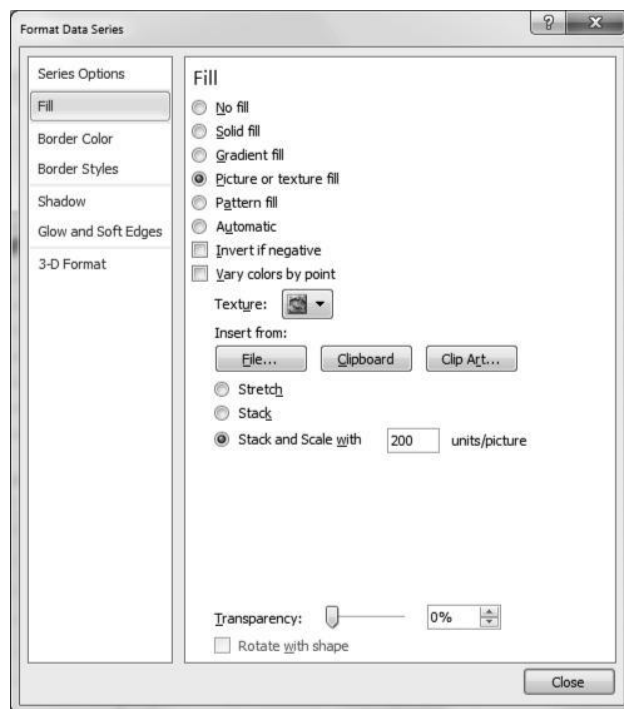


Figure 2-4: Dialog box for creating picture graph

NOTE If you choose Stretch in the Fill tab of the Format Data Series dialog, you can ensure that Excel represents each month's sales by a single car whose size is proportional to that month's sales.

%4 Part I: Using Excel to Summarizing Marketing Data

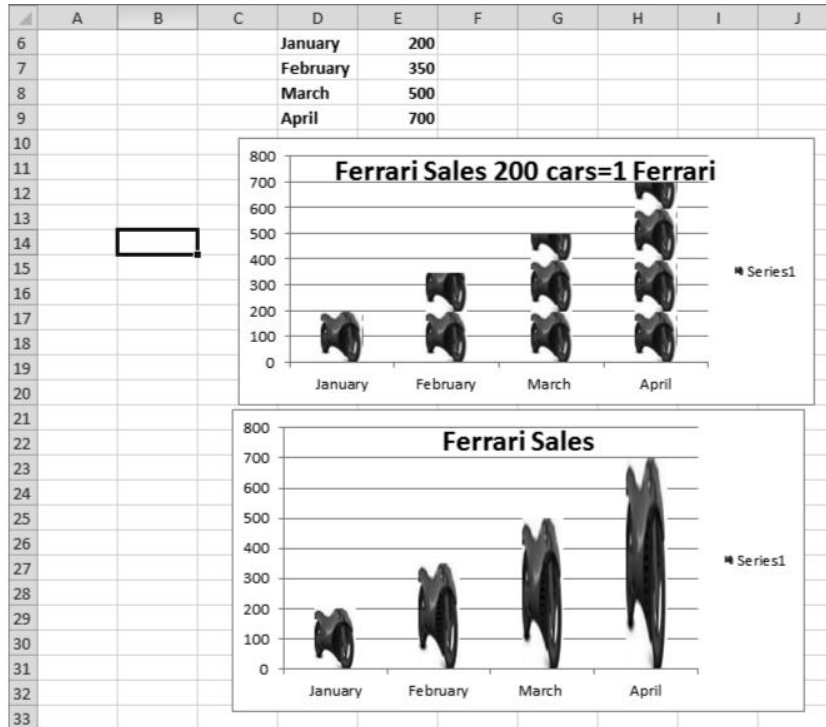


Figure 2-5: Ferrari picture graph

Adding Labels or Tables to Your Charts

Often you want to add data labels or a table to a graph to indicate the numerical values being graphed. You can learn to do this by using example data that shows monthly sales of product lines at the True Color Department Store. Refer to the Labels and Tables worksheet for this section's examples.

To begin adding labels to a graph, perform the following steps:

- %4 Select the data range C5:D9 and choose the first Column chart option (Clustered Column) from the Charts section of the Insert tab.
- %4 Now click on any column in the graph and choose Layout from the Chart Tools section of the ribbon.

%4 Choose the Data Labels option from the Layout section of Chart Tools and select More Data Label Options... Fill in the dialog boxes as shown in Figure 2-6.

%4 Include the Value and Category Name in the label, and put them on a different line.

The resulting graph with data labels is shown in Figure 2-7.

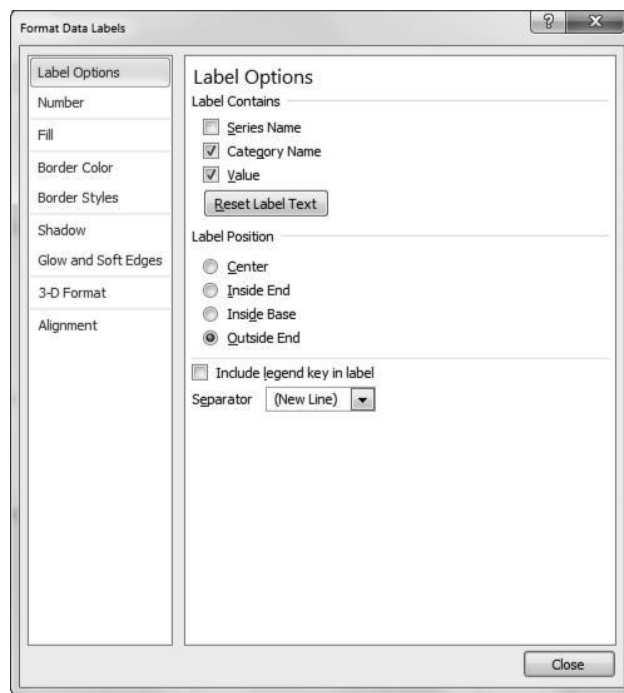


Figure 2-6: Dialog Box for creating chart with data labels

You can also add a Data Table to a column graph. To do so, again select any of the columns within the graph and perform the following steps:

%4 From the Chart Tools tab, select Layout.

%4 Choose Data Table.

%4 Choose the Show Data Table option to create the table shown in the second chart in Figure 2-7.

%4 Part I: Using Excel to Summarizing Marketing Data



Figure 2-7: Chart with data labels or data tables

Using a PivotChart to Summarize Market Research Surveys

To **determine a new product's viability**, market researchers often ask potential customers questions about a new product. PivotTables (discussed in Chapter 1, "Slicing and Dicing Marketing Data with PivotTables") are typically used to summarize this type of data. A PivotChart is a chart based on the contents of a PivotTable. As you will see, a PivotChart can be a highly effective tool to summarize any data gathered from this type of market research. The Survey PivotChart worksheet in the Chapter2chart.xlsx file uses a PivotChart to summarize a market research survey based on the data from the Survey Data worksheet (see Figure 2-8). The example data shows the answer to seven questions about a product (such as Likely to Buy) that was recorded on a 1–5 scale, with a higher score indicating a more favorable view of the product.

	I	J	K
5	Response	Question	
6	5	Likely To Buy	
7	3	Easy to cook	
8	3	Attractive packaging	
9	2	Too expensive	
10	5	Easy to cook	
11	1	Easy to cook	
12	3	Too expensive	
13	1	Likely To Buy	
14	2	Likely To Buy	
15	3	Better tasting	
16	2	Too expensive	
17	5	Too expensive	
18	3	Nutritious meal	
19	2	Recommend to friend	
20	2	Nutritious meal	
21	4	Likely To Buy	
22	2	Likely To Buy	
23	5	Attractive packaging	
24	4	Likely To Buy	
25	2	Likely To Buy	
26	5	Easy to cook	
27	5	Attractive packaging	
28	3	Too expensive	
29	4	Recommend to friend	
30	4	Likely To Buy	

Figure 2-8: Data for PivotChart example

To summarize this data, you can perform the following steps:

- %4 Select the data from the Survey Data worksheet and from the Tables section of the Insert tab choose PivotTable > PivotTable and click OK.
- %4 Next drag the Question field heading to the Row Labels zone and the Response field heading to the Column Labels and Values zones.
- %4 Assume you want to chart the fraction of 1, 2, 3, 4, and 5 responses for each question. Therefore, use Value Field Settings to change the Summarize Values By tab field to Count and change the Show Values As tab field to % of Row Total. You then obtain the PivotTable shown first in Figure 2-9.
- %4 You can now create a Pivot Chart from this table. Select PivotChart from the Options tab and choose the first Column chart option. This process yields the PivotChart shown second in Figure 2-9.
- %4 This chart is a bit cluttered however, so click the drop-down arrow by the Question button, and choose to display only responses to Likely to Buy and Recommend to a Friend. The resulting uncluttered chart is shown in Figure 2-10.

%4 Part I: Using Excel to Summarizing Marketing Data

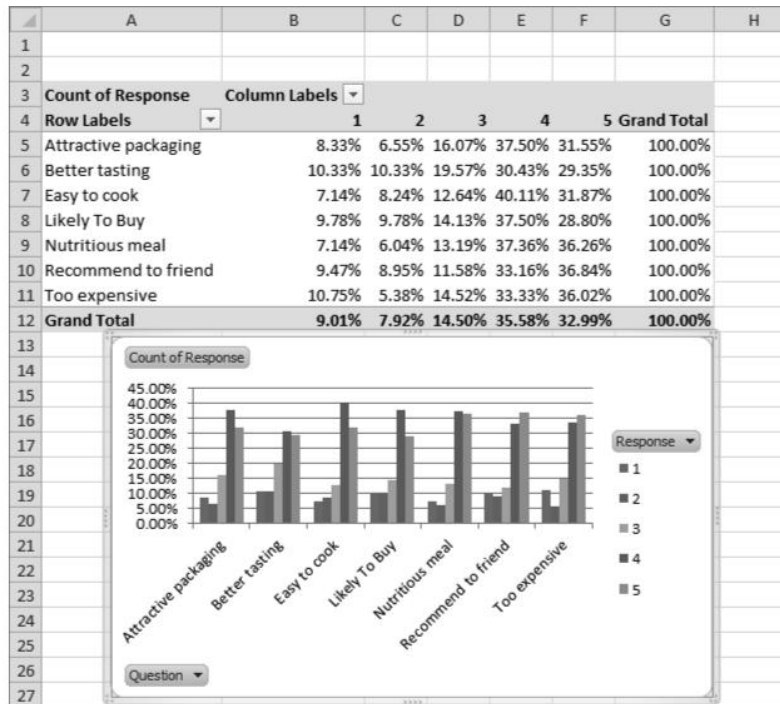


Figure 2-9: PivotTable and cluttered PivotChart

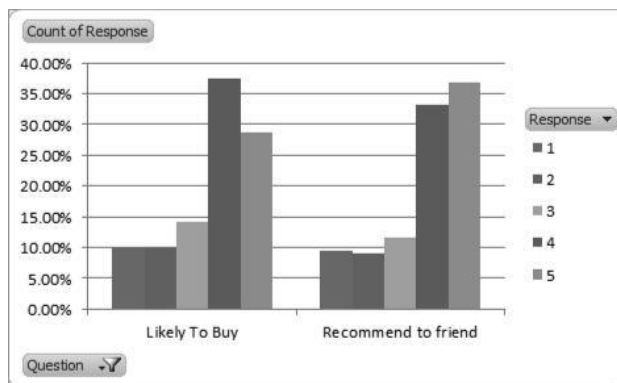


Figure 2-10: Uncluttered PivotChart

You can see that for both questions more than 60 percent of the potential customers gave the new product a 4 or 5, which is quite encouraging.

Ensuring Charts Update Automatically When New Data is Added

Most **marketing analysts** download sales data at the end of each month and individually update a slew of graphs that include the new data. If you have Excel 2007 or later, an easy way to do this is to use the TABLE feature to cause graphs to automatically update to include new data. To illustrate the idea, consider the Housing starts worksheet, which contains U.S. monthly housing starts (in 000s) for the time period January 2000 through May 2011. From the Insert tab, you can create the X-Y scatter chart, as shown in Figure 2-11. Just click anywhere in the data and then choose the first scatter plot option.



Figure 2-11: Housing chart through May 2011

If you add new data in columns D and E as it currently stands, the chart does not automatically incorporate the new data. To ensure that new data is automatically charted, you can instead make your data a TABLE before you create the chart. To do this, simply select the data (including headers) and press Ctrl+T. You can try this for yourself in the New Data worksheet by following these steps:

- %4 Select the data (D2:E141) and press Ctrl+T to make it a table.
- %4 Insert a scatter chart.
- %4 Add two new data points for June 2011 and July 2011 in Row 142 and Row 143.
- %4 The chart (as shown in Figure 2-12) automatically updates to include the new data. The difference between the two charts is apparent because June and July 2011 housing starts are much larger than April and May 2011 housing

40 Part I: Using Excel to Summarizing Marketing Data

starts. Imagine the time-savings if you have 15 or 20 charts keying off this data series!



Figure 2-12: Housing chart through July 2011

You can use the Ctrl+T trick for PivotTables too. If you use data to create a PivotTable, you should make the data a TABLE before creating the PivotTable. This ensures that when you right-click and select Refresh (to update a PivotTable) the PivotTable automatically incorporates the new data.

Making Chart Labels Dynamic

Data in **your charts comes from data in your spreadsheet**. Therefore changes to your spreadsheet will often alter your chart. If you do not set up the chart labels to also change, your chart will certainly confuse the user. This section shows how to make chart labels dynamically update to reflect changes in the chart.

Suppose you have been asked to project sales of a new product. If sales in a year are at least 2,000 units, the company will break even. There are two inputs in the Dynamic Labels worksheet: Year 1 sales (in cell D2) and annual growth in sales (in cell D3). The break-even target is given in cell D1. Use the Excel Create from Selection feature to name the data in D1:D3 with the names in the cell range C1:C3. To accomplish this, perform the following steps:

- %4 Select the cell range D1:D3.
- %4 Under the Formulas tab in the Defined Names section, select Create from Selection.

%4 Choose the Left Column. Now, for example, when you use the name `Year1sales` in a formula, Excel knows you are referring to D2.

NOTE A important reason for using range names is the fact that you can use the F3 key to Paste Range Names in formulas. This can save you a lot of time moving back and forth in your spreadsheet. For example, if your data is in Column A and you have named it you can use the F3 key to quickly and easily paste the data into a formula that is far away from Column A (for example, Column HZ.)

You can continue by graphing the annual sales that are based on initial sales and annual sales growth. As these assumptions change, the graph title and vertical axis legend need to reflect those changes. To accomplish this goal proceed as follows:

%4 Enter the Year 1 sales in cell F5 with the formula `=Year1sales` . Copy the formula `=F5*(1+anngrowth)` from F6 to F7:F15 to generate sales in Years 2–11.

%4 You want to find the year in which you break even, so copy the formula `=IF(F5>=target,"yes","no")` from G5 to G6:G15 to determine if you have broken even during the current year. For example, you can see that during Years 5–11 you broke even.

The key to creating dynamic labels that depend on your assumed initial sales and annual sales growth is to base your chart title and series legend key off formulas that update as your spreadsheet changes.

Now you need to determine the first year in which you broke even. This requires the use of the Excel MATCH function. The MATCH function has the following syntax: `MATCH(lookup_value,lookup_range,0)` . The *lookup_range* must be a row or column, and the MATCH function will return the position of the first occurrence of the *lookup_value* in the *lookup_range* . If the *lookup_value* does not occur in the *lookup_range*, the MATCH function returns an #N/A error. The last argument of 0 is necessary for the MATCH function to work properly.

In cell E19 the formula `=MATCH("yes",G5:G15,0)` returns the first year (in this case 5) in which you meet the breakeven target. In cell D22 several Excel functions are used to create the text string that will be the dynamic chart title. The formula entered in cell D22 is as follows:

`=IFERROR("We will break even in year "&TEXT(E19,"0"),"Never break even")`

%4 Part I: Using Excel to Summarizing Marketing Data

The functions used in this formula include the following:

- %4 IFERROR evaluates the formula before the comma. If the formula does not evaluate to an error, IFERROR returns the formula's result. If the formula evaluates to an error, IFERROR returns what comes after the comma.
- %4 The & or concatenate sign combines what is before the & sign with what comes after the & sign.
- %4 The TEXT function converts the contents of a cell to text in the desired format. The two most commonly used formats are "0", which indicates an integer format, and "0.0%", which indicates a percentage format.

If the MATCH formula in cell E19 does not return an error, the formula in cell D22 returns the text string We Will Break Even in Year followed by the first year you break even. If you never make it to break even, the formula returns the phrase Never Break Even. This is exactly the chart title you need.

You also need to create a title for the sales data that includes the annual sales growth rate. The series title is created in cell D23 with the following formula:

```
= "Sales (growth rate=" & TEXT(anngrowth,"0.0") & "%")"
```

NOTE Note that the growth rate has been changed to a single decimal.

Now you are ready to create the chart with dynamic labels. To do so, perform the following steps:

- %4 Select the source data (cell range E4:F15 in the Dynamic Labels worksheet) from the chart.
- %4 Navigate to the Insert tab and choose an X-Y Scatter Chart (the second option).
- %4 Go to the Layout section of the Chart Tools Group and select Chart Title and Centered Overlay Chart.
- %4 Place the cursor in the formula bar and type an equal sign (=), point to the chart title in D22, and press Enter. You now have a dynamic chart title that depends on your sales assumptions.
- %4 To create the series title, right-click any plotted data point, and choose Select Data.
- %4 Click Edit and choose Series name.

%4 Type an equals sign, point to the title in D23, and press Enter. You now have a dynamic series label. Figure 2-13 shows the resulting chart.

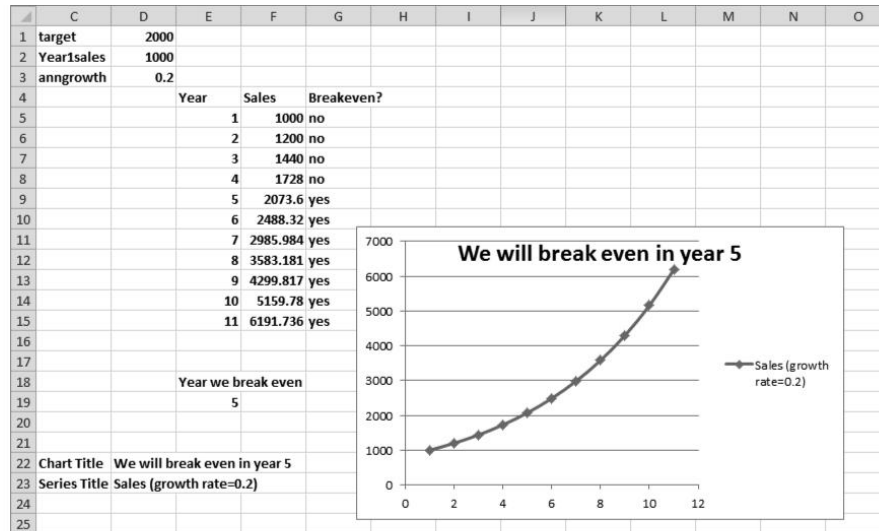


Figure 2-13: Chart with dynamic labels

Summarizing Monthly Sales-Force Rankings

If you **manage a sales force**, you need to determine if a salesperson's **performance** is improving or declining. Creating customized icon sets provides an easy way to track a salesperson's performance over time.

The data in the Sales Tracker worksheet list sales of salespeople during each month (see Figure 2-14). You can track each month with icons (up arrow, down arrow, or flat arrow) to determine whether a salesperson's ranking has improved, worsened, or stayed the same. You could use Excel's Conditional Formatting icon sets (see Chapter 24 of Winston's *Data Analysis and Business Modeling with Excel 2010* for a description of the Conditional Formatting icon sets), but then you would need to insert a set of icons separately for each column. A more efficient (although not as pretty) way to create icons is to use Wingdings 3 font and choose an "h" for an up arrow, an "i" for a down arrow, and a "g" for a flat arrow (see Figure 2-15).

%4 Part I: Using Excel to Summarizing Marketing Data

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	p	h	up												
2	q	i	down												
3	u	g	flat												
4							rank	rank	rank	rank	rank	trend	trend	trend	trend
5		January	February	March	April	May	January	February	March	April	May	Feb	March	April	May
6	Lebron	85	66	81	61	56	10	14	13	15	15	↓	↑	↓	→
7	Wade	82	63	74	78	75	12	15	15	12	14	↓	→	↑	↓
8	Dirk	45	100	115	127	150	15	4	3	2	1	↑	↑	↑	↑
9	Manning	75	88	89	76	83	13	10	11	13	12	↑	↓	↓	↑
10	Brady	96	90	98	76	93	9	9	9	13	10	→	→	↓	↑
11	Halliday	75	73	79	91	95	13	13	14	10	9	→	↓	↑	↑
12	Britney	98	91	109	99	84	8	8	4	8	11	→	↑	↓	↓
13	Lindsay	83	84	97	81	98	11	12	10	11	8	↓	↑	↓	↑
14	Paris	106	98	84	93	82	5	5	12	9	13	→	↓	↑	↓
15	JLO	104	88	109	101	115	6	10	4	6	3	↓	↑	↓	↑
16	Emma	115	94	105	101	107	3	7	6	6	6	↓	↑	→	→
17	Melo	118	98	128	126	108	2	5	1	3	4	↓	↑	↓	↓
18	KD	100	114	104	116	131	7	2	7	5	2	↑	↓	↑	↑
19	Vick	112	122	102	124	107	4	1	8	4	6	↑	↓	↑	↓
20	Rodgers	127	114	116	139	108	1	2	2	1	4	↓	→	↑	↓
21															

Figure 2-14: Monthly sales data

J	K	L
Letter	Wingdings 3	
a	↔	
b	↶	
c	↷	
d	↶	
e	↷	
f	←	
g	→	
h	↑	
i	↓	
j	↖	
k	↗	
l	↙	
m	↘	
n	↔	
o	↕	
p	▲	
q	▼	
r	△	
s	▽	
t	◀	
u	▶	
v	◁	
w	▷	
x	◀	
y	▶	
z	◀	

Figure 2-15: Icon sets created with Wingdings 3 font

To begin creating up, down, or flat arrows that reflect monthly changes in each salesperson's performance, perform the following steps:

- %4 Create each salesperson's rank during each month by copying the formula
 =RANK(E6,E\$6:E\$20,0) from J6 to J6:N20.

The last argument of 0 in the RANK function indicates that the largest number in E6:E20 will receive a rank of 1, the second largest number a rank of 2, and so on. If the last argument of the RANK function is 1, the smallest number receives a RANK of 1.

%4 Next create the h, i, and g that correspond to the arrows by copying the formula `=IF(K6<J6,"h",IF(K6>J6,"i","g"))` from O6 to O6:R20.

%4 Finally change the font in O6:R20 to Wingdings3.

You can now follow the progress of each salesperson (relative to her peers). For example, you can now see that Dirk Nowitzki improved his ranking each month.

Using Check Boxes to Control Data in a Chart

Often the marketing analyst wants to plot many series (such as sales of each product) in a single chart. This can result in a highly cluttered chart. Often it is convenient to use *check boxes* to control the series that are charted. The Checkboxes worksheet illustrates this idea. The data here shows weekly sales of chocolates, DVDs, magazines, soda, and hot dogs at the Quickie Mart convenience store. You can easily summarize this data in a single chart with five lines, as shown in Figure 2-16

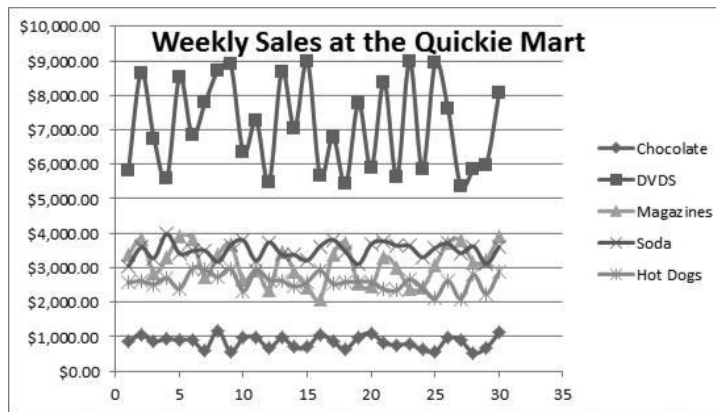


Figure 2-16: Sales at Quickie Mart

Now this chart is a good start, however, you might notice that the presence of five series appears cluttered, and therefore you would probably be better off if you

46 Part I: Using Excel to Summarizing Marketing Data

did not show the sales of every product. Check boxes make it easy to control which series show up in a chart. To create a check box, perform the following steps:

- %4 Ensure you see the Developer tab on the ribbon. If you do not see the Developer tab in Excel 2010 or 2013, select File > Options, and from Customize Ribbon select the Developer tab.
- %4 To place a check box in a worksheet, select the Developer tab and chose Insert.
- %4 From the Form Controls (Not ActiveX) select the check box that is third from the left in the top row.
- %4 After releasing the left mouse in the worksheet, you see a drawing tool. Hold down the left mouse to size the check box. If you want to resize the check box again later, simply hold down the control key and click the check box. A check box simply puts a TRUE or FALSE in a cell of your choosing. It serves as a Toggle switch that can be used to turn Excel features (such as a Function argument or Conditional Formatting) on or off.
- %4 Next, select the cell controlled by the check box by moving the cursor to the check box until you see the Pointer (a hand with a pointing finger).
- %4 Right-click, select the Format Control dialog box, and select cell F3 (by point-ing) as the cell link. Click OK.
- %4 Now when this check box is selected, F3 shows TRUE, and when the check box is not selected, cell F3 shows FALSE. Label this check box with the name of the series it controls, Chocolate in this case.

In a similar fashion four more check boxes were created that control G3, H3, I3, and J3. These check boxes control the DVDs, magazines, sodas, and hot dogs series, respectively. Figure 2-17 shows these check boxes.

After you have created the necessary check boxes, the trick is to not chart the original data but chart data controlled by the check boxes. The key idea is that Excel will ignore a cell containing an #N/A error when charting. To get started, copy the formula = IF(F\$3,F6,NA()) from O6 to O6:S35 to ensure that a series will only be charted when its check box is checked. For instance, check the Chocolate check box and a TRUE appears in F3, so O6 just picks up the actual week 1 chocolate sales in F6. If you uncheck the Chocolate check box, you get a FALSE in F3, O6 returns #N/A, and nothing shows up in the chart. To chart the data in question, select the modified data in the range O5:S35 and choose from the Insert tab the second X Y Scatter option. If you check only the Chocolate and Magazines check boxes, you

can obtain the chart shown in Figure 2-18, which shows only sales of Chocolate and Magazines.

	B	C	D	E	F	G	H	I	J
1									
2									
3					TRUE	FALSE	TRUE	FALSE	FALSE
4					Original Data				
5				Week	Chocolate	DVDS	Magazine	Soda	Hot Dogs
6				1	\$857.00	\$5,820.00	\$3,374.00	\$3,036.00	\$2,564.00
7				2	\$1,050.00	\$8,656.00	\$3,821.00	\$3,589.00	\$2,615.00
8	☒ Chocolate			3	\$863.00	\$6,749.00	\$2,857.00	\$3,279.00	\$2,512.00
9				4	\$933.00	\$5,580.00	\$3,284.00	\$3,981.00	\$2,705.00
10				5	\$905.00	\$8,534.00	\$3,923.00	\$3,411.00	\$2,386.00
11	☐ DVDS			6	\$901.00	\$6,837.00	\$3,837.00	\$3,484.00	\$2,977.00
12				7	\$597.00	\$7,794.00	\$2,713.00	\$3,498.00	\$2,946.00
13				8	\$1,185.00	\$8,708.00	\$3,375.00	\$3,169.00	\$2,729.00
14				9	\$565.00	\$8,892.00	\$3,674.00	\$3,660.00	\$2,977.00
15				10	\$964.00	\$6,338.00	\$2,640.00	\$3,797.00	\$2,322.00
16	☒ Magazines			11	\$982.00	\$7,258.00	\$2,926.00	\$3,201.00	\$2,939.00
17				12	\$658.00	\$5,468.00	\$2,360.00	\$3,738.00	\$2,648.00
18				13	\$961.00	\$8,690.00	\$3,466.00	\$3,374.00	\$2,619.00
19	☐ Soda			14	\$691.00	\$7,054.00	\$2,888.00	\$3,371.00	\$2,449.00
20				15	\$695.00	\$8,988.00	\$2,433.00	\$3,211.00	\$2,589.00
21				16	\$1,065.00	\$5,658.00	\$2,085.00	\$3,622.00	\$2,919.00
22				17	\$851.00	\$6,781.00	\$3,392.00	\$3,812.00	\$2,533.00
23	☐ Hot dogs			18	\$636.00	\$5,423.00	\$3,714.00	\$3,458.00	\$2,590.00
24				19	\$989.00	\$7,763.00	\$2,534.00	\$3,101.00	\$2,596.00

Figure 2-17: Check boxes for controlling which products show in chart

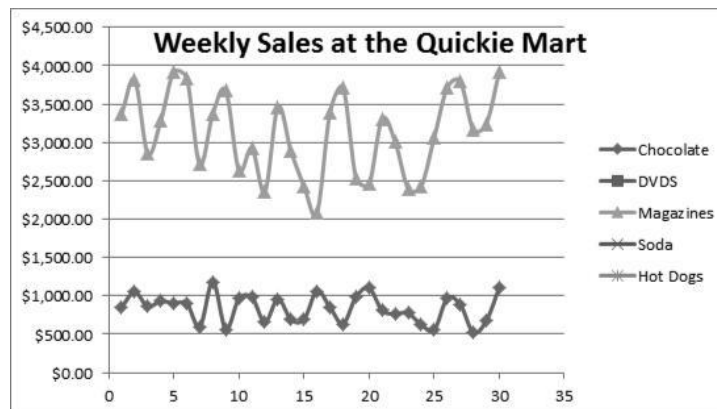


Figure 2-18: Charting only selected data

Using Sparklines to Summarize Multiple Data Series

Suppose you are charting daily sales of French Fries at each of the over 12,000 US McDonald's. Showing the sales for each restaurant in a single chart would result in a useless, cluttered graph. But think of the possibilities if you could summarize daily sales for each restaurant in a single cell! Fortunately, Excel 2010 and later enables you to create *sparklines*. A sparkline is a graph that summarizes a row or column of data in a single cell. This section shows you how to use Excel 2010 to easily create sparklines.

The Sparklines worksheet contains data that can be used to illustrate the concept of Sparklines. The data gives the number of engagement rings sold each day of the week in each city for a national jewelry store chain (see Figure 2-19).

	C	D	E	F	G	H	I	J	K
1									
2		Different scale for each sparkline							
3		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
4	Chicago	2520	3680	3852	3816	4800	5668	3600	
5	New York	6804	9630	8262	7290	10044	11700	9630	
6	Los Angeles	6489	8640	8910	8181	11340	12753	8550	
7	San Francisco	2184	3000	2673	2646	3528	4134	2970	
8	Dallas	3325	4650	4050	4050	6540	6630	5200	
9	Houston	4494	6120	5724	5778	6840	8580	5700	
10	Atlanta	3052	4240	3384	3924	4512	4680	4400	
11	Seattle	1890	3300	2700	2619	3960	3939	3030	
12	Miami	1316	1820	1800	1728	2328	2678	2120	
13	Boston	3948	6180	5562	5670	7056	7488	5940	
14	Nashville	1498	2020	1818	1962	2352	2600	1880	
15									
16		Same Scale for each sparkline							
17		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
18	Chicago	2520	3680	3852	3816	4800	5668	3600	
19	New York	6804	9630	8262	7290	10044	11700	9630	
20	Los Angeles	6489	8640	8910	8181	11340	12753	8550	
21	San Francisco	2184	3000	2673	2646	3528	4134	2970	
22	Dallas	3325	4650	4050	4050	6540	6630	5200	
23	Houston	4494	6120	5724	5778	6840	8580	5700	
24	Atlanta	3052	4240	3384	3924	4512	4680	4400	
25	Seattle	1890	3300	2700	2619	3960	3939	3030	
26	Miami	1316	1820	1800	1728	2328	2678	2120	
27	Boston	3948	6180	5562	5670	7056	7488	5940	
28	Nashville	1498	2020	1818	1962	2352	2600	1880	

Figure 2-19: Sparklines Example

You can summarize the daily sales by graphing the daily counts for each city in a single cell. To do so, perform the following steps:

- %4 First, select where you want your Sparklines to go (the Sparklines worksheet uses K4:K14; you can use L4:L14) and then from the Insert tab, select Line from the Sparklines section.
- %4 Fill in the dialog box shown in Figure 2-20 with the data range on which the Sparklines are based (D4:J14).

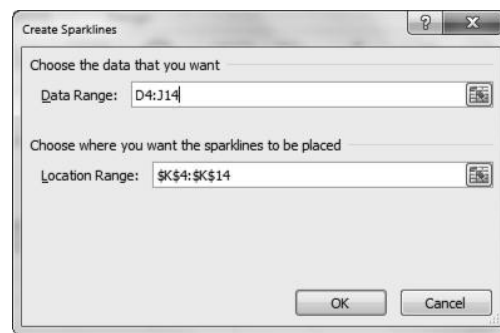


Figure 2-20: Sparkline dialog box

You now see a line graph (refer to Figure 2-19) that summarizes the daily sales in each city. The Sparklines make it clear that Saturday is the busiest day of the week.

If you click in any cell containing a Sparkline, the Sparkline Tools Design tab appears. Here, after selecting the Design tab, you can make many changes to your Sparklines. For example, Figure 2-21 shows selections for the high and low points to be marked and Figure 2-22 shows the Sparklines resulting from these selections.

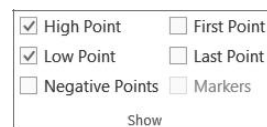


Figure 2-21: Selecting High and Low Point Markers

%4 Part I: Using Excel to Summarizing Marketing Data

	N	O	P	Q	R	S	T	U	V
1									
2		Different scale for each sparkline							
3		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
4	Chicago	2520	3680	3852	3816	4800	5668	3600	
5	New York	6804	9630	8262	7290	10044	11700	9630	
6	Los Angeles	6489	8640	8910	8181	11340	12753	8550	
7	San Francisco	2184	3000	2673	2646	3528	4134	2970	
8	Dallas	3325	4650	4050	4050	6540	6630	5200	
9	Houston	4494	6120	5724	5778	6840	8580	5700	
10	Atlanta	3052	4240	3384	3924	4512	4680	4400	
11	Seattle	1890	3300	2700	2619	3960	3939	3030	
12	Miami	1316	1820	1800	1728	2328	2678	2120	
13	Boston	3948	6180	5562	5670	7056	7488	5940	
14	Nashville	1498	2020	1818	1962	2352	2600	1880	

Figure 2-22: Sparklines with High and Low Markers

These Sparklines make it clear that Saturday was the busiest day for each branch and Monday was the slowest day.

The Design tab enables you to make the following changes to your Sparklines:

- %4 Alter the type of Sparkline (Line, Column, or Win-Loss). Column and Win-Loss Sparklines are discussed later in the chapter.
- %4 Use the Edit Data choice to change the data used to create the Sparkline. You can also change the default setting so that hidden data is included in your Sparkline.
- %4 Select any combination of the high point, low point, negative points, first point, or last point to be marked.
- %4 Change the style or color associated with the Sparklines and markers.
- %4 Use the Axis menu to change the way the axes are set for each Sparkline. For example, you may make the x-axis or y-axis scale the same for each Sparkline. This is the scaling used in the cell range K18:K28. Note this choice shows that Nashville, for example, sells fewer rings than the other cities. The default is for the scale for each axis to be based on the data for the individual Sparkline. This is the scaling used in the Sparklines in the cell range K4:K14 of Figure 2-22. The Custom Value choice enables you to pick an upper and lower limit for each axis.
- %4 When data points occur at irregularly spaced dates, you can select Data Axis Type from the Axis menu so that the graphed points are separated by an amount of space proportional to the differences in dates.

NOTE By clicking any of the Sparklines, you could change them to Column Sparklines simply by selecting Column from the Sparklines Design tab.

Excel also can create Win-Loss Sparklines. A Win-Loss Sparkline treats any positive number as an *up block* and any negative number as a *down block*. Any 0s are graphed as a gap. A Win-Loss Sparkline provides a perfect way to summarize performance against sales targets. In the range D32:J42 you can see a list of the daily targets for each city. By copying the formula =IF(D18>D32,1,-1) from D45 to D45:J55, you can create a 1 when a target is met and a -1 when a target is not met. To create the Win-Loss Sparklines, select the range where the Sparklines should be placed (cell range K45:K55) and from the Insert menu, select Win-Loss Sparklines. Then choose the data range D45:J55. Figure 2-23 shows your Win-Loss Sparklines.

	C	D	E	F	G	H	I	J	K
30		Target							
31		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
32	Chicago	3193	3592	3592	3991	3991	5188	3991	
33	New York	7241	8146	8146	9051	9051	11767	9051	
34	Los Angeles	7413	8340	8340	9266	9266	12046	9266	
35	San Francisco	2415	2717	2717	3019	3019	3925	3019	
36	Dallas	3937	4429	4429	4921	4921	6397	4921	
37	Houston	4941	5559	5559	6177	6177	8030	6177	
38	Atlanta	3222	3625	3625	4027	4027	5236	4027	
39	Seattle	2450	2756	2756	3063	3063	3981	3063	
40	Miami	1576	1773	1773	1970	1970	2561	1970	
41	Boston	4782	5380	5380	5978	5978	7771	5978	
42	Nashville	1615	1817	1817	2019	2019	2624	2019	
43		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
44		Met Goal?							
45	Chicago	-1	1	1	-1	1	1	-1	■ ■ ■ ■ ■
46	New York	-1	1	1	-1	1	-1	1	■ ■ ■ ■ ■
47	Los Angeles	-1	1	1	-1	1	1	-1	■ ■ ■ ■ ■
48	San Francisco	-1	1	-1	-1	1	1	-1	■ ■ ■ ■ ■
49	Dallas	-1	1	-1	-1	1	1	1	■ ■ ■ ■ ■
50	Houston	-1	1	1	-1	1	1	-1	■ ■ ■ ■ ■
51	Atlanta	-1	1	-1	-1	1	-1	1	■ ■ ■ ■ ■
52	Seattle	-1	1	-1	-1	1	-1	-1	■ ■ ■ ■ ■
53	Miami	-1	1	1	-1	1	1	1	■ ■ ■ ■ ■
54	Boston	-1	1	1	-1	1	-1	-1	■ ■ ■ ■ ■
55	Nashville	-1	1	1	-1	1	-1	-1	■ ■ ■ ■ ■

Figure 2-23: Win-Loss Sparklines

NOTE If you want your Sparklines to automatically update to include new data, you should make the data a table.

%4 Part I: Using Excel to Summarizing Marketing Data

Using GETPIVOTDATA to Create the End-of-Week Sales Report

Many marketing analysts download weekly sales data and want to summarize it in charts that update as you download the new data. The Excel Table feature and GETPIVOTDATA function can greatly ease this process. To illustrate this process, download the source data in columns D through G in the End of Month dashboard worksheet. Each row (see Figure 2-24) gives the week of sales, the sales category, the store in which sales were made, and the revenue generated. The goal is to create a line graph to summarize sales of selected product categories (for each store) that automatically update when new data is downloaded. The approach is as follows:

	D	E	F	G
4	WEEK	Catego	Store	Revenue
5	3	Abrasives	Lowe's	2043
6	12	Safety	Menards	2343
7	3	Tape	Home Depo	1414
8	12	Command	Target	1820
9	9	Tape	Home Depo	943
10	7	Tape	Target	1219
11	7	Command	Menards	1156
12	11	Abrasives	Lowe's	2127
13	12	Safety	Menards	1315
14	3	Tape	Target	1580
15	10	Abrasives	Home Depo	1598
16	4	Command	Lowe's	1000
17	7	Tape	Menards	1087
18	7	Abrasives	Menards	1728
19	1	Abrasives	Target	1911
20	7	Abrasives	Menards	1563
21	2	Tape	Target	2482
22	7	Safety	Lowe's	1534
23	12	Safety	Menards	1471
24	2	Abrasives	Lowe's	990
25	11	Tape	Lowe's	1580
26	1	Safety	Target	2389

Figure 2-24: Weekly sales data

%4 Make the source data a table; in this example you can do so by selecting Table from the Insert tab after selecting the range D4:G243.

%4 Create a PivotTable based on the source data by dragging Week to the Row Labels zone, Store and Category to the Column Labels zone, and Revenue to the Values zone. This PivotTable summarizes weekly sales in each store for each product category. Figure 2-25 shows a portion of this PivotTable.

	J	K	L	M	N	O	P	Q	R	S	T	U
11	Column Labels											
12	Lowe's				Lowe's Total	Menards			Menards Total	Home Depot		
13	Abrasives	Safety	Tape	Command		Abrasives	Safety	Tape	Command		Abrasives	Safety
14		1065		1247	2312	3772	526	3466		7764	3272	1984
15	2349	4510			6859	6220	3296	2159		11675	1921	2167
16	2797	1233	2046	6335	12411	3472		959	2369	6800		3352
17	1501	3822	1180	1000	7503	6018	2212			8230		
18	1940		833	2945	5718	708		501	597	1806	1758	3755
19			1918	6494	8412	4145	6205	4653		15003		2253
20	3049	3228			6277	3291	2303	6472	1156	13222	3580	4218
21	4208	2890	1873		8971	910		1494	3776	6180	2445	5496
22			2052	2825	4877	6265		1888		8153		
23	6117		1875		7992			2138		2138	4807	2325
24	2127	1971	4461	2394	10953	3596	1169			4765		810
25		1542	4786	850	7178	1124	7565	7264	2431	18384	2471	
26				1196	1196				3400	3400		
27	1455				1455							
28	25543	20261	21024	25286	92114	39521	23276	30994	13729	107520	20254	26360

Figure 2-25: PivotTable for Weekly sales report

- %4 Create a drop-down box in cell AG8 from which a store can be selected. To accomplish this, navigate to the Data tab and select Data Validation from the Data Tools group.
- %4 Choose Settings and fill in the Data Validation dialog box, as shown in Figure 2-26. This ensures that when the cursor is placed in cell AG8 you can see a drop-down box that lists the store names (pulled from cell range F5:F8)

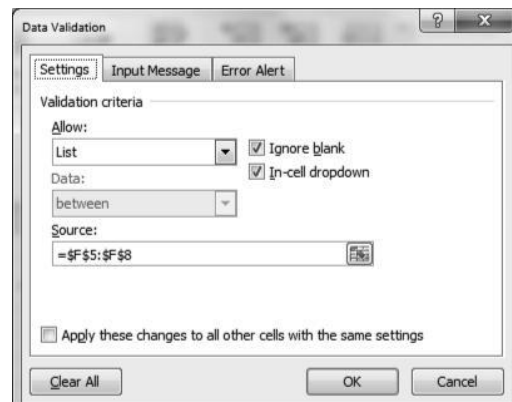


Figure 2-26: Creating Data Validation drop-down box

- %4 Enter week numbers in the range AG11:AG24 and product categories in AG10:AK10. Then use the GETPIVOTDATA function to pull the needed data from the PivotTable. Before entering the key formula, simply click anywhere in the PivotTable to get a GETPIVOTDATA function. Then copy that function and

%4 Part I: Using Excel to Summarizing Marketing Data

paste it in the following formula to modify the arguments for Week, Category, and Store, as shown here. Enter this important formula in cell AH11:

```
=IF(AH$9=FALSE,NA(),IFERROR(GETPIVOTDATA("Revenue",$I$11,"WEEK",
,$AG11,"Category",AH$10,"Store",$AG$8)," ")).
```

%4 Copy this formula to the range AH11:AK24. This formula uses GETPIVOTDATA to pull revenue for the store listed in AG8 and the category listed in row 10 if row 9 has a TRUE. The use of IFERROR ensures a blank is entered if the category were not sold in the selected store during the given week.

%4 Make the source data for the chart, AG10:AK24, a table, so new data is auto-matically included in the graph. The Table feature can also ensure that when you enter a new week the GETPIVOTDATA formulas automatically copy down and pull needed information from the updated PivotTable.

%4 Use check boxes to control the appearance of TRUE and FALSE in AH9:AK9. After making the range AG10:AK24 a table, the chart is simply an X-Y scatter graph with source data AG10:AK24. Figure 2-27 shows the result.

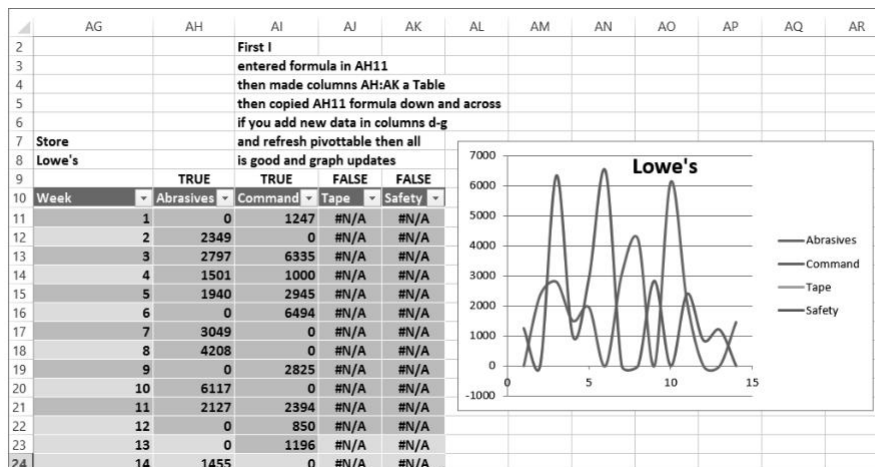


Figure 2-27: Sales summary report

%4 Now add new data for Week 15 and refresh the PivotTable. If you add Week 15 in cell AG25, the graph automatically updates. Because you made the source data for the chart a table, when you press Enter Excel “knows” you want the formulas copied down.

Summary

In this chapter you learned the following:

- %4 Right-clicking a series in a chart enables you to change the series' chart type and create a Combination chart.
- %4 Right-clicking a chart series and choosing Format Data Series enables you to create a secondary axis for a chart.
- %4 Right-clicking a chart series in a Column graph and selecting Fill followed by Picture or Texture Fill enables you to replace a bland column with a picture from a file or Clip Art.
- %4 On Chart Tools from the Layout tab, you can easily insert Data Labels and a Data Table in the chart.
- %4 If you make the source data for a chart a table (from the Insert tab choose Table) before creating a chart, the chart automatically updates to include new data.
- %4 PivotCharts are a great way to summarize a lengthy market research survey. Filtering on the questions enables you customize the results shown in your chart.
- %4 If you base your chart title and series labels on cell formulas, they dynamically update as you change the inputs or assumptions in your spreadsheet.
- %4 Clever use of IF formulas and the Wingdings 3 font can enable you to create a visually appealing summary of trends over time in sales data.
- %4 Use check boxes to control the series that appear in a chart.
- %4 Combining the Table feature, PivotTables, GETPIVOTDATA, check boxes, and Data Validation drop-down boxes makes it easy to create charts with custom-ized views that automatically update to include new data.

Exercises

Exercises 1–6 use data in the file Chapter2exercisesdata.xlsx

- %4 The Weather worksheet includes monthly average temperature and rainfall in Bloomington, Indiana. Create a combination chart involving a column and line graph with a secondary axis to chart the temperature and rainfall data.
- %4 The Weather worksheet includes monthly average temperature and rainfall in Bloomington, Indiana. Create a combination chart involving a column and area graph with a secondary axis to chart the temperature and rainfall data.

%4 Part I: Using Excel to Summarizing Marketing Data

The Pictures and Labels worksheet includes monthly tomato sales on Farmer Smith's farm. Summarize this data with pictures of tomatoes, data labels, and a data table.

The Survey worksheet contains results evaluating a training seminar for salespeople. Use a PivotChart to summarize the evaluation data.

The data in the checkboxes worksheet contains monthly sales during 2010 and 2011. Use check boxes to set up a chart in which the user can choose which series are charted.

The Income worksheet contains annual data on median income in each state for the years 1984–2010. Use Sparklines to summarize this data.

Jack Welch's GE performance evaluation system requires management to classify the top 20 percent of all workers, middle 70 percent of all workers, and bottom 10 percent of all workers. Use Icon sets to classify the salespeople in the Sales Tracker worksheet of file Chapter2data.xlsx according to the 20-70-10 rule.

NOTE You need the PERCENTILE function. For example, the function PERCENTILE(A1:A50,9) would return the 90th percentile of the data in the cell range A1:A50.

Exercises 8 and 9 deal with the data in the file LaPetitbakery.xlsx that was discussed in Chapter 1.

%4 Use the Excel Table feature to set up a chart of daily cake sales that updates automatically when new data is included.

%4 Set up a chart that can be customized to show total monthly sales for any subset of La Petit Bakery's products. Of course, you want the chart to update automatically when new data appears in the worksheet.

%4 The marketing product life cycle postulates that sales of a new product will increase for a while and then decrease. Specify the following five inputs:

Year 1 sales

Years of growth

Years of decline

Average annual growth rate during growth period

Average annual decline during decline period

Set up a Data Validation drop-down box that allows years of growth and decline to vary between 3 and 10. Then determine sales during Years 1–20. Suppose 10,000 units need to be sold in a year to break even. Chart your annual sales, and set up a dynamic chart title that shows the year (if any) in which you break even. Your series label should include the values of the five input parameters.



3

Using Excel Functions to Summarize Marketing Data

In Chapter 1, “Slicing and Dicing Marketing Data with PivotTables,” and Chapter 2, “Using Excel Charts to Summarize Marketing Data,” you learned how to summarize marketing data with PivotTables and charts. Excel also has a rich library of powerful functions that you can use to summarize marketing data. In this chapter you will learn how these Excel functions enable marketers to gain insights for their

data that can aid them to make informed and better marketing decisions.

This chapter covers the following:

- %4 Using the Excel Table feature to create dynamic histograms that summarize marketing data and update automatically as new data is added.
- %4 Using Excel statistical functions such as AVERAGE STDEV RANK PERCENTILE and PERCENTRANK to summarize marketing data.
- %4 Using the powerful “counting and summing functions” (COUNTIF COUNTIFS SUMIF SUMIFS AVERAGEIF, and AVERAGEIFS) to summarize marketing data.
- %4 Writing array formulas to perform complicated statistical calculations on any subset of your data.

Summarizing Data with a Histogram

%4 *histogram* is a commonly used tool to summarize data. A histogram essentially tells you how many data points fall in various ranges of values. Several Excel functions, including *array functions* such as the TRANSPOSE and the FREQUENCY function, can be used to create a histogram that automatically updates when new data is entered into the spreadsheet.

%4 Part I: Using Excel to Summarizing Marketing Data

When using an array function, you need to observe the following rules:

- %4 Select the range of cells that will be populated by the array function. Often an array function populates more than one cell, so this is important.
- %4 Type in the function just like an ordinary formula.
- %4 To complete the entry of the formula, do not just press Enter; instead press Control+Shift+Enter. This is called *array entering* a formula.

Because the FREQUENCY function is a rather difficult array function, let's begin with a discussion of a simpler array function, the TRANSPOSE function.

Using the TRANSPOSE Function

The TRANSPOSE function is a great place to start when illustrating the use of array functions. To begin, take a look at the Histogram worksheet in the Chapter3bakery. xlsx file. In Figure 3-1, cells N27:Q27 list some of my best students. This list would be easier to digest though if the names were listed in a column. To do this, copy N27:Q27 and move the cursor to R28, then right-click and select Paste Special... Select the Transpose checkbox to paste the names in a column.

	N	O	P	Q	R	S
27	Scarlett	Lindsay	Britney	Paris	PASTE SPECIAL	TRANSPOSE FUNCTION
28					Scarlett	Scarlett
29					Lindsay	Lindsay
30					Britney	Britney
31					Paris	Paris

Figure 3-1: Use of the Transpose function

Unfortunately, if you change the source data (for example, Scarlett to Blake) the transposed data in column R does not reflect the change. If you want the transposed data to reflect changes in the source data, you need to use the TRANSPOSE function. To do so perform the following steps:

- %4 Select S28:S31.
- %4 Enter in cell S28 the formula = TRANSPOSE (N 2 7 : Q 2 7) and press Control+Shift+Enter. You will know an array formula has been entered when you see a { in the Formula bar.
- %4 Change Scarlett in N27 to Blake, and you can see that Scarlett changes to Blake in S28 but not in R28.

Using the FREQUENCY Function

The FREQUENCY function can be used to count how many values in a data set fit into various ranges. For example, given a list of heights of NBA basketball players the FREQUENCY function could be used to determine how many players are at least 7' tall, how many players are between 6' 10" and 7' tall, and so on. The syntax of the FREQUENCY function is `FREQUENCY(array,bin range)`. In this section you will utilize the FREQUENCY function to create a histogram that charts the number of days in which cake sales fall into different numerical ranges. Moreover, your chart will automatically update to include new data.

On the Histogram worksheet you are given daily sales of cakes at La Petit Bakery. You can summarize this data with a histogram. The completed histogram is provided for you in the worksheet, but if you want to follow along with the steps simply copy the data in D7:E1102 to the same cell range in a blank worksheet. Begin by dividing a range of data into 5–15 equal size ranges (called a *bin range*) to key a histogram. To do so, perform the following steps:

- %4 Select your data (the range D7:E1102) and make this range a table by select-ing Insert > Table. If a cell range is already a table then under Insert the table option will be grayed out.
- %4 In cell E5, point to all the data and compute the minimum daily cake sales with the formula `=MIN(Table1[Cakes])`
- %4 Similarly, determine the maximum daily cake sales in E6. The range of daily cake sales is between 32 and 165. Select your bin ranges to be `<= 30, 31–50, 51–70, ..., 151–170..., > 170` cakes.
- %4 List the boundaries for these bin ranges in the cell range I9:I17. Figure 3-2 shows this process in action.

You can now use the FREQUENCY function to count the number of days in which cake sales fall in each bin range. The syntax of the FREQUENCY function is `FREQUENCY(array,bin range)`. When an array is entered, this formula can count how many values in the array range fall into the bins defined by the bin range. In this example the function will count the number of days in which cake sales fall in each bin range. This is exemplified in Figure 3-2. To begin using the FREQUENCY function, perform the following steps:

- %4 Select the range J9:J17 in the Histogram worksheet and type the formula `=FREQUENCY(E8:E1102,I9:I17)`

%4 Part I: Using Excel to Summarizing Marketing Data

Array enter this formula by pressing Control+Shift+Enter. This formula computes how many numbers in the table are ≤ 30 , between 31 and 50,..., between 151 and 170, and more than 170.

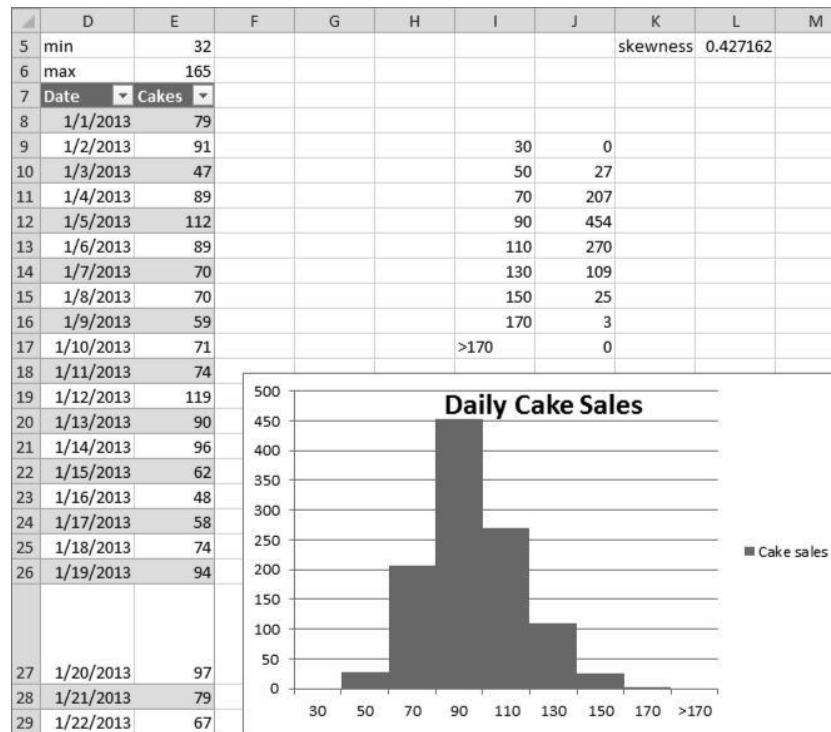


Figure 3-2: Dynamic histogram for cake sales

For example, you can find that there were three days in which between 151 and 170 cakes were sold. The counts update automatically if new days of sales are added.

- %4 Select the range I9:J17, and from the Insert tab, select the first Column Graph option.
- %4 Right-click any column, select Format Data Series..., and change Gap Width to 0. This enables you to obtain the histogram, as shown in Figure 3-2.
- %4 If you enter more data in Column E (say 10 numbers >170) you can see that the histogram automatically updates to include the new data. This would not have happened if you had failed to make the data a Table.

Skewness and Histogram Shapes

There are a variety of histograms that are seen when examining marketing data, but the three most commonly seen histogram shapes are listed here (see Figure 3-3 and file *Skewexamples.xlsx*):

- %4 *Symmetric histogram*: A histogram is symmetric and looks approximately the same to the left of the peak and the right of the peak. In Figure 3-3 the IQ scores yield an asymmetric histogram.
- %4 *Positively skewed*: (skewed right) A histogram is positively skewed (or skewed right) if it has a single peak and the values of the data set extend much further to the right of the peak than to the left of the peak. In Figure 3-3 the histogram of family income is positively skewed.
- %4 *Negatively skewed*: (skewed left) A histogram is negatively skewed if it has a single peak and the values of the data set extend much further to the left of the peak than to the right of the peak. In Figure 3-3, the histogram of Days from Conception to Birth is negatively skewed.

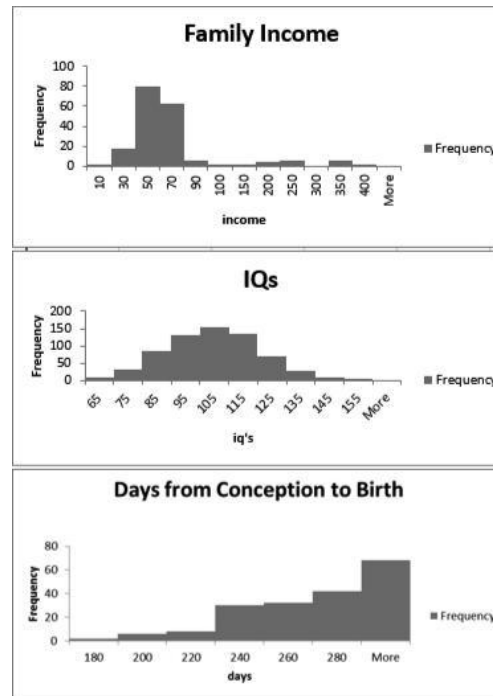


Figure 3-3: Symmetric, positively skewed, and negatively skewed histograms

%4 Part I: Using Excel to Summarizing Marketing Data

The Excel SKEW function provides a measure of the skewness of a data set. A skewness value $> +1$ indicates mostly positive skewness, < -1 mostly negative skewness, and a skewness value between -1 and $+1$ indicates a mostly symmetric histogram. This function can be used to obtain a quick characterization of whether a histogram is symmetric, positively skewed, or negatively skewed. To see this function in action, enter the formula `=SKEW(Table1[Cakes])` into cell L5 in the Histogram worksheet to yield a skewness of .43, implying that cake cells follow a symmetric distribution. Because you made the source data an Excel table, the skewness calculation in L5 would automatically update if new data is added.

In the next section you will see that the degree of skewness (or lack thereof) in a data set enables the marketing analyst to determine how to best describe the typical value in a data set.

Using Statistical Functions to Summarize Marketing Data

In her **daily work** the marketing analyst will often encounter large datasets. It is often difficult to make sense of these data sets because of their vastness, so it is helpful to summarize the data on two dimensions:

- %4 **A typical value of the data:** For example, what single number best summarizes the typical amount a customer spends on a trip to the supermarket.
- %4 **Spread or variation about the typical value:** For example, there is usually much more spread about the typical amount spent on a visit to a grocery superstore than a visit to a small convenience store.

Excel contains many statistical functions that can summarize a data set's typical value and the spread of the data about its mean. This section discusses these statistical functions in detail.

Using Excel Functions to Compute the Typical Value for a Data Set

The file `Chapter3bakery.xlsx` contains the La Petit Bakery sales data that was discussed in Chapter 1. For each product, suppose you want to summarize the typical

number of that product sold in a day. Three measures are often used to summarize the *typical value* for a data set.

%4 The mean, or *average*, is simply the sum of the numbers in the data set divided by the number of values in the dataset. The average is computed with the Excel AVERAGE function. For example, the average of the numbers 1, 3, and 5 is $\frac{1+3+5}{3} = 3$. You can often use $\bar{x}$ to denote the mean or average of a data set.

%4 The *median* is approximately the 50th percentile of the data set. Roughly one-half the data is below the median and one-half the data is above the median. More precisely, suppose there are n values that when ordered from smallest to largest are $x_1, x_2, \dots, x_n$. If n is odd the median is $x_{(n+1)/2}$, and if n is even, the median is $(x_{n/2} + x_{(n/2)+1})/2$. For example, for the data set 1, 3, 5, the median is 3, and for the data set 1, 2, 3, 4 the median is 2.5. The MEDIAN function can be used to compute the median for a data set.

%4 The *mode* of a data set is simply the most frequently occurring value in the data set. If no value appears more than once, the data set has no mode. The MODE function can be used to compute the mode of a data set. A data set can have more than one mode. If you have Excel 2010 or later, then the array function MODE.mult (see *Excel 2010 Data Analysis and Business Modeling* by Wayne Winston [Microsoft Press, 2011]) can be used to compute all modes. For example, the data set 1, 3, and 5 has no mode, whereas the data set 1, 2, 2, 3 has a mode of 2.

The Descriptive stats worksheet uses Excel functions to compute three measures of typical value for daily sales of cakes, pies, cookies, smoothies, and coffee. Complete the following steps and see Figure 3-4 for details.

%4 To compute the average for daily cake sales in cell H5, enter the formula
=AVERAGE(H12:H1106)

%4 To compute the median for daily cake sales in cell H6, enter the formula
=MEDIAN(H12:H1106)

%4 To compute the mode of daily cake sales in cell H7, enter the formula
=MODE(H12:H1106)

%4 Copy these formulas to the range I5:L7 to compute these measures for La Petit Bakery's other products.

You find, for example, that on average 86.28 cakes were sold each day, around one-half the time fewer than 85 cakes were sold, and the most frequently occurring number of cakes sold in a day is 90.

%4 Part I: Using Excel to Summarizing Marketing Data

	G	H	I	J	K	L	M
1	5th largest cake day	148					
2	3rd smallest cake day	38					
3	rank of 136 cakes	0.979					
4							
5	mean	86.28037	53.54795	540.4511	218.4666667	391.6265	
6	median	85	52	533	206	388	
7	mode	90	46	397	166	375	
8	standard deviation	20.27667	13.37189	132.3544	74.35093017	102.9305	
9	95th percentile	122	77.2	764.8	352.2	568.4	
10	Skewness	0.427162	0.474088	0.440287	0.785999912	0.360745	
11	Date	Cakes	Pies	Cookies	Smoothies	Coffee	promotion
12	1/1/2013	79	46	518	60	233	none
13	1/2/2013	91	50	539	161	427	none
14	1/3/2013	47	60	222	166	347	none
15	1/4/2013	89	64	734	153	358	none
16	1/5/2013	112	73	764	240	392	none
17	1/6/2013	89	57	922	259	510	none
18	1/7/2013	70	50	476	120	334	none
19	1/8/2013	70	48	496	222	316	none
20	1/9/2013	59	37	587	181	156	none
21	1/10/2013	71	36	488	178	298	none
22	1/11/2013	74	50	645	100	490	none
23	1/12/2013	119	71	438	162	416	none
24	1/13/2013	90	51	568	137	434	none
25	1/14/2013	96	48	585	194	573	promotion

Figure 3-4: Descriptive statistics

Which Measure of Typical Value Is Best?

Although there is no single best answer to this question, there are a few rules of thumb that are summarized here. The mode might be relevant if a hat store wanted to maximize sales and was going to stock only one size of hats. In most cases, however, the analyst who wants to summarize a data set by a single number must choose either the mean or the median. In the presence of substantial positive or negative skewness, extreme values tend to distort the mean, and the median is a better choice as a summary of a typical data value. In other situations using the median throws out important information, so the mean should be used to summarize a data set. To summarize:

- %4 If a skew statistic is greater than 1 or less than -1 , substantial skewness exists and you should use median as a measure of typical value.
- %4 If skew statistic is between -1 and 1 , use mean as a measure of typical value.

For the cake example, data skewness is between -1 and 1 , so you would use the mean of 86.28 cakes as a summary value for daily cake sales.

For an example of how skewness can distort the mean, consider the starting salaries of North Carolina students who graduated in 1984. Geography majors had by far the highest average starting salary. At virtually every other school, business majors had the highest average starting salary. You may guess that Michael Jordan was a UNC geography major, and his multimillion dollar salary greatly distorted the average salary for geography majors. The median salary at UNC in 1984 was, as expected, highest for business majors.

Using the VAR and STDEV Functions to Summarize Variation

Knowledge of the typical value characterizing a data set is not enough to completely describe a data set. For example, suppose every week Customer 1 spends \$40 on groceries and during a week Customer 2 is equally likely to spend \$10, \$20, \$30, or \$100 on groceries. Each customer spends an average of \$40 per week on groceries but we are much less certain about the amount of money Customer 2 will spend on groceries during a week. For this reason, the description of a data set is not complete unless you can measure the spread of a data set about its mean. This section discusses how the variance and standard deviation can be used to measure the spread of a data set about its mean.

Given a data set $x_1, x_2, \dots, x_n$ the *sample variance* (s^2) of a data set is approximately the average squared deviation of the mean and may be computed with the formula

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

set 1, 3, 5 the average is 3, so the sample variance is as follows:

$$\frac{1}{2} \{ (1-3)^2 + (3-3)^2 + (5-3)^2 \} = 4$$

Note that if all data points are identical, then the sample variance is 0. You should square deviations from the mean to ensure that positive and negative deviations from the mean will not cancel each other out.

Because variance is not in the same units as dollars, the analyst will usually take the square root of the sample variance to obtain the *sample standard deviation* (s). Because the data set 1, 3, and 5, has a sample variance of 4, $s=2$.

The Excel VAR function computes sample variance, and the Excel STDEV function computes sample standard deviation. As shown in Figure 3-4 (see Descriptive stats worksheet), copying the formula =STDEV(H12:H1106) from H8 to K8:L8 computes the standard deviation of each product's daily sales. For example, the standard deviation of daily cake sales was 20.28 cakes.

%4 Part I: Using Excel to Summarizing Marketing Data

The Rule of Thumb for Summarizing Data Sets

If a data set does not exhibit substantial skewness, then the statisticians' interesting and important rule of thumb lets you easily characterize a data set. This rule states the following:

- %4 Approximately 68 percent of all data points are within 1 sample standard deviation of the mean.
- %4 Approximately 95 percent of all data points are within 2 standard deviations ($2s$) of the mean.
- %4 Approximately 99.7 percent of the data points are within 3 standard deviations of the mean.

Any data point that is more than two sample standard deviations ($2s$) from the mean is unusual and is called an *outlier*. Later in the chapter (see the “Verifying the Rule of Thumb” section) you can find that for daily cake sales, the rule of thumb implies that 95 percent of all daily cake sales should be between 45.3 and 126.83. Thus cake sales less than 45 or greater than 127 would be an outlier. You can also find that 95.62 percent of all daily cake sales are within $2s$ of the mean. This is in close agreement with the rule of thumb.

The Percentile.exc and Percentrank.exc Functions

A common reorder policy in a supply chain is to produce enough to have a low percent chance, say, a 5 percent, of running out of a product. For cakes, this would imply that you should produce x cakes, where there is a 5 percent chance demand for cakes that exceed x or a 95 percent chance demand for cakes that is less than or equal to x . The value x is defined to be the 95th percentile of cake demand. The Excel 2010 and later function `PERCENTILE.EXC(range, k)` (here k is between 0 and

%4 returns the $100 * k$ th percentile for data in a range. In earlier versions of Excel, you should use the `PERCENTILE` function. By copying the formula `=PERCENTILE.EXC(H12:H1106,0.95)` from H9 to I9:L9 of the Descriptive stats worksheet you can obtain the 95th percentile of daily sales for each product. For example, there is a 5 percent chance that more than 122 cakes will be sold in a day.

Sometimes you want to know how unusual an observation is. For example, on December 27, 2015, 136 cakes were sold. This is more than two standard deviations above average daily cake sales, so this is an outlier or unusual observation. The `PERCENTRANK.EXC` function in Excel 2010 or later (or `PERCENTRANK` in earlier versions of Excel) gives the ranking of an observation relative to all values in a data

set. The syntax of PERCENTRANK.EXC is PERCENTRANK(range, x,[significance]) This returns the percentile rank (as a decimal) of x in the given range. Significance is an optional argument that gives the number of decimal points returned by the function. Entering in cell H3 the formula =PERCENTRANK.EXC(H12:H1106,H1102) you find that you sold more than 136 cakes on only 2.1 percent of all days.

The LARGE and SMALL Functions

Often you want to find, say, the fifth largest or third smallest value in an array. The function LARGE(range,k) returns the k th largest number in a range, whereas the function SMALL(range,k) returns the k th smallest number in a range. As shown in Figure 3-4, entering in cell H1 the formula =LARGE(H12:H1106,5) tells you the fifth largest daily sales of cakes was 148. Similarly, entering in cell H2 the formula =SMALL(H12:H1106,3) tells you the third smallest daily demand for cakes was 38.

These same powerful statistical functions can be used by marketing managers to gain important insights such as who are the 5 percent most profitable customers, who are the three most costly customers, or even what percentage of customers are unprofitable.

Using the COUNTIF and SUMIF Functions

In addition to mathematical functions such as SUM and AVERAGE, the COUNTIF and SUMIF functions might be (or should be!) two of the most used functions by marketing analysts. These functions provide powerful tools that enable the analyst to select a subset of data in a spreadsheet and perform a calculation (count sum or average) on any column in the spreadsheet. The wide variety of computations performed in this section should convince the marketer that these functions are a necessary part of your toolkit.

The first few examples involve the COUNTIF and SUMIF functions. The COUNTIF function has the syntax COUNTIF(range, criteria) . Then the COUNTIF function

counts how many cells in the range meet the criteria. The

function has the
SUMIF

syntax SUMIF (range criteria , and sum range). Then the SUMIF function adds up the entries in the sumrange column for every row in which the cell in the range column meets the desired criteria.

The following sections provide some examples of these functions in action. The examples deal with the La Petit Bakery data, and the work is in the Sumif Countif worksheet. The work is also shown in Figure 3-5. The Create From Selection feature (discussed in Chapter 2) is used in the examples to name columns F through O with their row 10 headings.

70 Part I: Using Excel to Summarizing Marketing Data

	F	G	H	I	J	K	L	M	N	O	P	Q	R
1													
2			Outside										
3	MEAN-2SIGMA	45.72702047	13										
4	MEAN+2SIGMA	126.8337101	35										
5	TOTAL	1095											
6	WITHIN 2 SIGMA	1047											
7	FRACTION	95.62%											
8													
9													
10	weekday	Date	Cakes	Pies	Cookies	Smoothies	Coffee	promotion	month	Named month			
11	2	1/1/2013	79	46	518	60	233	none	1	January			
12	3	1/2/2013	91	50	539	161	427	none	1	January			
13	4	1/3/2013	47	60	222	166	347	none	1	January		How many promotions?	
14	5	1/4/2013	89	64	734	153	358	none	1	January		Promotion	110
15	6	1/5/2013	112	73	764	240	392	none	1	January		None	985
16	7	1/6/2013	89	57	922	259	510	none	1	January			
17	1	1/7/2013	70	50	476	120	334	none	1	January		Days sold>=122 cakes	56

Figure 3-5: Illustrating the Rule of Thumb

Counting the Number of Promotions

The Excel COUNTIF function can be used to count the number of rows in a range of cells that meet a single criterion. For example, you can compute the number of days on which La Petit Bakery had a promotion and did not have a promotion by using the COUNTIF function. Simply copy the formula =COUNTIF(promotion,Q14) from R14 to R15. You will see that there were 110 days with a promotion and 985 without a promotion. Note that Q14 as a criterion ensures that you only count days on which there was a promotion and Q15 as a criterion ensures that you only count days on which there was no promotion.

Verifying the Rule of Thumb

In the “Rule of Thumb” section of this chapter you learned that the rule of thumb tells you that for approximately 95 percent of all days, cake sales should be within 2s of the mean. You can use the COUNTIF function to check if this is the case. In cell G3 you can compute for daily cake sales the Average $-2s$ with the formula AVERAGE(Cakes) - 2*STDEV(Cakes). In G4 you can compute the Average $+2s$ with the formula AVERAGE(Cakes) + 2*STDEV(Cakes).

You can also find the number of days (13) on which cakes were more than 2s below the mean. To do so, enter in cell H3 the formula COUNTIF(Cakes,"<"&G3). The & (or concatenate) sign combines with the less than sign (in quotes because < is text) to ensure you only count entries in the Cakes column that are more than 2s (s = sample standard deviation) below the mean.

In cell H4 the formula COUNTIF(Cakes,">"&G4) computes the number of days (35) on which cake sales were more than 2s above the mean. Any data point that differs from the dataset's mean by more than 2s is called an outlier. Chapters 10 and 11 will make extensive use of the concept of an outlier.

The COUNT function can be used to count the number of numeric entries in a range while the COUNTA function can be used to count the number of non-blanks in a range. Also the COUNTBLANK function counts the number of blank cells in a range. For example, the function =COUNT(Cakes) in cell G5 counts how many numbers appear in the Cake column (1095). The COUNTA function counts the number of nonblank (text or numbers) cells in a range, and the COUNTBLANK function counts the number of blanks in a range. In G6 compute (with formula =G5/H3/H4) the number of cells (1047) within 2s of the mean. In G7 you find that 95.62 percent of the days had cake cells within 2s of the mean. This is in close agreement with the rule of thumb.

Computing Average Daily Cake Sales

You can also use SUMIF and COUNTIF functions to calculate the average sales of cakes for each day of the week. To do so, perform the following steps:

- %4 Copy the formula =SUMIF(daywk,Q20,Cakes) from S20 to S21:S26 to compute the total cake sales for each day of the week.
- %4 Copy the formula =COUNTIF(daywk,Q20) from R20 to R21:R26 to compute for each day of the week the number of times the day of the week occurs in your data set.
- %4 Finally, in Column T, divide the SUMIF result for each day by the COUNTIF result, and this gives the average cake sales for each day of the week.

To rank the average cake sales for each day of the week, copy the formula =RANK(T20,\$T\$20:\$T\$26,0) from U20 to U21:U26. The formula in U20 ranks the Monday average sales among all days of the week. Dollar signing the range T20:T26 ensures that when you copy the formula in U21, each day's average sales is ranked against all 7 days of the week. If you had not dollar signed the range T20:T26, then in U21 the range would have changed to T21:T27, which would be incorrect because this range excludes Monday's average sales. The last argument of 0 ensures that the day of the week with the largest sales is given a rank of 1. If you use a last argument of 1, then sales are ranked on a basis that makes the smallest number have a rank of 1, which is not appropriate in this situation. Note that Saturday is the best day for sales and Wednesday was the worst day for sales.

In an identical manner, as shown in Figure 3-6 you can compute the average daily sales of smoothies during each month of the year in the cell range Q28:U40. The results show that the summer months (June–August) are the best for smoothie sales.

%4 Part I: Using Excel to Summarizing Marketing Data

	Q	R	S	T	U
18	Cakes				
	Total Sales by day of	How many	Total	Average	
19	week	days	Sales	Daily	Rank
20	Monday	156	12259	78.58333	4
21	Tuesday	157	12079	76.93631	6
22	Wednesday	157	11986	76.34395	7
23	Thursday	157	12148	77.3758	5
24	Friday	156	14770	94.67949	3
25	Saturday	156	16181	103.7244	1
26	Sunday	156	15054	96.5	2
27	Smoothies				
	Total sales by month of	How many	Total	Average	
28	year	days	Sales	Daily	Rank
29	January	93	15955	171.5591	12
30	February	84	14692	174.9048	10
31	March	93	16930	182.043	8
32	April	90	19016	211.2889	7
33	May	93	22434	241.2258	4
34	June	90	24971	277.4556	3
35	July	93	26654	286.6022	2
36	August	93	26724	287.3548	1
37	September	90	19737	219.3	5
38	October	93	19813	213.043	6
39	November	90	15463	171.8111	11
40	December	93	16832	180.9892	9

Figure 3-6: Daily summary of cake sales

In the next section you will learn how the AVERAGEIF function makes these calculations a bit simpler.

Using the COUNTIFS, SUMIFS, AVERAGEIF, and AVERAGEIFS Functions

COUNTIF and SUMIF functions are great, but they are limited to calculations based on a single criteria. The next few examples involve the use of four Excel functions that were first introduced in Excel 2007: COUNTIFS, SUMIFS, AVERAGEIF, and AVERAGEIFS (see the New 2007 Functions worksheet). These functions enable you to do calculations involving multiple (up to 127!) criteria. A brief description of the syntax of these functions follows:

%4 The syntax of COUNTIFS is COUNTIFS(range1,criteria1, range2,criteria2, .. range_n, criteria_n) COUNTIFS counts the number of rows in which

the *range1* entry meets *criterial* , the *range2* entry meets *criteria2* , and ...
the *range_n* entry meets *criteria_n*

%4 The syntax of SUMIFS is SUMIFS(sum_range , range1, criterial, range2, criteria2, ,range_n,criteria_n) SUMIFS sums up every entry in the sum_range for which *criterial* (based on *range1*), *criteria2* (based on *range2*), *criteria_n* (based on *range_n*) are all met.

%4 The AVERAGEIF function has the syntax AVERAGEIF(range, criteria, average_range) AVERAGEIF averages the range of cells in the average range for which the entry in the range column meets the criteria.

%4 The syntax of AVERAGEIFS is AVERAGEIFS(average_range, criteria_range1, criteria_range2,criteria_range_n) AVERAGEIFS averages every entry in the average range for which *criterial* (based on *range1*), *criteria2* (based on *range2*), ... *criteria_n* (based on *range_n*) are all met.

Now you can use these powerful functions to perform many important computations. The following examples are shown in the New 2007 Functions worksheet. The work is shown in Figure 3-7.

	Q	R	S	T	U	V
1	Promotions on each day of week			Average Cookies Sales by day of week		
2	COUNTIFS	How Many		AVERAGEIF	Mean sales	
3	Monday	23		Monday	482.0962	
4	Tuesday	14		Tuesday	480.6242	
5	Wednesday	17		Wednesday	486.8662	
6	Thursday	13		Thursday	488.6752	
7	Friday	13		Friday	597.3397	
8	Saturday	14		Saturday	654.7179	
9	Sunday	16		Sunday	593.8974	
10						
11	Total Sales					
12	SUMIFS					
13		Cakes	Pies	Cookies	Smoothie	Coffee
14	January	7726	4803	48070	15955	37039
15	February	7089	4507	44886	14692	33514
16	March	7848	4999	50764	16930	36317
17	April	7601	4647	46740	19016	34550
18	May	8360	4846	49865	22434	34396
19	June	7760	4948	47298	24971	34938
20	July	8085	4881	50228	26654	37015
21	August	7898	5130	50674	26724	37784
22	September	7842	4712	49605	19737	34081
23	October	8041	5117	51527	19813	36518
24	November	8045	4955	51617	15463	35713
25	December	8182	5090	50520	16832	36966

Figure 3-7: Monthly sales summary

%4 Part I: Using Excel to Summarizing Marketing Data

Calculating the Number of Promotions on Each Day of the Week

In R3 the formula `COUNTIFS(daywk,Q3,promotion,"promotion")` counts the number of promotions (23) on Monday. Copying this formula to R4:R9 calculates the number of promotions on each day of the week.

Calculating the Average Cookie Sales on Each Day of the Week

The formula in U3 `=AVERAGEIF(daywk,T3,Cookies)` averages the number of cookie sales on Monday. Copying this formula to the range U4:U9 calculates average cookie sales for all other days of the week.

Computing Monthly Sales for each Product

The formula `=SUMIFS(INDIRECT(R$13),Namemonth,$Q14)` in cell R14 computes total cake sales in January (7726).

%4 Place the `INDIRECT` function before the reference to R13 to enable Excel to recognize the word “Cakes” as a range name.

%4 Copy this formula to the range R14:V25. This calculates the total sales for each product during each month.

NOTE Note the `INDIRECT` function makes it easy to copy formulas involving range names. If you do not use the `INDIRECT` function then Excel will not recognize text in a cell such as R13 as a range name.

Computing Average Daily Sales by Month for Each Product

The formula `=AVERAGEIFS(INDIRECT(R$28),Namemonth,$Q29)` in cell R29 computes average cake sales in January. Copying this formula to the range R29:V40 calculates average sales during each month for each product. Figure 3-8 shows this example.

Summarizing Data with Subtotals

The final method to summarize market data discussed in this chapter is with *subtotals*. The *subtotals* feature yields a great looking summary of data. Unfortunately, if new data is added, subtotals are difficult to update. Suppose you want to get a breakdown for each day of the week and for each product showing how sales differ on days with and without promotions (see the *Subtotals Bakery worksheet*). Before computing subtotals you need to sort your data in the order in which you want your subtotals to be computed. In the subtotals you want to see the day of the week, and

then “promotion” or “not promotion.” Therefore, begin by sorting the data in this order. The dialog box to create this sort is shown in Figure 3-9.

	Q	R	S	T	U	V
27	Average Sales					
28	AVERAGEIFS	Cakes	Pies	Cookies	Smoothie	Coffee
29	January	83.0752688	51.64516	516.8817204	171.5591	398.2688
30	February	84.3928571	53.65476	534.3571429	174.9048	398.9762
31	March	84.3870968	53.75269	545.8494624	182.043	390.5054
32	April	84.4555556	51.63333	519.3333333	211.2889	383.8889
33	May	89.8924731	52.10753	536.1827957	241.2258	369.8495
34	June	86.2222222	54.97778	525.5333333	277.4556	388.2
35	July	86.9354839	52.48387	540.0860215	286.6022	398.0108
36	August	84.9247312	55.16129	544.8817204	287.3548	406.2796
37	September	87.1333333	52.35556	551.1666667	219.3	378.6778
38	October	86.4623656	55.02151	554.0537634	213.043	392.6667
39	November	89.3888889	55.05556	573.5222222	171.8111	396.8111
40	December	87.9784946	54.73118	543.2258065	180.9892	397.4839

Figure 3-8: Using AVERAGEIFS to summarize monthly sales

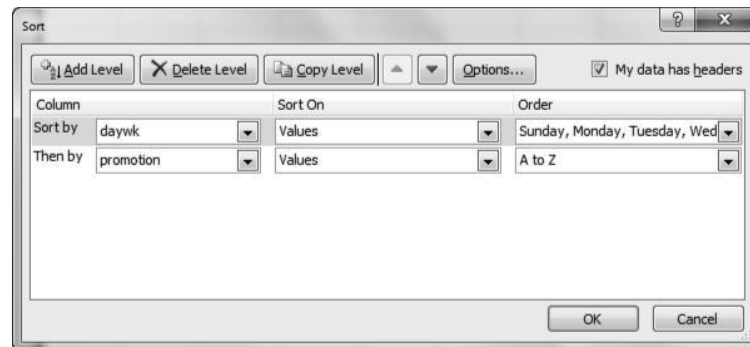


Figure 3-9: Sorting in preparation for subtotals

After performing the sort you see all the Sundays with no promotion, followed by the Sundays with a promotion, then the Mondays with no promotion days, and so on. Perform the following steps to continue:

- %4 Place your cursor anywhere in the data and select Subtotal from the Data tab on the ribbon.
- %4 Compute Subtotals for each product for each day of the week by filling in the Subtotals dialog box, as shown in Figure 3-10.

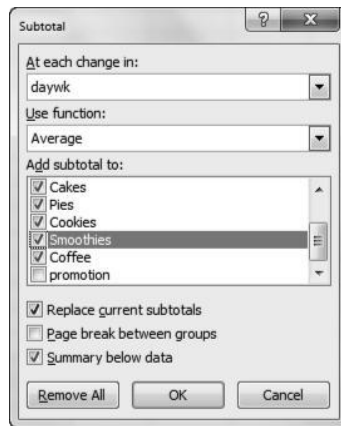


Figure 3-10: Subtotals settings for daily summary

%4 Next you “nest” these subtotals with totals each day of the week for no pro-motion and promotion days. To compute the nested subtotals giving the breakdown of average product sales for each day of the week on promotion and no promotion days, select Subtotal from the Data tab and fill in the dialog box, as shown in Figure 3-11.

%4 Select OK and the Subtotals feature creates for each day of the week total sales for each product. This summary (as shown in Figure 3-12) can be seen by clicking the number 2 in the upper left corner of your screen.

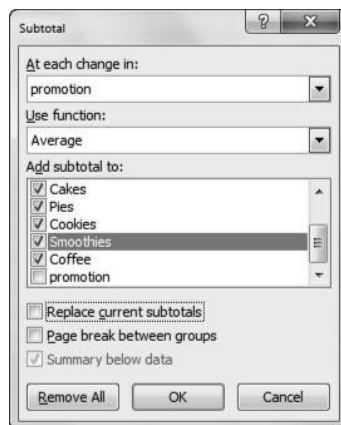


Figure 3-11: Final subtotals dialog box

By unchecking Replace Current Subtotals, you ensure that the subtotals on Promotions will build on and not replace the daily calculations. By clicking the 3 in the upper-left corner of the screen, you can find the final breakdown (see Figure

3-13) of average sales on no promotion versus promotion sales for each day of the week. It is comforting to note that for each product and day of week combination, average sales are higher with the promotion than without.

	D	E	F	G	H	I	J	K	L	M
9										
10		daywk	weekday	Date	Cakes	Pies	Cookies	Smoothies	Coffee	promotion
169		Sunday Average			96.5	59.46795	593.8974	240.2820513	434.5128	
328		Monday Average			78.58333	48.19231	482.0962	195.3653846	353.7564	
488		Tuesday Average			76.93631	46.76433	480.6242	198.5732484	348.3631	
648		Wednesday Average			76.34395	47.75159	486.8662	198.2101911	347.5287	
808		Thursday Average			77.3758	48.10828	488.6752	197.1719745	362.9427	
967		Friday Average			94.67949	59.6859	597.3397	235.0064103	427.0321	
1126		Saturday Average			103.7244	64.98077	654.7179	265.0512821	467.9936	
1127		Grand Average			86.28037	53.54795	540.4511	218.4666667	391.6265	

Figure 3-12: Daily subtotals summary

	D	E	F	G	H	I	J	K	L	M	N
9											
10		daywk	weekday	Date	Cakes	Pies	Cookies	Smoothies	Coffee	promotion	
151					95.37857	58.67857	584.7571	236.0428571	428.4786	none Average	
168					106.3125	66.375	673.875	277.375	487.3125	promotion Average	
169		Sunday Average			96.5	59.46795	593.8974	240.2820513	434.5128		
303					76.04511	47.30827	469.015	194.4135338	342.6241	none Average	
327					93.26087	53.30435	557.7391	200.8695652	418.1304	promotion Average	
328		Monday Average			78.58333	48.19231	482.0962	195.3653846	353.7564		
472					76.13986	46.33566	469.9441	194.8391608	346.014	none Average	
487					85.07143	51.14286	589.7143	236.7142857	372.3571	promotion Average	
488		Tuesday Average			76.93631	46.76433	480.6242	198.5732484	348.3631		
629					75.02143	47.06429	476.9786	195.75	343.2643	none Average	
647					87.23529	53.41176	568.2941	218.4705882	382.6471	promotion Average	
648		Wednesday Average			76.34395	47.75159	486.8662	198.2101911	347.5287		
793					76.56944	47.34722	478.25	194.1597222	360.9722	none Average	
807					86.30769	56.53846	604.1538	230.5384615	384.7692	promotion Average	
808		Thursday Average			77.3758	48.10828	488.6752	197.1719745	362.9427		
952					93.96503	59.07692	589.8392	227.2377622	421.8112	none Average	
966					102.5385	66.38462	679.8462	320.4615385	484.4615	promotion Average	
967		Friday Average			94.67949	59.6859	597.3397	235.0064103	427.0321		
1110					102.1479	63.78169	650.7183	258.028169	459.5775	none Average	
1125					119.7143	77.14286	695.2857	336.2857143	553.3571	promotion Average	
1126		Saturday Average			103.7244	64.98077	654.7179	265.0512821	467.9936		
1127		Grand Average			86.28037	53.54795	540.4511	218.4666667	391.6265		

Figure 3-13: Final subtotals summary

USING EXCEL OUTLINES WITH THE SUBTOTALS FEATURE

Whenever you use the Subtotals feature you will see numbers (in this case 1,2, 3, and 4) in the upper left hand corner of your spreadsheet. This is an example of an Excel outline. The higher the number in the outline, the less aggregated is the data. Thus selecting 4 gives the original data (with subtotals below the relevant data), selecting 3 gives a sales breakdown by day of week and promotion and lack thereof, selecting 2 gives a sales breakdown by day of the week, and selecting 1 gives an overall breakdown of average sales by product. If your cursor is within the original data you can click the Remove All button in the dialog box to remove all subtotals.

Using Array Formulas to Summarize *ESPN The Magazine* Subscriber Demographics

The six wonderful functions previously discussed are great ways to calculate conditional counts, sums, and averages. Sometimes, however, you might want to compute a conditional median, standard deviation, percentile, or some other statistical function. Writing your own array formulas you can easily create your own version of a MEDIANIF, STDEVIF, or other conditional statistical functions.

NOTE The rules for array formulas discussed in “Summarizing Data with a Histogram” also apply to array formulas that you might write.

To illustrate the idea, look at the Chapter 1 data on ESPN subscribers located in the Slicing with arrays worksheet (see also Figure 3-14).

	A	B	C	D	E	F	G	H	I	J	K
1											
2											
3											
4											
5											
6											
7											
8	Age	Gender	Income	Location							
9	72	m	72	rural		77	Median income subscribers >= 50				
10	29	m	68	suburban		80	Median income subscribers < 50				
11	33	m	57	suburban		27.36277	Standard deviation income rural subscribers				
12	25	m	62	suburban							
13	38	m	164	urban							
14	33	m	44	urban							
15	18	f	62	rural							
16	17	m	68	urban							
17	32	f	53	urban							
18	24	f	92	urban							
19	26	f	54	suburban							
20	40	m	88	urban							
21	26	f	46	rural							
22	21	m	83	suburban							
23	29	m	144	rural							
24	51	f	30	suburban							

Figure 3-14: Using arrays to compute conditional medians and standard deviations

Suppose you have been asked to determine the median income for subscribers over 50. To accomplish this goal, perform the following steps:

%4 Array enter in cell F9 the formula `=MEDIAN(IF(Age>=50,Income,""))`. This formula loops through the Income column and creates an array as follows: whenever the Age column contains a number greater than or equal to 50, the array returns the income value; otherwise the array returns a blank. Now the MEDIAN function is simply being applied to the rows corresponding to subscribers who are at least 50 years old. The median income for subscribers who are at least 50 is \$77,000. Note that since your array formula populates a single cell you do not need to select a range containing more than one cell before typing in the formula.

%4 In a similar fashion, array enter in cell F10 the formula `=MEDIAN(IF(Age<50,Income,""))` to compute the median salary (\$80,000) for subscribers who are under 50 years old.

%4 Finally, in cell F11, array enter the formula `=STDEV(IF(Location="rural",Income,""))` to compute (27.36) the standard deviation of the Income for Rural subscribers. This formula produces an array that contains only the Income values for rural subscribers and takes the standard deviation of the values in the new array.

Summary

In this chapter you learned the following:

%4 Using the FREQUENCY array function and the TABLE feature, you can create a histogram that summarizes data and automatically updates to include new data.

%4 The median (computed with the MEDIAN function) is used to summarize a typical value for a highly skewed data set. Otherwise, the mean (computed with the AVERAGE function) is used to summarize a typical value from a data set.

%4 To measure a data set's spread about the mean, you can use either variance (computed with the Excel VAR function) or standard deviation (computed with the Excel STDEV function). Standard deviation is the usual measure because it is in the same units as the data.

%4 Part I: Using Excel to Summarizing Marketing Data

For data sets that do not exhibit significant skewness, approximately 95 percent of your data is within two standard deviations of the mean.

The PERCENTILE.EXC function returns a given percentile for a data set.

The PERCENTRANK.EXC function returns the percentile rank for a given value in a data set.

The LARGE and SMALL functions enable you to compute either the k th largest or k th smallest value in a data set.

The COUNTIF and COUNTIFS functions enable you to count how many rows in a range meet a single or multiple criteria.

The SUMIF and SUMIFS enable you to sum a set of rows in a range that meets a single or multiple criteria.

The AVERAGEIF and AVERAGEIFS functions enable you to average a set of rows in a range that meets a single or multiple criteria.

The SUBTOTALS feature enables you to create an attractive summary of data that closely resembles a PivotTable. Data must be sorted before you invoke the SUBTOTALS feature.

Using array functions you can create formulas that mimic a STDEVIF, MEDIANIF, or PERCENTILEIF function.

Exercises

%4 Exercises 1-6 use the data in the Descriptive stats worksheet of the Chapter3bakery.xlsx workbook. Create a dynamically updating histogram for daily cookie sales.

%4 Are daily cookie sales symmetric or skewed?

%4 Fill in the blank: By the rule of thumb you would expect on 95 percent of all days daily smoothie demand will be between ____ and ____.

%4 Determine the fraction of days for which smoothie demand is an outlier.

%4 Fill in the blank: There is a 10 percent chance daily demand for smoothies will exceed ____.

%4 Fill in the blank: There is a ____ chance that at least 600 cookies will be sold in a day.

%4 Exercises 7 and 8 use the data in the Data worksheet of the ESPN.xlsx workbook. Find the average age of all ESPN subscribers who make at least \$100,000 a year.

%4 For each location and age group (under 25, 25–39, 40–54, 55, and over) determine the fraction of ESPN subscribers for each location that are in each age group.

- %4 Exercises 9-12 use the data in the Descriptive stats worksheet of the Chapter3bakery.xlsx workbook. Determine the percentage of La Petit Bakery cookie sales for each day of the week and month combination. For example, your final result should tell you the fraction of cookie sales that occur on a Monday in January, and so on.
- %4 Determine the average profit earned for each day of the week and month combination. Assume the profit earned by La Petit Bakery on each product is as follows:
- Cakes: \$2
 - Cookie: \$0.50
 - Pie: \$1.50
 - Smoothie: \$1.00
 - Coffee: \$0.80
- %4 Find the median cake sales on days in which at least 500 cookies were sold.
- %4 Fill in the blank: On days in which at least 500 cookies are sold there is a 5 percent chance at least _____ cookies are sold.
- %4 In the years 1980–2012 median U.S. family income (after inflation) has dropped but mean family income has sharply increased. Can you explain this seeming anomaly?
- Compute and interpret the skewness for the three data sets in the file Skewexamples.xlsx.





Pricing

Chapter 4: Estimating Demand Curves and Using Solver to Optimize Price

Chapter 5: Price Bundling

Chapter 6: Nonlinear Pricing

Chapter 7: Price Skimming and Sales

Chapter 8: Revenue Management



4

Estimating Demand Curves and Using Solver to Optimize Price

Understanding how pricing impacts revenues and profitability is one of the most important issues faced by managers. To do so, managers need to understand how consumers' willingness to purchase changes at different price levels and how these changes impact profitability. (That is, managers need to understand the demand curve.) To do so effectively, this chapter covers the following topics:

- %4 Using the Excel Trend Curve and Goal Seek to obtain back-of-the-envelope estimation of a demand curve.
- %4 Using the Excel Solver to determine the profit maximizing price.
- %4 Examining the effect of product tie-ins on the optimal product price.
- %4 Using the SolverTable add-in to quickly price thousands of products!

Estimating Linear and Power Demand Curves

In this section **you learn how to fit the two most frequently used demand curves** (linear and power) to a particular marketing situation. These estimations are used to determine a profit-maximizing price, and to do so you need to know two things:

- %4 The variable cost to produce each unit of the product. (Call this UC .)
- %4 The product's demand curve. Simply put, a demand curve tells you the number of units of your product a customer will demand at each price. In short, if you charge a price of $\$p$ per unit, the demand curve gives you a number $D(p)$, which equals the number of units of your product that will be demanded at price $\$p$. Of course, a firm's demand curve is constantly changing and often depends on factors beyond the firm's control (such as the state of the economy and a competitor's price). Part III, "Forecasting," addresses these factors.

When you know UC and the demand curve, the profit corresponding to a price of $\$p$ is simply $(p - UC) * D(p)$. After you have an equation for $D(p)$, which gives the

%4 Part II: Pricing

quantity of the product demanded for each price, you can use the Microsoft Office Excel Solver feature to find the profit-maximizing price.

Price Elasticity

Given a demand curve, the *price elasticity* for demand is the percentage of decrease in demand resulting from a 1 percent increase in price. When elasticity is larger than 1, demand is price elastic. When demand is price elastic, a price cut will increase revenue. When elasticity is less than 1, demand is price inelastic. When demand is price inelastic, a price cut will decrease revenue. Studies by economists have obtained the following estimates of elasticity:

- %4 **Salt:** 0.1 (very inelastic)
- %4 **Coffee:** 0.25 (inelastic)
- %4 **Legal fees:** 0.4 (inelastic)
- %4 **TV sets:** 1.2 (slightly elastic)
- %4 **Restaurant meals:** 2.3 (elastic)
- %4 **Foreign travel:** 4.0 (very elastic)

A 1 percent decrease in the cost of foreign travel, for example, can result in a 4 percent increase in demand for foreign travel. Managers need to understand the price elasticity at each price point to make optimal pricing decisions. In the next section you will use price elasticity to estimate a product's demand curve.

Forms of Demand Curves

There are multiple forms of demand curves that you can use to analyze marketing data. Using q to represent the quantity demanded of a product, the two most commonly used forms for estimating demand curves are as follows:

- %4 **Linear demand curve:** In this case, demand follows a straight line relationship of the form $q = a - bp$. Here q = quantity demanded and p = unit price. For example, $q = 10 - p$ is a linear demand curve. (Here a and b can be determined by using a method described in the “Estimating a Linear Demand Curve” section of this chapter.) When the demand curve is linear, the elasticity is changing.
- %4 **Power demand curve:** In this situation, the demand curve is described by a power curve of the form $q = ap^b$, $a > 0$, $b < 0$. Again, a and b can be determined by the method described later in the chapter. The equation $q = 100p^{-2}$ is an example of a power demand curve. If demand follows a power curve, for any price, the elasticity equals $-b$. See Exercise 11 for an explanation of this

important property of the power demand curve. Thus, for the demand curve $q = 100p^{-2}$ the price elasticity of demand always equals 2.

Estimating a Linear Demand Curve

Suppose that a product’s demand curve follows a linear demand curve. Given the current price and demand for a product and the product’s price elasticity of demand, determining the product’s demand curve is a simple matter. The following example illustrates how to fit a linear demand curve.

Suppose a product is currently selling for \$100 and demand equals 500 units. The product’s price elasticity for demand is 2. Assuming the demand curve is linear, you can determine the equation of the demand curve. The solution is in the Linearfit.xls file, which is shown in Figure 4-1.

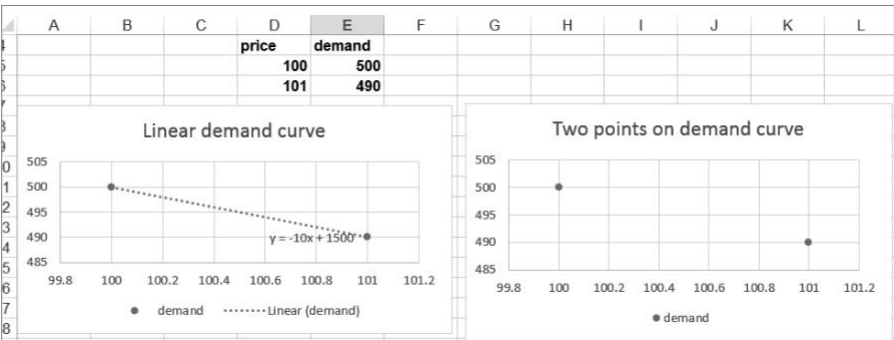


Figure 4-1: Fitting linear demand curve

Given two points, you know that there is a unique straight line that passes through those two points. You actually know two points on your demand curve. One point is $p=100$ and $q=500$. Because elasticity of demand equals 2, a 1 percent increase in price results in a 2 percent decrease in demand. Thus, if $p=101$ (a 1 per-cent increase), demand drops by 2 percent of 500 (10 units) to 490. Thus $p=101$ and $q=490$ is a second point on the demand curve. You can now use the Excel Trendline to find the straight line that passes through the points (100,500) and (101,490).

Begin by entering these points in the worksheet in the cell range D5:E6 (refer to Figure 4-1). Then select the range D4:E6, and on the Ribbon, in the Charts group choose Scatter > Scatter with Only Markers.

- %4 Begin by entering the points in the Linearfit.xls worksheet in the cell range D5:E6 (refer to Figure 4-1).
- %4 Select the range D4:E6.

%4 Part II: Pricing

Go to the Insert tab and in the Charts group choose Scatter > Scatter with Only Markers. You should now see that the graph has a positive slope. This would imply that higher prices lead to higher demand, which cannot be correct. The problem is that with only two data points, Excel assumes that the data points you want to graph are in separate columns, not separate rows.

To ensure Excel understands that the individual points are in separate rows, click inside the graph.

On the Ribbon click the Design tab in the Chart Tools section.

Click Switch Row/Column in the Data section of the Design tab.

NOTE Note that by clicking the Select Data button, you can change the source data that generates your chart.

%4 Next, right-click one of the points and then click Add Trendline.

%4 Click the Linear button and click Display Equation on Chart option.

%4 Click Close in the Format Trendline box.

You see the straight line plot, complete with the equation referred to in Figure 4-1. Because x is price and y is demand, the equation for your demand curve is

%4 $= 1500 - 10p$. This equation means that each \$1 increase in price costs 10 units of demand. Of course, demand cannot be linear for all values of p because for large values of p , a linear demand curve yields negative demand. For prices near the current price, however, the linear demand curve is usually a good approximation to the product's true demand curve.

Estimating a Power Demand Curve

Recall that for a linear demand curve the price elasticity is different for each price point. If the marketing analyst believes that elasticity remains relatively constant as price changes, then she can use a power demand curve (which has constant price elasticity) to model demand for a product.

Again assume that a product is currently selling for \$100 and demand equals 500 units. Assume also that the product's price elasticity for demand is known to equal 2. (In Chapters 31 and 32 you will learn some more advanced methods to estimate price elasticity). You can fit a power demand curve to this information by performing the following steps in the Powerfit.xls file, as shown in Figure 4-2.

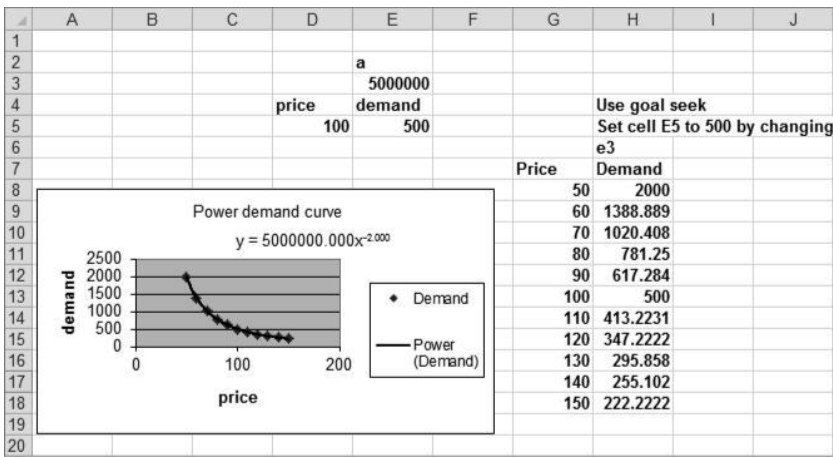


Figure 4-2: Fitting Power Demand Curve

- %4 After naming cell E3 as a , enter a trial value for a
- %4 In cell D5, enter the current price of \$100. Because elasticity of demand equals 2, you know that the demand curve has the form $q=ap^{-2}$ where a is unknown.
- %4 In cell E5, enter a formula that computes the demand when the unit price equals \$100. The demand is linked to your choice of the value of a in cell E3 with the formula $a*D5^2$
- %4 Now use the Goal Seek command to determine the value of a . This formula makes the demand for price \$100 equal to 500 units. Goal Seek enables you to find a value of a cell in a spreadsheet (called the By Changing Cell) that makes a formula (called the Set Cell) hit a wanted Value. In the example you want to change cell E3 so that the formula in cell E5 equals 500. Set cell E5 to the value of 500 by changing cell E3.
- %4 To invoke Goal Seek, switch to the Data tab and select What-If analysis from the Data Tools Group, then choose Goal Seek from the drop down arrow. Fill the dialog box as shown in Figure 4-3. With these settings, Excel changes the changing cell (E3) until the value of the set cell (E5) matches your desired value of 500.

A value for $a = 5$ million yields a demand of 500 at a price of \$100. Thus, the demand curve (refer to Figure 4-2) is given by $q = 5,000,000p^{-2}$. For any price, the price elasticity of demand on this demand curve equals 2.

%4 Part II: Pricing

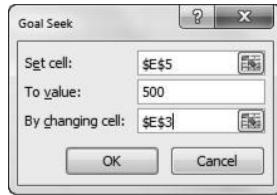


Figure 4-3: Goal Seek dialog box

Using the Excel Solver to Optimize Price

Often in marketing you want to **maximize profit, minimize, or optimize** some other objective. The *Excel Solver* is a powerful tool that you can use to solve many marketing (and other!) optimization problems. The Solver is an add-in. To activate the Solver proceed as follows:

For Excel 2010 or Excel 2013:

- %4 Select File and then Options.
- %4 Select Add-Ins, click Go, check the Solver add-in, and select OK.
- %4 Now click the Data tab and you should see the Solver add-in on the right side of the Ribbon.

For Excel 2007:

- %4 Click the Office Button (the oval in the left side of the ribbon) and choose Excel Options.
- %4 Selecting Add-Ins, click Go, check the Solver add-in, and select OK.
- %4 Now click the Data tab and you should see the Solver add-in on the right side of the Ribbon.

In this chapter the examples work with the Excel 2010 Solver, which is more powerful than the Solver included with previous versions of Excel. If you select Solver from the Data tab, the Solver window appears, as shown in Figure 4-4.

To define a Solver model, you must specify in the Solver dialog box the following three parts of the model:

- %4 **Set Objective or Target Cell:** The objective cell contains the goal you want to maximize (like profit) or minimize (production cost).

- %4 **Changing Variable Cells:** These cells are the cells that you can change or adjust to optimize the target cell. In this chapter the changing cells will be each product's price.
- %4 **Constraints:** These are restrictions on the changing cells. For example, you might want to constrain the price for each of your products to be within 10 percent of the competitor's price.



Figure 4-4: Solver window

The Excel Solver has been greatly revamped and improved in Excel 2010. The primary change is the presence of the Select a Solving Method drop-down list. From this list you must select the appropriate solution engine for your optimization problem. You can choose from the following options:

- %4 The *Simplex LP engine* is used to solve linear optimization problems. A linear optimization problem is one in which the target cell and constraints are all created by adding together terms of the form *(changing cell)*(constant)*. Most

%4 Part II: Pricing

marketing models are not linear. An exception is the classic advertising media selection model discussed in Chapter 35, “Media Selection Models.”

%4 The *GRG Nonlinear engine* is used to solve optimization problems in which the target cell and/or some of the constraints are not linear and are computed by using typical mathematical operations such as multiplying or dividing changing cells, raising changing cells to a power, using exponential or trig functions involving changing cells, and so on. The GRG engine includes a powerful Multistart option that enables users to solve many problems that were solved incorrectly with previous versions of Excel. The Multistart option will be used extensively throughout this book.

%4 The *Evolutionary engine* is used when your target cell and/or constraints contain nonsmooth functions that reference changing cells. For example, if your target cell and/or constraints contain IF, SUMIF, COUNTIF, SUMIFS, COUNTIFS, AVERAGEIF, AVERAGEIFS, ABS, MAX, or MIN functions that reference the changing cells, then the Evolutionary engine probably has the best shot at finding a good solution to your optimization problem. The Evolutionary engine is extensively used throughout this book.

After you have input the target cell, changing cells, and constraints, what does Solver do? A set of values for the changing cells is a *feasible solution* if it meets all constraints, and the Solver essentially searches through all feasible solutions and finds the set of feasible solution changing cell values (called the *optimal solution*) that has the best value for the target cell (largest in a maximization and smallest in a minimization). If there is more than one optimal solution, the Solver stops at the first one it finds.

Pricing Razors (No Blades!)

Using the techniques described in the “Estimating a Linear Demand Curve” section, it’s easy to determine a demand curve for the product that’s originally purchased. You can then use the Microsoft Office Excel Solver to determine the original product price that maximizes the sum of the profit earned from razors. Then you can show how the fact that purchasers of razors also buy blades reduces the profit maximizing price for razors.

Suppose that you currently charge \$5.00 for a razor and you sell 6 million razors. Assume that the variable cost to produce a razor is \$2.00. Finally, suppose that the

price elasticity of demand for razors is 2, and the demand curve is linear. What price should you charge for razors?

You can determine a demand curve (assuming a linear demand curve), as shown in Figure 4-5. (You can find this data and the chart on the no blades worksheet in the file *razorsandblades.xls*.) Two points on the demand curve are price = \$5.00, demand = 6 million razors and price = \$5.05 (an increase of 1 percent), demand = 5.88 million (2 percent less than 6 million).

%4 Begin by drawing a chart and inserting a linear trend line, as shown in the section “Estimating a Linear Demand Curve.” You’ll find the demand curve equation is $y = 18 - 2.4x$. Because x equals price and y equals demand, you can write the demand curve for razors as follows:

$$\text{Demand (in millions)} = 18 - 2.4(\text{price})$$

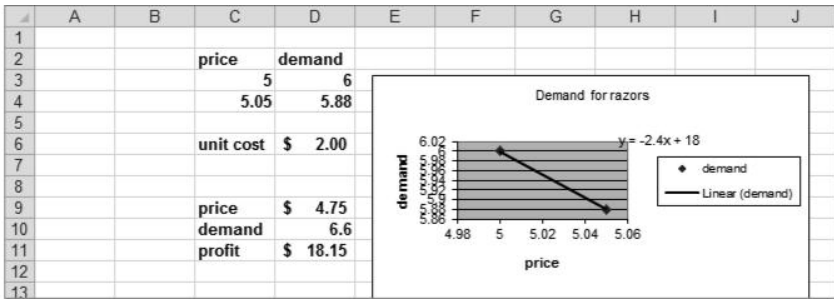


Figure 4-5: Optimizing razor price: no blades

- %4 Associate the names in cell C6 and the range C9:C11 with cells D6 and D9:D11.
- %4 Next, enter a trial price in D9 to determine demand for that price in cell D10 with the formula $=18 - 2.4 * \text{price}$
- %4 Determine in cell D11 the profit for razors by using the formula $=\text{demand} * (\text{price} - \text{unit\_cost})$
- %4 You can now use Solver to determine the profit-maximizing price. The Solver Parameters dialog box is shown in Figure 4-6. Choose to maximize the profit cell (cell D11) by changing the price (cell D9).
- %4 The model is not linear because the target cell multiplies together two quantities—*demand* and *(price–cost)*—each depending on the changing cell. Therefore choose the GRG Nonlinear option. Solver finds that by charging \$4.75 for a razor, you can maximize the profit. (The maximum profit is \$18.15 million.)



Figure 4-6: Solver window for razor price: no blades

Incorporating Complementary Products

Certain consumer product purchases frequently result in the purchase of related products, or *complementary products*. Table 4-1 provides some examples:

Table 4-1: Examples of Complementary Product

Original Purchase	Tie-in Complementary Product
Men's suit	Tie or shirt
Inkjet printer	Printer cartridge
Xbox console	Video game
Cell phone	Case

If the profit from complementary products is included in the target cell, the profit maximizing price for the original product will decrease. Suppose that the average purchaser of a razor buys 50 blades and you earn \$0.15 of profit per blade

purchased. You can use the Excel Solver to determine how this changes the price you should charge for a razor. Assume that the price of a blade is fixed. (In Exercise 3 at the end of the chapter, the blade price changes.) The analysis is in the blades worksheet of the razorsandblades.xls file, as shown in Figure 4-7.

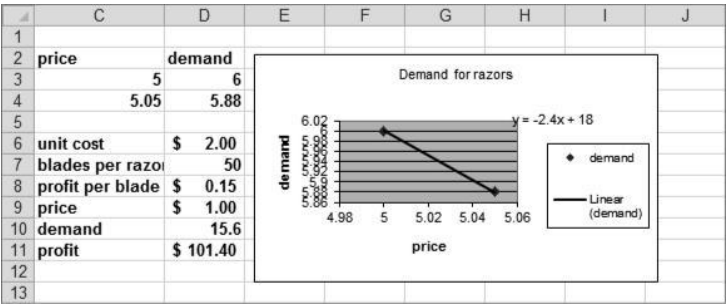


Figure 4-7: Optimizing razor price: blade profit included

To maximize profit perform the following steps:

- %4 Use the Create from Selection command in the Defined Names group on the Formulas tab to associate the names in cells C6:C11 with cells D6:D11. (For example, cell D10 is named Demand.)

NOTE Recall that cell D10 of the no blades worksheet is also named Demand. When you use the range name Demand in a formula Excel simply refers to the cell named Demand in the current worksheet. In other words, when you use the range name Demand in the blades worksheet, Excel refers to cell D10 of that

- %4 In cells D7 and D8, enter the relevant information about blades.
- %4 In D9, enter a trial price for razors, and in D10, compute the demand with the formula $18 - 2.4 * \text{price}$
- %4 Next, in cell D11, compute the total profit from razors and blades with the following formula:

$$\text{demand} * (\text{price} - \text{unit\_cost}) + \text{demand} * \text{blades\_per\_razor} * \text{profit\_per\_blade}$$

Here the $\text{demand} * \text{blades\_per\_razor} * \text{profit\_per\_blade}$ is the profit from blades.

- %4 The Solver setup is exactly as shown earlier in Figure 4-6: Change the price to maximize the profit. Of course, now the profit formula includes the profit earned from blades.

%4 Part II: Pricing

Profit is maximized by charging only \$1.00 (half the variable cost!) for a razor. This price results from making so much money from the blades. You are much better off ensuring that many people have razors even though you lose \$1.00 on each razor sold. Many companies do not understand the importance of the profit from tie-in products. This leads them to overprice their primary product and not maximize their total profit.

Pricing Using Subjectively Estimated Demand Curves

In situations when you don't know the price elasticity for a product or don't think you can rely on a linear or power demand curve, a good way to determine a product's demand curve is to identify the lowest price and highest price that seem reasonable. You can then try to estimate the product's demand with the high price, the low price, and a price midway between the high and low prices. This approach is based on a discussion in the book *Power Pricing*, by Robert Dolan. Given these three points on the product's demand curve, you can use the Microsoft Office Excel Trendline feature to fit a quadratic demand curve with the following equation:

$$(1) \text{ Demand} = a(\text{price})^2 + b(\text{price}) + c$$

Fitting a quadratic demand curve in this manner enables the slope of the demand curve to either become steeper or flatter, which is much more realistic than the linear demand curve that requires the slope to remain constant.

For any three specified points on the demand curve, values of a , b , and c exist that makes Equation 1 *exactly* fit the three specified points. Because Equation 1 fits three points on the demand curve, it seems reasonable to believe that the equation can give an accurate representation of demand for other prices. You can then use Equation 1 and Solver to maximize profit, which is given by the formula (price– unit cost)*demand. The following example shows how this process works.

Suppose that a drugstore pays \$0.90 for each unit of ChapStick it orders. The store is considering charging from \$1.50 through \$2.50 for a unit of ChapStick. It thinks that at a price of \$1.50, it can sell 60 units per week. (See the ChapStickprice.xls file.) At a price of \$2.00, it thinks it can sell 51 units per week and at a price of \$2.50, 20 units per week. To determine what price the store should charge for ChapStick, perform the following steps.

- %4 Begin by entering the three points with which to chart your demand curve in the cell range E3:F6.
- %4 Select E3:F6, click the Scatter option on the Charts group on the Ribbon, and then select the first option for a Scatter chart.

%4 Right-click a data point and select Add Trendline.

%4 In the Format Trendline dialog box, choose Polynomial, and select 2 in the Order box to obtain a quadratic curve of the form of Equation 1. Then select the option Display Equation on Chart. Figure 4-8 shows the required Trend Curve Settings. The chart containing the demand curve is shown in Figure 4-9. The estimated demand curve Equation 2 is as follows:

$$\text{Demand} = -44 * \text{Price}^2 + 136 * \text{Price} - 45$$

%4 Next, insert a trial price in cell I2. Compute the product demand by using Equation 2 in cell I3 with the formula $=-44*\text{price}^2+136*\text{price}-45$. (Cell I2 is named Price.)

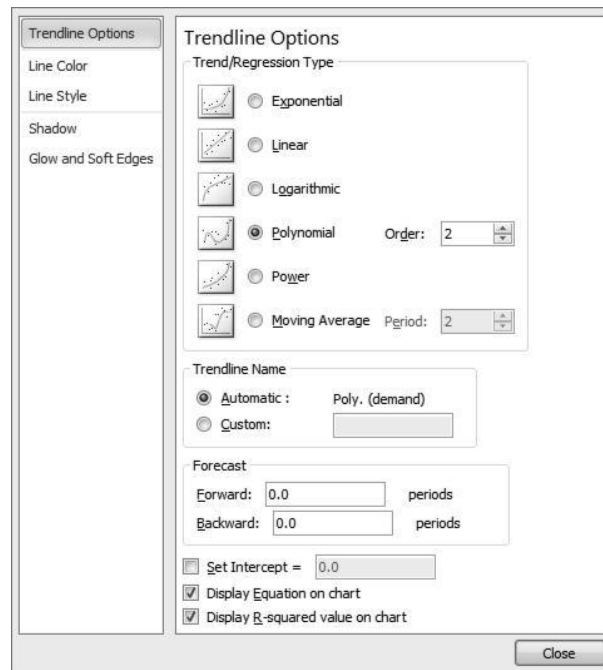


Figure 4-8: Pricing with a quadratic demand curve

%4 Compute the weekly profit from ChapStick sales in cell I4 with the formula $=\text{demand}*(\text{price}-\text{unit\_cost})$. (Cell E2 is named Unit_Cost and cell I3 is named Demand.)

%4 Use Solver to determine the price that maximizes profit. The Solver Parameters dialog box is shown in Figure 4-10. The price is constrained to be from the lowest through the highest specified prices (\$1.50 through \$2.50). If you allow Solver to consider prices outside this range, the quadratic demand curve

%4 Part II: Pricing

might slope upward, which implies that a higher price would result in larger demand. This result is unreasonable, which is why you constrain the price.

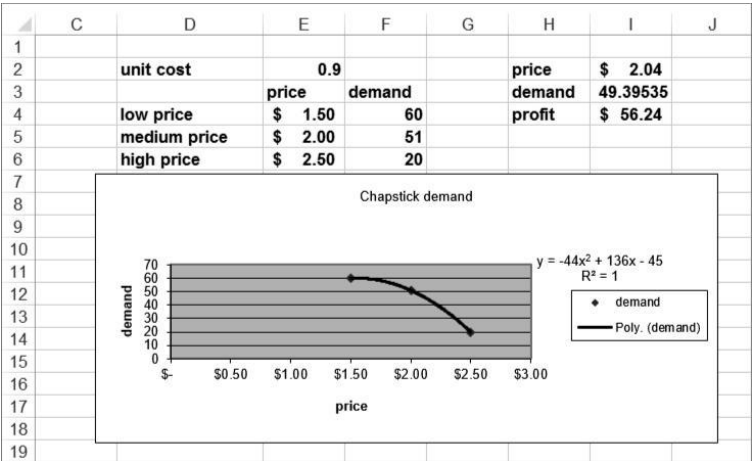


Figure 4-9: ChapStick demand curve



Figure 4-10: Solver window for quadratic demand curve example

You find that the drugstore should charge \$2.04 for a ChapStick unit. This yields sales of 49.4 units per week and a weekly profit of \$56.24.

NOTE The approach to pricing outlined in this section requires no knowledge of the concept of price elasticity. Inherently, the Solver considers the elasticity for each price when it determines the profit-maximizing price.

NOTE For the quadratic demand model to be useful, the minimum and maximum prices must be consistent with consumer preferences. A knowledgeable sales force should be able to come up with realistic minimum and maximum prices.

Using SolverTable to Price Multiple Products

The approach developed to price a product in the preceding section can be extended to enable a company to easily price hundreds or thousands of products. The only information required for each product is the unit cost, estimated demand for lowest possible price, estimated demand for an intermediate price, and estimated demand. Using the SolverTable add-in written by Chris Albright (available for download at www.kelley.iu.edu/albright/Free_downloads.htm) you can easily fit the quadratic demand curve to each product. SolverTable is an Excel add-in that enables you to easily vary the inputs to a Solver model and track a wanted set of outputs.

After fitting the demand curves, you can use Solver to set a price for each product to maximize the total product generated from all products by completing the following steps (the Data worksheet in the *FittingmultipleDemandCurves.xls* file shows the estimated demand for three products at a low (\$1.10), medium (\$1.30), and high (\$1.50) price):

- %4 Use an HLOOKUP function that keys off the entry in cell A11 to place the demands for each product in E14:E16. The Solver model chooses constant a , a coefficient b for price, and a coefficient c for price² that exactly passes through the demand points for each product.
- %4 Use SolverTable to loop through each product by changing the value in cell A11 in the range 1, 2, ..., where n = number of products.
- %4 Name the range F6:H10 Lookup.
- %4 Copy the formula =HLOOKUP(\$A\$11,Lookup,C14) from E14 to E15:E16 to pull the demand for each product corresponding to the product index in A11.
- %4 Copy the formula =E\$3+E\$4*D14+E\$5*(D14^2) from F14 to F15:F16 to compute the forecasted demand based on the quadratic demand curve based on the values in E3:E5.

%4 Part II: Pricing

Copy the formula $= (E14 - F14)^2$ from G14 to G15:G16 to compute the squared error in the demand forecast for each price. For each price the estimation error is simply actual demand minus demand estimated from quadratic demand curve.

In cell G12 use the formula $= \text{SUM}(G14:G16)$ to compute the sum of the squared estimation errors.

If you minimize G12 by changing E3:E5, Solver finds the values of a , b , and c that make the sum of squared errors equal to 0. You minimize the sum of squared errors instead of minimizing the sum of errors because if you do not square errors, then the sum of positive and sum of negative errors cancels each other out. Minimizing the sum of squared errors mimics the action of the Trend Curve Polynomial option described in the preceding section. The appropriate Solver Window is shown in Figure 4-11.

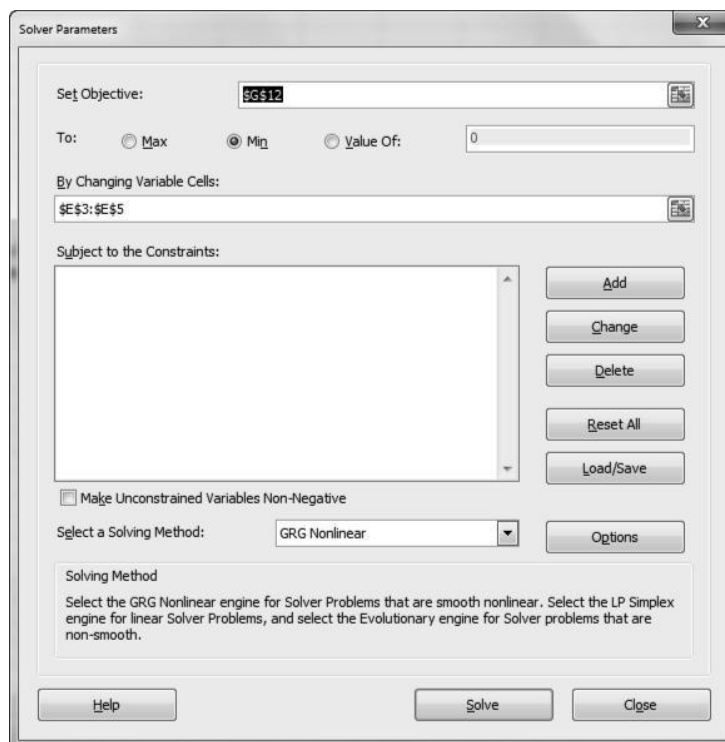


Figure 4-11: Solver window for SolverTable example

- %4 Choose the GRG Nonlinear option because the target cell is not constructed by adding up terms of the form (changing cell)*constant.
- %4 For Solver to obtain a correct solution to this problem, go to Options and check Use Automatic Scaling; then go to the GRG tab and select Central Derivatives.
- %4 Select Solve and you find the answer (see Figure 4-12). The Sum of Squared errors is 0 to 26 decimal places and the quadratic demand curve is $73.625 + 195\text{price} + 87.5\text{price}^2$. Because a , b , and c can be negative, do not check the Make Unconstrained Variables Non-Negative box.

	A	B	C	D	E	F	G	H
1								
2								
3				a	-73.625			
4				b	195			
5				c	-87.5			
6						1	2	3
7					Price	Product 1 demand	Product 2 demand	Product 3 demand
8				Low	1.1	35	32	24
9				Medium	1.3	32	27	17
10	Data set			High	1.5	22	16	9
11	1						SSE	
12							1.00974E-27	
13				Price	Demand	Forecast	Squared Error	
14			3	1.1	35	35	2.01948E-28	
15			4	1.3	32	32	8.07794E-28	
16			5	1.5	22	22	0	

Figure 4-12: Use of SolverTable for price optimization

Using SolverTable to Find the Demand Curve for All Products

After you have a Solver model in a worksheet, you can use SolverTable to “loop through” the values of one (in a One-way SolverTable) or two (in a Two-way SolverTable) inputs and track any wanted outputs. To get started, assuming you have installed SolverTable, select SolverTable from the Ribbon, choose a One-way Table, and fill in the SolverTable dialog box, as shown in Figure 4-13.

To begin Solver place a 1 in the input cell (A11) and run the Solver to track the output cells (a , b , and c and Sum of Squared Errors [SSE]). Then Solver places a 2 in A11 and finally a 3 in A11. The results are placed in a new worksheet (STS_1), as shown in Figure 4-14.

%4 Part II: Pricing

Parameters for oneway table

Specify the following information about the input to be varied and the outputs to be captured.

Input cell:

(Optional) Descriptive name for input:

Values of input to use for table

☒ Base input values on following:

Minimum value:

Maximum value:

Increment:

☐ Use the values from the following range:

Input value range:

☐ Use the values below (separate with commas)

Input values:

Output cell(s):

Note about specifying output cells: The safest way to select multiple output cells or ranges is to put your finger on the Ctrl key and then drag as many output cell ranges as you like. This will automatically insert commas between the ranges you select.

Figure 4-13: SolverTable settings

	A	B	C	D	E	F	G	H	I
1	Oneway analysis for Solver model in data worksheet								
2									33.23375
3	Input (cell \$A\$11) values along side, output cell(s) along top								
4					\$G\$12	Price	demand	unit cost	Profit
5	1	-73.625	195	-87.5	1.04441E-22	1.389493	28.3907	0.8	16.73611
6	2	-47.75	155	-75	6.31897E-25	1.370568	23.80379	0.9	11.2013
7	3	44.625	-5	-12.5	7.98128E-21	1.367619	14.40715	1	5.296336

Figure 4-14: SolverTable results

You can now set up a Solver model that determines the profit maximizing price for each product. Proceed as follows:

- %4 Enter trial prices for each product in the range F5:F7.
- %4 Based on the prices in F5:F7, compute the demand for each product by copy-ing the formula $=B5+C5*F5+F5^2*D5$ from G5 to G6:G7.
- %4 Enter the unit cost for each product in H5:H7, and compute the profit for each product as $(price-unit\ cost)*demand$ by copying the formula $=(F5-H5)*G5$ from I5 to I6:I7.
- %4 Compute the total profit for all products in cell I2 with the formula $=SUM(I5:I7)$

%4 Now use Solver (see Figure 4-15) to choose the prices in F5:F7 to maximize total profit (I2). This, of course, chooses the profit maximizing price for each product.



Figure 4-15: Finding profit maximizing price for each product

You find that you should charge \$1.39 for Product 1 and \$1.37 for Products 2 and

%4 The SolverTable provides a powerful tool that can enable the marketing analyst to quickly estimate a demand curve for several products and determine a profit maximizing price for each product.

Summary

In this chapter you learned the following:

- %4 Price elasticity of demand is the percentage decrease in demand for a 1 percent increase in price.
- %4 Given the current price and demand and price elasticity, you can use the Excel Trend Line feature to fit a linear demand curve ($Demand = a - b * price$).

%4 Part II: Pricing

For a linear demand curve price, elasticity is different at every point on the demand curve.

%4 Given the current price and demand and price elasticity, you can use the Excel Goal Seek feature to fit a power or constant elasticity demand curve ($Demand = a(price)^{-b}$)

%4 Given a demand curve you can use Solver to find a profit maximizing price by maximizing $(price - unit\ cost) * demand$

%4 If you do not know the price elasticity, you can use the Polynomial portion of the Trend Line feature to fit a quadratic to three points on the demand curve.

%4 Using the SolverTable add-in you can quickly fit the demand curve for a huge number of products and then use Solver to find the profit maximizing price for each product.

Exercises

%4 **Your company** charges \$60 for a board game it invented and has sold 3,000 copies during the last year. Elasticity for board games is known to equal 3. Use this information to determine a linear and power demand curve.

%4 You need to determine the profit-maximizing price for a video game console. Currently you charge \$180 and sell 2 million consoles per year. It costs \$150 to produce a console, and the price elasticity of demand for consoles is 3. What price should you charge for a console?

%4 Now assume that, on average, a purchaser of your video game console buys 10 video games, and you earn \$10 profit on each video game. What is the correct price for consoles?

%4 In the `razorsandblades.xls` file example, suppose the cost to produce a blade is \$0.20. If you charge \$0.35 for a blade, a customer buys an average of 50 blades from you. Assume the price elasticity of demand for blades is 3. What price should you charge for a razor and for a blade?

%4 You manage a movie theater that can handle up to 8,000 patrons per week. The current demand, price, and elasticity for ticket sales, popcorn, soda, and candy are given in Figure 4-16. The theater keeps 45 percent of ticket revenues. Unit cost per ticket, popcorn sales, candy sales, and soda sales are also given. Assuming linear demand curves, how can the theater maximize profits? Demand for foods is the fraction of patrons who purchase the given food.

	B	C	D	E	F	G	H	I
2								
3			elasticity	current price	demand	cost	ticket	percentage
4	keep 45%	ticket	3	8	3000	0	0.45	
5		popcorn	1.3	3.5	0.5	0.4		
6		soda	1.5	3	0.6	0.6		
7		candy	2.5	2.5	0.2	1		

Figure 4-16: Data for Exercise 5

%4 A prescription drug is produced in the United States and sold internationally. Each unit of the drug costs \$60 to produce. In the German market, you sell the drug for 150 euros per unit. The current exchange rate is 0.667 U.S. dollars per euro. Current demand for the drug is 100 units, and the estimated elasticity is 2.5. Assuming a linear demand curve, determine the appropriate sales price (in euros) for the drug.

%4 Suppose it costs \$250 to produce a video game console. A price between \$200 and \$400 is under consideration. Estimated demand for the game console is shown in the following table. What price should you charge for game console?

Price	Demand
\$200	50,000
\$300	25,000
\$400	12,000

%4 Use the demand information given in Exercise 7 for this exercise. Each game owner buys an average of 10 video games. You earn \$10 profit per video game. What price should you charge for the game console?

%4 You want to determine the correct price for a new weekly magazine. The variable cost of printing and distributing a copy of the magazine is \$0.50. You are thinking of charging from \$0.50 through \$1.30 per copy. The estimated weekly sales of the magazine are shown in the following table. What price should you charge for the magazine?

Price	Demand (in Millions)
\$0.50	2
\$0.90	1.2
\$1.30	.3

%4 Part II: Pricing

Given the following information in the table for four products, find the profit-maximizing price for each product.

Product	Lowest Price	Medium Price	High Price	Low Price Demand	Medium Price Demand	High Price Demand	Unit cost
Product 1	\$1.40	\$2.20	\$3.00	100	40	5	\$1.10
Product 2	\$2.20	\$3.00	\$4.00	200	130	25	\$1.50
Product 3	\$45	\$75	\$95	400	300	130	\$40
Product 4	\$200	\$250	\$300	600	300	50	\$120

%4 (Requires calculus) Show that if the demand for a product is given by the power curve $q = ap^{-b}$, then for any price a , a 1 percent increase in price will decrease demand by 1 percent.

%4 For the demand curve $q = 100p^{-2}$ show that for $p = 1, 2, 4, 8$, and 16 that a 1 percent increase in price will result in approximately a 2 percent decrease in demand.

5

Price Bundling

Chapter 4, “Estimating Demand Curves and Using Solver to Optimize Price,” discussed situations in which a customer pays the same price for each unit she purchases of a product. When a customer pays the same price for each unit the seller is using *linear pricing*. This chapter and Chapter 6, “Nonlinear Pricing,” describe some *nonlinear pricing* models in which the total amount a customer pays for a set of products is not equal to the sum of the individual product prices. One of the most common instances of nonlinear pricing is price bundling. This chapter shows that analytic models can help companies use bundling to maximize their profits.

Why Bundle?

Companies often bundle products in an attempt to get customers to purchase more products than they would have without bundling. A few examples of bundling include the following:

- %4 Cable companies bundle landlines, cell phone service, TV service, and Internet service.
- %4 Automobile companies often bundle popular options such as navigation, sat-ellite radio, and keyless entry.
- %4 Computer mail order companies often bundle computers with printers, scan-ners, and monitors.
- %4 Microsoft Office has been a highly successful bundling of software products such as Excel, Word, Access and Outlook.

In this chapter you will make the assumption that customers make decisions based on the concept of *consumer surplus*. The consumer surplus of a product is simply the value a consumer attaches to an available product minus the actual cost of the product.

%4 Part II: Pricing

NOTE In Chapters 16, “Conjoint Analysis” and 18, “Discrete Choice Analysis,” you learn how to determine the value a consumer attaches to a particular product or combination of products.

The consumer’s goal is to make a choice that maximizes her consumer surplus. Therefore, if each product combination has a negative surplus, no product combination is purchased; however, as long as at least one option has a non-negative consumer surplus, the consumer will choose the option with the largest consumer surplus. The possibility of ties can be ignored. This section examines three examples that portray how bundling can increase profits.

Bundling Products to Extract Consumer Surplus

Suppose you own Disney World and each customer values the five rides in the park as follows:

- %4 Space Mountain: \$10
- %4 Matterhorn: \$8
- %4 Tower of Terror: \$6
- %4 Small World: \$4
- %4 Mr. Toad: \$2

How can Disney World maximize revenue from these rides? First assume a single price is charged for each ride. If Disney World charges \$10 a ride, each customer would go on only Space Mountain and you make \$10 per customer. If Disney World charges \$8 per ride, each customer would go on two rides, and you make \$16. If you charge \$6 per ride, then each customer goes on three rides and you make \$18. If you charge \$4 per ride, then each customer goes on four rides and you make \$16. Finally, charging \$2 per ride, you make \$10 in revenue. Therefore with a single price for each ride, revenue is maximized at \$18 per customer.

Now suppose Disney World does not offer per ride tickets but simply offers a five-ride ticket for \$30. Because the consumer values all five rides at \$30 (the sum of the ride values) he will pay \$30 for park admission. This is a 67 percent increase in revenue. Of course, single park admission is the strategy Disney World has adapted. Single park admission has other benefits such as reducing lines, but this simple exam-ple shows that product bundling can help companies “extract” consumer surplus from customers. Later in the chapter you see that the assumption that all customers are identical is not important to your analysis.

Pure Bundling

Now consider another example. Suppose you sell two products and have two potential customers that value each product, as shown in Table 5-1:

Table 5-1: Pure Bundling Works Well with Negatively Correlated Valuations

Product	Customer 1	Customer 2
Computer	\$1,000	\$500
Monitor	\$500	\$1,000

For simplicity, assume the cost of producing each product is negligible and can be assumed to equal \$0. Therefore revenue and profit are identical. In this example customer valuations are negatively correlated; Customer 1 values the computer more than Customer 2 but values the monitor less. If you price each product separately, you would charge \$1,000 for each product and make \$2,000 revenue; however, if you charge \$1,500 for a bundle of both products, you make \$3,000. Thus, when customer valuations for products are negatively correlated, bundling can result in a significant increase in profit.

The reason bundling works in this case is that bundling enables the seller to entirely extract the value the consumers attach to the products. If the seller only offers the customer a choice between purchasing all products or nothing, the situation is called *pure bundling*. Movie rental companies usually give theatres a choice between renting an assortment of some blockbuster movies and some not-so-popular movies, or renting no movies at all. In this case if the theatre owner wants the blockbuster movies she has to also rent the less-popular movies.

If customer valuations are positively correlated then bundling usually offers no benefit, because bundling does not allow the seller to extract more surplus than separate prices. For example, in Table 5-2 valuations are positively correlated because Customer 2 is willing to pay more for each product than Customer 1. In this case, if the seller offers only separate prices she would charge \$1,000 for the computer and \$500 for the monitor. Then each customer buys both products and total revenue is \$3,000. If the seller offers a pure bundle she should charge \$1,500 for the bundle. Then both customers purchase the bundle for \$1,500 yielding a total revenue of \$3,000. In this case pure bundling and separate prices yield identical profits.

%4 Part II: Pricing

Table 5-2: Pure Bundling Works Poorly with Positively Correlated Valuations

Product	Customer 1	Customer 2
Computer	\$1,000	\$1,250
Monitor	\$500	\$750

Mixed Bundling

Mixed bundling means the seller offers a different price for each available combination of products. For the data in Table 5-3 mixed bundling is optimal.

Table 5-3: Mixed Bundling is Optimal

Customer	Computer	Monitor
1	\$900	\$800
2	\$1,100	\$600
3	\$1,300	\$400
4	\$1,500	\$200
Unit Cost	\$1,000	\$300

To maximize profits, the seller can do the following (see also Exercise 7):

- %4 To maximize profit from separate prices the seller charges \$600 for the monitor and \$1,300 for the computer. In this scenario Customers 3 and 4 buy the computer and Customers 1 and 2 buy the monitor yielding a profit of \$1,200.
- %4 To maximize profit with mixed bundling the seller can charge \$799 for the monitor, \$1,499 for the computer, and \$1,700 for the bundle. Then Customer 1 buys just a monitor, Customer 4 buys just a computer, and Customers 2 and 3 buy the product bundle. The seller earns a profit of \$1,798.
- %4 If only the monitor and computer pure bundle is offered, the seller should sell the pure bundle for \$1,700. Then each customer will purchase the pure bundle and the seller earns a profit of \$1,600.

Using Evolutionary Solver to Find Optimal Bundle Prices

When you combine the consumer surplus decision-making assumptions explained in the previous section with the power of the Evolutionary Solver Engine of Excel 2010 or later, you can easily solve complex bundling problems.

For this section's example, suppose Verizon sells cell phone service, Internet access, and FIOS TV service to customers. Customers can buy any combination of these three products (or not buy any). The seven available product combinations include the following:

- %4 Internet (Combination 1)
- %4 TV (Combination 2)
- %4 Cell phone (Combination 3)
- %4 Internet and TV (Combination 4)
- %4 Internet and cell phone (Combination 5)
- %4 TV and cell phone (Combination 6)
- %4 All three products (Combination 7)

The file phone.xls gives the amount 77 representative customers are willing to pay per month for each service. Use the model in the initial solver worksheet to follow along with the example (also shown in Figure 5-1). In row 6 you see that the first customer is willing to pay up to \$3.50 per month for Internet, \$7 per month for TV service, and \$3.50 per month for cell phone service. (It's hard to make money off this customer!) You can use the Evolutionary Solver Engine and the willingness to pay data to determine a price for each product combination that maximizes revenue.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3																
4			Product	1	2	3	4	5	6	7		total	3413.696			
5			Price	74.35303	35	82.16	69.9942	69.99101	69.991	89.945						
6	Internet	TV	Cell Phon	Internet	TV	Cell	+TV	+Cell	Internet	TV+Cell	All 3	max surp	bought?	revenue		
7	\$3.50	\$7.00	\$3.50	-70.853	-28	-78.7	-59.494	-62.991	-59.491	-75.945	-27.9987	0	0			
8	\$17.50	\$35.00	\$3.50	-56.853	0	-78.7	-17.494	-48.991	-31.491	-33.945	0.0013	2	34.9987			
9	\$28.00	\$28.00	\$49.00	-46.353	-7	-33.2	-13.994	7.008987	7.0087	15.055	15.05482	7	89.94518		Product	Frequency
10	\$70.00	\$70.00	\$0.00	-4.35303	35	-82.2	70.0058	0.008987	0.0087	50.055	70.00576	4	69.99424		0	25
11	\$0.00	\$7.00	\$14.00	-74.353	-28	-68.2	-62.994	-55.991	-48.991	-68.945	-27.9987	0	0		1	0
12	\$0.00	\$70.00	\$0.00	-74.353	35	-82.2	0.00576	-69.991	0.0087	-19.945	35.0013	2	34.9987		2	19
13	\$21.00	\$35.00	\$10.50	-53.353	0	-71.7	-13.994	-38.491	-24.491	-23.445	0.0013	2	34.9987		3	0
14	\$7.00	\$21.00	\$0.00	-67.353	-14	-82.2	-41.994	-62.991	-48.991	-61.945	-13.9987	0	0		4	8
15	\$5.25	\$7.00	\$2.10	-69.103	-28	-80.1	-57.744	-62.641	-60.891	-75.595	-27.9987	0	0		5	2
16	\$21.00	\$28.00	\$28.00	-53.353	-7	-54.2	-20.994	-20.991	-13.991	-12.945	-6.9987	0	0		6	1
17	\$35.00	\$49.00	\$21.00	-39.353	14	-61.2	14.0058	-13.991	0.0087	15.055	15.05482	7	89.94518		7	22
18	\$21.00	\$21.00	\$21.00	-53.353	-14	-61.2	-27.994	-27.991	-27.991	-26.945	-13.9987	0	0			
19	\$14.00	\$35.00	\$21.00	-60.353	0	-61.2	-20.994	-34.991	-13.991	-19.945	0.0013	2	34.9987			
20	\$14.00	\$35.00	\$14.00	-60.353	0	-68.2	-20.994	-41.991	-20.991	-26.945	0.0013	2	34.9987			
21	\$70.00	\$0.00	\$49.00	-4.35303	-35	-33.2	0.00576	49.00899	-20.991	29.055	49.00899	5	69.99101			
22	\$7.00	\$35.00	\$14.00	-67.353	0	-68.2	-27.994	-48.991	-20.991	-33.945	0.0013	2	34.9987			
23	\$21.00	\$35.00	\$42.00	-53.353	0	-40.2	-13.994	-6.99101	7.0087	8.0548	8.054824	7	89.94518			

Figure 5-1: Verizon Internet example

%4 Part II: Pricing

The key to your model is to set up a spreadsheet that tells you, for any set of prices for each possible product combination, how much revenue you can obtain from your sample of customers. Then you can use the Evolutionary Solver to find the set of prices for the product combinations that maximize your revenue. To find how much revenue you can generate for any set of product combination prices, proceed as follows:

- %4 In D4:J4 enter trial prices for each of the possible seven product combinations.
- %4 In Row 6, determine the first's customer's consumer surplus by computing her value for the products in a combination and subtracting the cost of the product combination. For example, the first customer's consumer surplus for the Internet +TV product combination is computed in cell G6 with the formula `=A6+B6-G$4`. Copy the formulas in row 6 to the cell range D7:J82 to compute each customer's consumer surplus for each product combination.
- %4 Determine for the set of prices in row 4 which product combination, if any, is purchased. Copy the formula `=MAX(D6:J6)` from K6 to K7:K82 to find each consumer's surplus.
- %4 Now here's the key to your spreadsheet. In Column L you use the MATCH function (introduced in Chapter 2, "Using Excel Charts to Summarize Marketing Data") combined with an IF statement to determine which product combination each customer purchases. Use product combination 0 to denote no purchase, whereas the actual product combinations are indicated by the integers 1–7. Copy the formula `=IF(K6<0,0,MATCH(K6,D6:J6,0))` from L6 to L7:L82 yields the product combination (if any) bought by each customer.
- %4 Copy the formula `=IF(L6=0,0,HLOOKUP(L6,$D$3:$J$4,2))` from M6 to M7:M82 to compute for each person the revenue generated.
- %4 In cell M4, compute your total revenue with the formula `=SUM(M6:M82)`

To find the product combination prices that maximize revenue you need to use the Evolutionary Solver, so some discussion of the Evolutionary Solver is in order.

Introduction to the Evolutionary Solver

As explained in Chapter 4, "Estimating Demand Curves and Using Solver to Optimize Price," the Evolutionary Engine is used when your target cell and constraints contain nonsmooth functions that reference changing cells. For example, if your target cell and constraints contain IF SUMIF COUNTIF SUMIFS COUNTIFS AVERAGEIF AVERAGEIFS ABS MAX, or MIN functions that reference the changing cells,

then the Evolutionary engine probably has the best chance to find a good solution to your optimization problem. The model makes extensive use of IF statements, so it is a good choice to use the Evolutionary Solver. The target cell is to maximize revenue (cell M3) by changing the product prices (cell range D4:J4).

When using the Evolutionary Solver, you should follow these rules:

%4 Place upper and lower bounds on each of your changing cells. This makes the Solver's search for an optimal solution much easier. In the Verizon bundling model use a lower bound of 0 for each price and an upper bound of 100.

NOTE If, when running Evolutionary Solver, a changing cell hits an upper or lower bound, you should relax that bound because Solver was trying to push the changing cell outside the range you specified.

%4 Use no constraints other than the bounds on the changing cells. You will soon see how to use penalties in the target cell to enforce constraints.

%4 In the Options tab increase Max Time to 3600 seconds. This causes Solver to run for 60 minutes, even if you leave the room. This should be plenty of time for Solver to find a good solution. Also add a few 000s to Max Subproblems and Max Feasible Solutions. This ensures that Solver does not stop when you leave the room. Figure 5-2 summarizes these settings.

%4 In the Options tab select the Evolutionary Solver tab and change Mutation Rate to 0.5 and Maximum Time Without Improvement to 300 seconds. (Mutation rate is explained soon.) Setting a Maximum Time Without Improvement to 5 minutes ensures that if Solver fails in 5 minutes to improve the current solution, the Solver stops. Hitting the Escape key at any time stops the Solver.

Evolutionary algorithms (often called *genetic algorithms*) were discovered by John Holland, a Michigan computer science professor. Evolutionary Solver begins with between 50 and 200 “solutions,” which are randomly chosen values of the changing cells that lie between the lower and upper bounds for each changing cell. The exact number of solutions used is specified in the Population Size field in Figure 5-3. The default value is 100, which is fine. Then the target cell value is computed for each solution. By a process of “reproduction” explained in David Goldberg’s textbook *Genetic Algorithms* (Addison-Wesley, 1989), a new set of 100 solutions is obtained. Previous solutions that have “good” values (large in a Max problem and small in a Min problem) have a better chance of surviving to the next generation of solutions. This is the mathematical implementation of Darwin’s survival of the fittest principle.

%4 Part II: Pricing

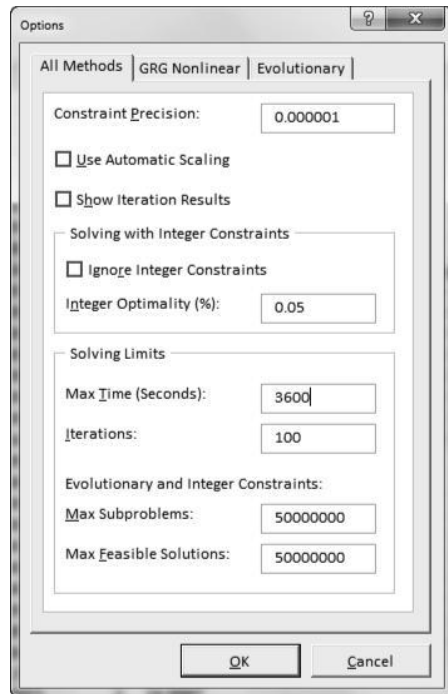


Figure 5-2: Evolutionary Solver settings

The Excel Solver also enables solutions to improve via Mutation. To understand how Mutation works you need to know that Excel codes each value of a changing cell in binary notation. For example, the number 9 is coded as 1001 ($1 * 2^3 + 0 * 2^2 + 0 * 2^1 + 1$).

A Mutation in the Evolutionary Solver changes a 0 to a 1 or a 1 to a 0. A higher value for the Mutation rate moves you around more in the feasible region but exposes you to a larger probability of being led on a wild goose chase after moving to a part of the feasible region with poor changing cell values. A Mutation rate of .5 usually yields the best results. Figure 5-3 shows the Mutation rate changed to .5. The amazingly simple Evolutionary Solver search procedure based on survival of the fittest can solve extraordinarily complex problems! When solving a problem the Evolutionary Solver must make many random choices. This means that if two different people run the same model, then after 5 minutes or so they may see different optimal solutions. Eventually they should see target cell values that are virtually identical.

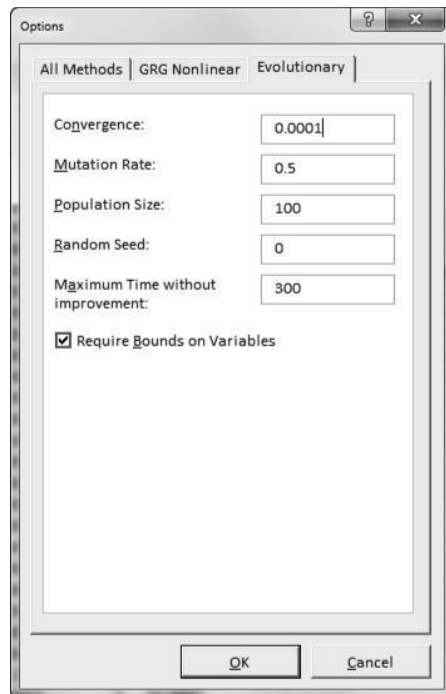


Figure 5-3: Changing the Mutation rate

Finding the Optimal Bundle Prices

The Solver window shown in Figure 5-4 enables you to find revenue maximizing the set of prices for each product combination. A maximum profit of \$3,413.70 (see worksheet initial solver) is found with the product combination prices, as shown in Figure 5-5.

NOTE Unlike the GRG or Simplex Solver engine, the Evolutionary Solver engine is only guaranteed to find a near optimal solution (as opposed to an optimal solution). Therefore, when you run the Evolutionary Solver on the book's examples, do not be surprised if your optimal target cell value differs slightly from the target cell value shown in the book.

%4 Part II: Pricing

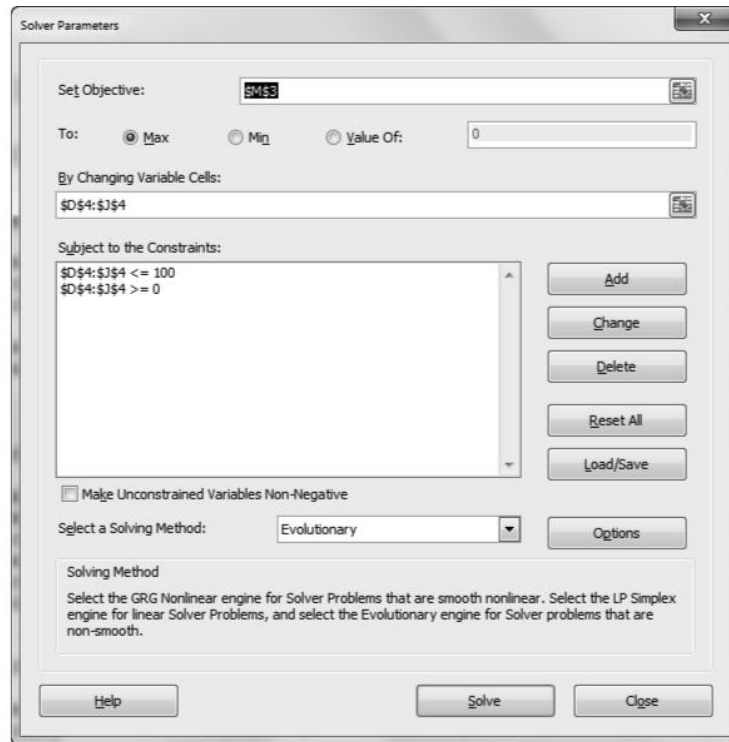


Figure 5-4: Initial bundling Solver model

	C	D	E	F	G	H	I	J
3	Product	1	2	3	4	5	6	7
4	Price	74.35303	35	82.16	69.9942	69.99101	69.991267	89.945
5	Cell Phone	Internet	TV	Cell	Internet +TV	Internet +Cell	TV+Cell	All 3

Figure 5-5: Solution to initial bundling model

Unfortunately, you could not go to market charging \$74.35 for Internet service and \$69.99 for Internet + TV service because it is unreasonable to provide two services for a lower price than a single service. This situation is called a *price reversal*. As you soon see, the reason Solver charged a higher price for Internet service than for Internet and TV service is because the revenue maximizing prices involve nobody just buying Internet service, so a high price helps ensure that nobody buys just Internet service.

Take a look at the final solver worksheet, and you see it contains the same prices found in the initial solver worksheet. The cell range P70:Q81 in the final

solver worksheet (as shown in Figure 5-6) computes the price reversals for each possible combination of a product with a larger combination of products. In column P, the price of the smaller product combination is subtracted from the price of the larger product combination. For example, in P71 the formula `=D4-G4` computes the Internet Price – (Internet + TV Price). The difference of \$4.36 indicates that the Internet + TV price is \$4.36 lower than the Internet price. To ensure that no price reversals occur, the target cell is penalized \$500 for each dollar of price reversal. Then, survival of the fittest should ensure that no price reversals survive. The choice of penalty amount is an art, not a science. Too large a penalty (such as \$1 million) seems to work poorly, and too small a penalty does not “kill off” the things you want to avoid.

	O	P	Q
68			
69	Penalty	dev	penalty
70	TV-(I+TV)	-34.9955	0
71	I-(I+TV)	4.358796	4.358796
72	Cell-(I+Cell)	12.16776	12.16776
73	I-(I+C)	4.36202	4.36202
74	TV-(TV+Cell)	-34.9926	0
75	Cell-(TV+Cell)	12.1675	12.1675
76	(I+TV)-All	-19.9509	0
77	(I+C)-All	-19.9542	0
78	(TV+Cell)-All	-19.9539	0
79	TV-All	-54.9465	0
80	Cell-All	-7.7864	0
81	I-All	-15.5921	0
82	Total		33.05608

Figure 5-6: Price reversals for initial Solver solution

You can copy the formula `=IF (P70>0,P70,0)` from Q70 to Q71:Q82, to track the price reversal because it yields a value of 0 if the product comparison does not have a price reversal. With the formula `=SUM(M6:M82)-500*Q82` you can incorporate your penalty in the target cell. The Solver window is the same (refer to Figure 5-4). After running the Solver you can find the optimal solution, as shown in Figure 5-7.

Therefore, after completing all the calculations, a maximum profit of \$3,413.90 is obtained by charging the following:

- %4 \$47.41 for Internet service
- %4 \$35 for TV service
- %4 \$67.87 for cell phone service
- %4 \$70 for any two product combination
- %4 \$89.95 for all three products

%4 Part II: Pricing

	C	D	E	F	G	H	I	J	K	L	M
1											
2											
3		1	2	3	4	5	6	7		total	3413.9
4	Price	47.41	35	67.87	70	70	70	89.95			
5	Cell Phone	Internet	TV	Cell	Internet +TV	Internet +Cell	TV+Cell	All 3	max surp	bought?	revenue
6	\$3.50	-43.91	-28	-64.4	-59.5	-63	-59.5	-75.95	-28	0	0
7	\$3.50	-29.91	0	-64.4	-17.5	-49	-31.5	-33.95	1.39E-09	2	35
8	\$49.00	-19.41	-7	-18.9	-14	7.000001	7	15.05	15.05	7	89.95
9	\$0.00	22.59	35	-67.9	70	6.01E-07	5E-08	50.05	70	4	70
10	\$14.00	-47.41	-28	-53.9	-63	-56	-49	-68.95	-28	0	0
11	\$0.00	-47.41	35	-67.9	1.9E-08	-70	5E-08	-19.95	35	2	35
12	\$10.50	-26.41	0	-57.4	-14	-38.5	-24.5	-23.45	1.39E-09	2	35
13	\$0.00	-40.41	-14	-67.9	-42	-63	-49	-61.95	-14	0	0
14	\$2.10	-42.16	-28	-65.8	-57.75	-62.65	-60.9	-75.6	-28	0	0
15	\$28.00	-26.41	-7	-39.9	-21	-21	-14	-12.95	-7	0	0
16	\$21.00	-12.41	14	-46.9	14	-14	5E-08	15.05	15.05	7	89.95
17	\$21.00	-26.41	-14	-46.9	-28	-28	-28	-26.95	-14	0	0

Figure 5-7: Final bundling solution

The consumer is given a substantial discount if she buys two or more products. You can copy the formula `=COUNTIF($L$6:$L$82,N9)` from O9 to O10:O16 to find the number of people purchasing each product combination. As shown in Figure 5-8, 25 people buy nothing; nobody buys just the Internet; 19 buy just TV; nobody buys just the cell phone; 8 buy the Internet + TV; 2 buy Internet + cell phone; 1 buys TV + cell phone; and 22 people buy the bundle! Your pricing helped extract the high value people place on the Internet and cell phone service by incentivizing these people to buy more than one service. Note the high prices for the Internet and cell phone are designed to give customers an incentive to purchase more products.

	N	O
8	Product	Frequenc
9		25
10	1	0
11	2	19
12	3	0
13	4	8
14	5	2
15	6	1
16	7	22

Figure 5-8: Number of purchases of each product combination

This method will automatically determine whether separate prices, mixed bundling, or pure bundling is optimal.

Summary

In this chapter you learned the following:

- %4 Bundling products often allows companies to extract more consumer surplus from customers by incentivizing the customers to purchase more products.
- %4 If you assume that each customer will purchase the product with the largest (non-negative) surplus, then you can set up a spreadsheet that tells you the revenue obtained from your customers for any set of prices.
- %4 You can use Evolutionary Solver to maximize Revenue (or Profit) obtained from customers.
- %4 When using Evolutionary Solver you need to implement bounds on changing cells, use a Mutation rate of 0.5, and handle other constraints by incorporating penalties in the target cell.

Exercises

- %4 **A German machine company** sells industrial machinery and maintenance policies for the machine. There are four market segments. Figure 5-9 shows the size of each market segment and the amount each segment will pay for the machine separately, maintenance separately, or the bundle of machinery + maintenance. The variable cost is \$550 per machine and \$470 per maintenance agreement. What set of prices maximize profits?

	B	C	D	E
1	fixed	Machine	Maintenance	Bundle
2	Costs	550	470	1020
3	Size	Machine Use Max	Maintenance Max	Bundle Max
4	12	1250	1500	2600
5	23	1450	540	1750
6	22	1080	1200	2090
7	43	1390	1100	2350

Figure 5-9 Data for Exercise 1

- %4 The file `songdata.xlsx` gives the values several hundred people place on 15 downloadable songs. A subset of the data is shown in Figure 5-10.

%4 Part II: Pricing

If you charge a single price for all songs, how can you maximize revenue?

If you can charge two different prices for songs, how can you maximize revenue?

	C	D	E	F	G	H	I	J
10	Person	Song 1	Song 2	Song 3	Song 4	Song 5	Song 6	Song 7
11	1	\$0.76	\$0.68	\$1.61	\$1.82	\$0.26	\$0.66	\$1.51
12	2	\$0.74	\$0.89	\$1.78	\$1.87	\$0.49	\$0.68	\$1.42
13	3	\$0.36	\$0.49	\$1.40	\$1.42	-\$0.02	\$0.33	\$1.14
14	4	\$0.34	\$0.45	\$1.42	\$1.50	\$0.07	\$0.34	\$1.14
15	5	\$1.27	\$1.31	\$2.14	\$2.38	\$1.02	\$1.19	\$2.05
16	6	\$1.46	\$1.55	\$2.53	\$2.52	\$1.15	\$1.38	\$2.18
17	7	\$0.57	\$0.70	\$1.51	\$1.48	\$0.11	\$0.46	\$1.23
18	8	\$0.62	\$0.74	\$1.48	\$1.71	\$0.31	\$0.47	\$1.29
19	9	\$1.37	\$1.50	\$2.31	\$2.57	\$1.09	\$1.31	\$2.14
20	10	\$0.47	\$0.38	\$1.39	\$1.51	\$0.10	\$0.31	\$1.11
21	11	\$1.50	\$1.61	\$2.42	\$2.56	\$1.13	\$1.43	\$2.10
22	12	\$0.57	\$0.73	\$1.53	\$1.65	\$0.37	\$0.62	\$1.36
23	13	\$0.92	\$0.87	\$1.76	\$1.82	\$0.51	\$0.82	\$1.63
24	14	\$1.02	\$1.13	\$2.07	\$2.22	\$0.65	\$1.11	\$1.91
25	15	\$1.27	\$1.32	\$2.10	\$2.28	\$0.77	\$1.23	\$1.94

Figure 5-10 Data for Exercise 2

%4 Kroger is trying to determine which types of Vanilla Wafer cookies to stock. The wholesale cost Kroger is charged for a box of each type of cookie is shown in the following table:

National 1	National 2	Generic	Copypat
\$1.5	\$2.00	\$0.90	\$1.25

The wholesale price of a box of cookies is assumed equal to the product's quality. Each customer assigns a different value to quality, as shown in file Wafersdata.xlsx and Figure 5-11. For example, Customer 7 would value National 1 at $1.02 * (1.5)$, and so on. Your task is to determine how to price each type of cookie and then to recommend which brands Kroger should stock to maximize profit from cookies.

	E	F	G
7	Customer	Quality value	
8		1	1
9		2	1.003333
10		3	1.006667
11		4	1.01
12		5	1.013333
13		6	1.016667
14		7	1.02
15		8	1.023333
16		9	1.026667
17		10	1.03
18		11	1.033333
19		12	1.036667
20		13	1.04
21		14	1.043333
22		15	1.046667
23		16	1.05

Figure 5-11 Valuations for Exercise 3

%4 Microsoft is going to sell a Student version of Office (excluding Access and Outlook) and a full version of Office. Assume there are three market segments whose size and valuations for each version of Office are given in the following table. What price for each version of Office can maximize Microsoft's revenue?

Segment	Student Version	Full Version	Size
Students	110	160	20%
Individuals (nonstudent)	170	300	40%
Businesses	240	500	40%

%4 *The New York Times* wants to price home subscriptions and web access. Of course, people can buy both home subscriptions and web access. Assume there are three market segments whose size and valuations (per month) for each product combination are given here. The cost to provide a home subscription is \$15 and the variable cost to provide web access is \$1.

Segment	Home Subscription	Web Access	Home +Web	Size
1	\$30	\$35	\$50	25%
2	\$35	\$15	\$40	40%
3	\$20	\$20	\$25	35%

What prices maximize monthly profits?

4 Part II: Pricing

Before publishing a hardcover book, a publisher wants to determine the proper price to charge for the hardcover and paperback versions. The file `paperback.xlsx` contains the valuations of 50 representative potential purchasers of the book. Suppose the unit cost of the hardcover book is \$4 and the unit cost of the paperback is \$2. If the bookstore charges double what it pays the publisher, what price should the publisher charge the bookstore for each version of the book?

For the data in Table 5-3 verify that the given mixed bundling strategy maximizes the seller's profit.

Describe a situation in which a bank can benefit from a bundling strategy.

Suppose your company sells four products and during a year a customer will buy each product at most once. Your company would like to give an end-of-year rebate of 10 percent off to customers who purchase at least \$1,000 of products. How can the methods of this chapter be used to maximize your company's profits?

6

Nonlinear Pricing

Nonlinear pricing is used in a situation in which the cost of purchasing q units is not equal to the unit cost c per unit times q . Bundling (discussed in Chapter 5, “Price Bundling”) is a special type of nonlinear pricing because the price of three items is not equal to the sum of the individual prices.

Other common examples of nonlinear pricing strategies include:

%4 Quantity discounts: If customers buy $\leq CUT$ units, they pay high price (HP) per unit, and if they buy $> CUT$ units, they pay low price (LP) per unit. CUT is simply the “cutoff point at which the charged price changes.” For example, if customers buy ≤ 1000 units, you charge \$10 a unit, but if they buy more than 1000 units, you charge \$8 per unit for all units bought. This form of nonlinear pricing is called the *nonstandard quantity discount*. Another type of quantity discount strategy is as follows: Charge HP for the first CUT units bought, and charge LP for remaining units bought. For example, you charge \$10 per unit for first 1000 units and \$8 per unit for remaining units bought. This form of nonlinear pricing is called the *standard quantity discount*. In both examples the value of $CUT = 1000$

%4 Two-part tariff: The cost of buying q units is a fixed charge K plus $\$c$ per unit purchased. For example, it may cost \$500 to join a golf club and \$30 per round of golf.

Many companies use quantity discounts and two-part tariffs. Microsoft does not charge twice as much for 200 units, for example, as for 100 units. Supermarkets charge less per ounce for a 2-pound jar of peanut butter as for a 1-pound jar of peanut butter. Golf courses often use a two-part tariff by charging an annual membership fee and a charge for each round of golf.

Just as in Chapter 5 you used the Evolutionary Solver to find optimal price bundling strategies, you can use the Evolutionary Solver to find the profit maximizing parameters of a nonlinear pricing strategy. As in Chapter 5, you can assume the

Part II: Pricing

consumer will choose an option giving her the maximum (non-negative) consumer surplus. You can see that nonlinear pricing can often significantly increase your profits, seemingly creating profits out of nothing at all.

In this chapter you will first learn how a consumer's demand curve yields the consumer's willingness to pay for each unit of a product. Using this information you will learn how to use the Evolutionary Solver to determine profit or revenue maximizing nonlinear pricing strategies.

Demand Curves and Willingness to Pay

A **demand curve** tells you for each possible price how many units a customer is willing to buy. A consumer's willingness to pay curve is defined as the maximum amount a customer is willing to pay for each unit of the product. In this section you will learn how to extract the willingness to pay curve from a demand curve.

Suppose you want to sell a software program to a Fortune 500 company. Let q equal the number of copies of the program the company demands, and let p equal the price charged for the software. Suppose you estimated that the demand curve for software is given by $q = 400 - p$. Clearly, your customer is willing to pay less for each additional unit of your software program. Locked inside this demand curve is information about how much the company is willing to pay for each unit of your program. This information is crucial to maximize profitability of sales.

Now rewrite your demand curve as $p = 400 - q$. Thus, when $q = 1$, $p = \$399$, and so on. Now try to figure out the value your customer attaches to each of the first two units of your program. Assuming that the customer is rational, the customer will buy a unit if and only if the value of the unit exceeds your price. At a price of \$400, demand equals 0, so the first unit cannot be worth \$400. At a price of \$399, however, demand equals 1 unit. Therefore, the first unit must be worth somewhere between \$399 and \$400. Similarly, at a price of \$399, the customer does not purchase the second unit. At a price of \$398, however, the customer is purchasing two units, so the customer does purchase the second unit. Therefore, the customer values the second unit somewhere between \$399 and \$398. The customer's willingness to pay for a unit of a product is often referred to as the unit's *reservation price*.

It can be shown that a good approximation to the value of the i^{th} unit purchased by the customer is the price that makes demand equal to $i - 0.5$. For example, by setting q equal to 0.5, you can find that the value of the first unit is $400 - 0.5 = \$399.50$. Similarly, by setting $q = 1.5$, you can find that the value of the second unit is $400 - 1.5 = \$398.50$.

Suppose the demand curve can be written as $p = D(q)$. In your example this looks like $D(q) = 400 - q$. The reader who knows integral calculus can exactly determine the value a consumer places on the first n items by computing $\int_0^n D(q) dq$. In your example you would find the value of the first two units to be the following:

$$\int_0^2 (400 - q) dq = [400q - .5q^2]_0^2 = 800 - 2 = 798$$

This agrees with your approximate method that yields a value of $399.5 + 398.5 = 798$ for the first two units.

Profit Maximizing with Nonlinear Pricing Strategies

Throughout this chapter assume a power company (Atlantis Power and Light, APL for short) wants to determine how to maximize the profit earned from a customer whose demand in kilowatt hours (kwh) for power is given by $q = 20 - 2p$. It costs \$2 to produce a unit of power. Your analysis begins by assuming APL will use linear pricing; that is, charging the same price for each unit sold. You will find that with linear pricing the maximum profit that can be obtained is \$32. Then you will find the surprising result that proper use of quantity discounts or a two-part tariff doubles APL's profit! The work for this chapter is in the file *Powerblockprice.xls*. To determine the profit maximizing linear pricing rule, you simply want to maximize $(20 - 2p) \times (p - 2)$. In the *oneprice* worksheet from the *Powerblockprice.xls* file, a price of \$6 yields a maximum profit of \$32 (see Figure 6-1). The Solver model simply chooses a non-negative price (changing the cell that maximizes profit [Cell D12]). Charging \$6 per kwh yields a maximum profit of \$32.00.

	C	D	E
8	price	\$6.00	
9	demand	8	
10	unit cost	\$2.00	
11			
12	profit	\$32.00	
13			

Figure 6-1: Finding the profit maximizing single price strategy

Optimizing the Standard Quantity Discount

To determine a profit maximizing pricing strategy that uses the standard quantity discount, assume the quantity discount pricing policy is defined as

%4 Part II: Pricing

follows: All units up to a value CUT are sold at a high price (HP). Recall that CUT is simply the cutoff point at which the per unit price is lowered. All other units sell at a lower price (LP). Assuming the customer chooses the number of kwh with the highest non-negative surplus, you can use the Evolutionary Solver to determine profit maximizing values of CUT , HP , and LP . The work for this task is shown in sheet qd of file Powerblockprice.xls. Also, Figure 6-2 shows that the demand curve may be written as $p = 10 - (q/2)$, so the first unit is valued at $10 - (.5/2) = \$9.75$, the second unit is valued at $10 - (1.5/2) = \$9.25$, and so on.

	C	D	E	F	G	H	I	J
1			cutoff	5		Units bought	16	
2	cost		HP	14.30422		Revenue	\$ 96.00	
3	2		LP	2.2253		Prod Cost	\$ 32.00	
4					Max surplus	0.000611745		
5	Midpoint	Unit	Value	Cum Valu	Price paid	Surplus		Profit
6	0.5	1	9.75	9.75	14.30421665	-4.554216652		\$ 64.00
7	1.5	2	9.25	19	28.6084333	-9.608433303		
8	2.5	3	8.75	27.75	42.91264995	-15.16264995		
9	3.5	4	8.25	36	57.21686661	-21.21686661		
10	4.5	5	7.75	43.75	71.52108326	-27.77108326		
11	5.5	6	7.25	51	73.74638371	-22.74638371		
12	6.5	7	6.75	57.75	75.97168417	-18.22168417		
13	7.5	8	6.25	64	78.19698462	-14.19698462		
14	8.5	9	5.75	69.75	80.42228508	-10.67228508		
15	9.5	10	5.25	75	82.64758553	-7.647585529		
16	10.5	11	4.75	79.75	84.87288598	-5.122885984		
17	11.5	12	4.25	84	87.09818644	-3.098186438		
18	12.5	13	3.75	87.75	89.32348689	-1.573486892		
19	13.5	14	3.25	91	91.54878735	-0.548787347		
20	14.5	15	2.75	93.75	93.7740878	-0.024087801		
21	15.5	16	2.25	96	95.99938826	0.000611745		
22	16.5	17	1.75	97.75	98.22468871	-0.47468871		
23	17.5	18	1.25	99	100.4499892	-1.449989164		
24	18.5	19	0.75	99.75	102.6752896	-2.925289618		
25	19.5	20	0.25	100	104.9005901	-4.900590073		

Figure 6-2: Finding the profit maximizing standard quantity discount strategy

To complete the determination of the profit maximizing standard quantity discount strategy, complete the following steps:

%4 Copy the formula $=10-0.5*C6$ from E6 to E7:E25 to determine the value of each unit.

%4 In column F compute the cumulative value associated with buying 1, 2,...19 units. In F6 compute the value of the first unit with formula $=E6$. Copy the formula $=F6+E7$ from F7 to F8:F25 to compute the cumulative value of buying 2, 3,...20 units.

- %4 Copy the formula $=IF(D6 \leq cutoff, HP * D6, HP * cutoff + LP * (D6 - cutoff))$ from G6 to G7:G25 to compute the total cost the consumer pays for purchasing each number of units. If you want to analyze different nonlinear pricing strategies, just change this column to a formula that computes the price the customer is charged for each number of units purchased.
- %4 Copy the formula $=F6 - G6$ from H6 to H7:H25 to compute the consumer surplus associated with each purchase quantity. In cell H4 compute the maximum surplus with the formula $=IF(MAX(H6:H25) >= 0, MAX(H6:H25), 0)$; if it is negative no units will be bought, and in this case you enter a surplus of 0.
- %4 In cell I1 use the match function to compute the number of units bought with the function $=IF(H4 > 0, MATCH(H4, H6:H25, 0), 0)$
- %4 In cell I2 compute the sales revenue with the formula $=IF(I1 = 0, 0, VLOOKUP(I1, lookup, 4))$. The range D5:G25 is named Lookup.
- %4 In cell I3 compute production cost with the formula $=I1 * C3$
- %4 In cell J6 compute profit with the formula $=I2 - I3$
- %4 Use Solver to find the values of *CUT*, *HP*, and *LP* that maximize profit. The Solver window is shown in Figure 6-3.

You can constrain the value of *CUT* and each price to be at most \$20. To do so, charge \$14.30 per unit for the first 5 units and \$2.22 for remaining units. A total profit of \$64 is earned. The quantity discount pricing structure here is an incentive for the customer to purchase 16 units, whereas linear pricing results in the customer purchasing only 6 units. After all, how can the customer stop at 6 units when later units are so inexpensive?

Optimizing the Nonstandard Quantity Discounts

Now assume that your quantity discount strategy is to charge a high price (*HP*) if at most *CUT* units are purchased, and if more than *CUT* units are purchased, all units are sold at a low price (*LP*). The only change needed to solve for the profit maximizing strategy of this form (see sheet qd2) is to copy the formula $=IF(D6 \leq cutoff, HP * D6, LP * D6)$ from G6 to G7:G25. Then just rerun the Solver window, as shown in Figure 6-3. This pricing strategy, shown in Figure 6-4, maximizes the profit: If the customer buys up to 15 units, he pays \$16.79, and if he buys at least 16 units, he pays \$6 per unit. Then the customer buys 16 units, and you make \$64 in profit.

%4 Part II: Pricing

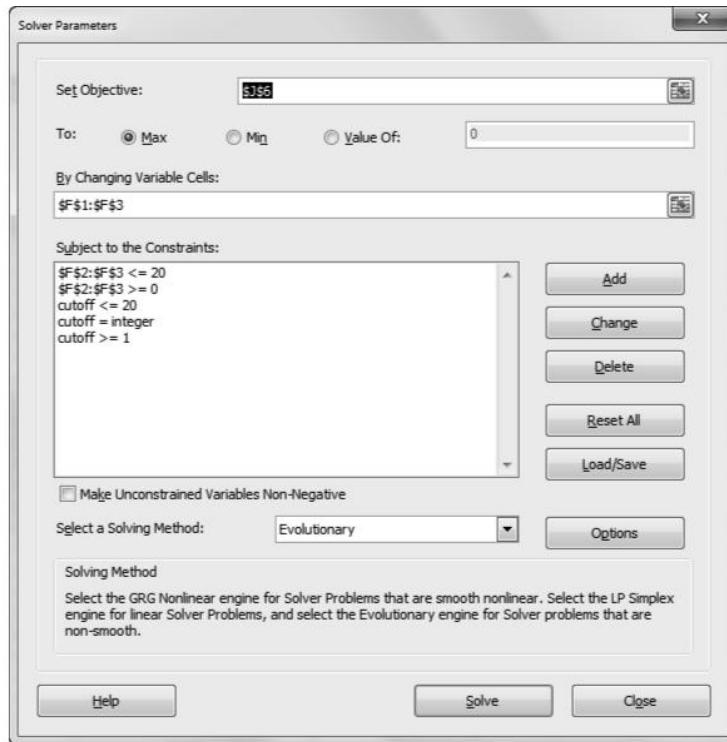


Figure 6-3: Determining the profit maximizing standard quantity discount

	C	D	E	F	G	H	I	J
1			cutoff	15		Units bought	16	
2	cost		HP	16.79609		Revenue	\$ 96.00	
3	2		LP	5.999873		Prod Cost	\$ 32.00	
4					Max surplus	0.002031561		
5	Midpoint	Unit	Value	Cum Value	Price paid	Surplus		Profit
6	0.5	1	9.75	9.75	16.79608788	-7.046087876		\$ 64.00
7	1.5	2	9.25	19	33.59217575	-14.59217575		
8	2.5	3	8.75	27.75	50.38826363	-22.63826363		
9	3.5	4	8.25	36	67.1843515	-31.1843515		
10	4.5	5	7.75	43.75	83.98043938	-40.23043938		
11	5.5	6	7.25	51	100.7765273	-49.77652726		
12	6.5	7	6.75	57.75	117.5726151	-59.82261513		
13	7.5	8	6.25	64	134.368703	-70.36870301		
14	8.5	9	5.75	69.75	151.1647909	-81.41479088		
15	9.5	10	5.25	75	167.9608788	-92.96087876		
16	10.5	11	4.75	79.75	184.7569666	-105.0069666		
17	11.5	12	4.25	84	201.5530545	-117.5530545		
18	12.5	13	3.75	87.75	218.3491424	-130.5991424		
19	13.5	14	3.25	91	235.1452303	-144.1452303		
20	14.5	15	2.75	93.75	251.9413181	-158.1913181		
21	15.5	16	2.25	96	95.99796844	0.002031561		
22	16.5	17	1.75	97.75	101.9978415	-4.247841466		
23	17.5	18	1.25	99	107.9977145	-8.997714493		
24	18.5	19	0.75	99.75	113.9975875	-14.24758752		
25	19.5	20	0.25	100	119.9974605	-19.99746055		

Figure 6-4: Profit maximizing nonstandard quantity discount

Optimizing a Two-Part Tariff

Organizations such as a golf country club or Sam's Club find it convenient to give the consumer an incentive to purchase many units of a product by charging a fixed fee to purchase any number of units combined with a low constant per unit cost. This method of nonlinear pricing is called a *two-part tariff*. To justify paying the fixed fee the customer needs to buy many units. Of course, the customer will not buy any units whose unit cost exceeds the consumer's reservation price, so the organization must be careful to not charge a unit price that is too high.

For example, when you join a country club, you often pay a fixed fee for joining and then a constant fee per round of golf. You might pay a \$500 per year fixed fee to be a club member and also pay \$30 per round of golf. Therefore one round of golf would cost \$530 and two rounds of golf would cost \$560. If a two-part tariff were linear then for any number of purchased units, the cost per unit would be the same, but because one round cost \$530 per round and two rounds costs $\$530/2 = \$265$, the two-part tariff is not consistent with linear pricing.

To dig a little deeper into the concept of a two-part tariff, look at the worksheet *tp*. You will see how a power company can optimize a two-part tariff by performing the following steps:

- %4 To begin, make a copy of your QD or QD2 worksheet.
- %4 In F2 enter a trial value for the fixed fee, and in F3 enter a trial value for the price per unit. Name cell F2 *Fixed* and cell F3 *var*. The only formulas that you need to change are the formulas in the Price Paid formula in Column G (see Figure 6-5).
- %4 Copy the formula `=Fixed+D6*Var` from G6 to G7:G25, to compute the cost of buying each number of units.
- %4 Adjust the Solver window, as shown in Figure 6-6.

You need to constrain the fixed and variable costs between \$0 and \$100. The profit is maximized by charging a fixed fee of \$60.27 and a cost of \$2.21 per unit. Again a maximum profit of \$64 is made, and the customer purchases 16 units.

%4 Part II: Pricing

	C	D	E	F	G	H	I	J
1				0		Units bought	16	
2	cost		Fixed	60.57433887		Revenue	\$ 96.00	
3		2	var	2.214067061		Prod Cost	\$ 32.00	
4					Max surplus	0.000588158		
5	Midpoint	Unit	Value	Cum Value	Price paid	Surplus		Profit
6	0.5	1	9.75	9.75	62.78840593	-53.03840593		\$ 64.00
7	1.5	2	9.25	19	65.00247299	-46.00247299		
8	2.5	3	8.75	27.75	67.21654005	-39.46654005		
9	3.5	4	8.25	36	69.43060711	-33.43060711		
10	4.5	5	7.75	43.75	71.64467417	-27.89467417		
11	5.5	6	7.25	51	73.85874123	-22.85874123		
12	6.5	7	6.75	57.75	76.07280829	-18.32280829		
13	7.5	8	6.25	64	78.28687535	-14.28687535		
14	8.5	9	5.75	69.75	80.50094241	-10.75094241		
15	9.5	10	5.25	75	82.71500948	-7.715009476		
16	10.5	11	4.75	79.75	84.92907654	-5.179076537		
17	11.5	12	4.25	84	87.1431436	-3.143143598		
18	12.5	13	3.75	87.75	89.35721066	-1.607210659		
19	13.5	14	3.25	91	91.57127772	-0.57127772		
20	14.5	15	2.75	93.75	93.78534478	-0.035344781		
21	15.5	16	2.25	96	95.99941184	0.000588158		
22	16.5	17	1.75	97.75	98.2134789	-0.463478903		
23	17.5	18	1.25	99	100.427546	-1.427545964		
24	18.5	19	0.75	99.75	102.641613	-2.891613025		
25	19.5	20	0.25	100	104.8556801	-4.855680086		

Figure 6-5: Optimizing a two-part tariff

Solver Parameters

Set Objective:

To: ☒ Max ☐ Min ☐ Value Of:

By Changing Variable Cells:

Subject to the Constraints:

☐ Make Unconstrained Variables Non-Negative

Select a Solving Method:

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 6-6: Solver window for two-part tariff

All three of the fairly simple nonlinear pricing strategies result in doubling the linear price profit to \$64. There is no nonlinear pricing plan that can make more than \$64 off the customer. You can easily show the exact value of the sixteenth unit is \$2.25, and the exact value of the seventeenth unit is \$1.75, which is less than the unit cost of \$2.00. Because the customer values unit 16 more than the cost, you should be able to get her to buy 16 units. Because the area under the demand curve (see Exercise 5) from 0 to 16 is 96, the most you can make the consumer pay for 16 units is \$96. Therefore your maximum profit is $96 - (\$2)(16) = \$64$, which is just what you obtained using all three nonlinear pricing strategies. The problem with the linear pricing strategy is that it does not extract the surplus value in excess of the \$2 cost that the customer places on earlier units. The nonlinear pricing strategies manage to extract all available consumer surpluses.

For a monopolist (like a power company) it may be realistic to assume that an organization can extract all consumer surplus. In the presence of competition, it may be unrealistic to assume that an organization can extract all consumer surplus. To illustrate how the models of this chapter can be modified in the presence of competition, suppose you are planning to enter a business where there is already a single company selling a product. Assume that this company is currently extracting 80 percent of the consumer surplus and leaving 20 percent of the consumer surplus with the consumer. By adding a constraint to your models that leaves, say 25 percent of the consumer surplus in the hands of the customer, you can derive a reasonable nonlinear pricing strategy that undercuts the competition. For an illustration of this idea see Exercise 8.

Summary

In this chapter you learned the following:

- %4 Nonlinear pricing strategies involve not charging the same price for each unit sold.
- %4 If you assume customers purchase the available option with the largest (if non-negative) surplus, then you can use the Evolutionary Solver to find profit maximizing nonlinear pricing strategies such as quantity discounts and two-part tariffs.
- %4 For a linear demand curve, nonlinear pricing earns twice as much profit as linear pricing.

%4 Part II: Pricing

Exercises

%4 **You own** a small country club. You have three types of customers who value each round of golf they play during a month as shown in the following table:

Round	Segment 1	Segment 2	Segment 3
1	\$60	\$60	\$40
2	\$50	\$45	\$30
3	\$40	\$30	\$20
4	\$30	\$15	\$10
5	\$20	\$0	\$0
6	\$10	\$0	\$0

It costs you \$5 in variable costs to provide a customer with a round of golf. Find a profit maximizing a two-part tariff. Each market segment is of equal size.

%4 The demand curve for a product is $q = 4000 - 40p$. It costs \$5 to produce each unit.

If you charge a single price for each unit, how can you maximize profit?

If you use a two-part tariff, how can you maximize the profit?

%4 Using the data in Exercise 1, determine the profit maximizing the quantity discount scheme in which all units up to a point sell for a high price, and all remaining units sell for a low price.

%4 The file `Finalmusicdata.xls` (available on the companion website) contains data showing the most 1,000 people were willing to pay for 10 songs. For example, as shown in Figure 6-7, Person 1 is willing to pay up to 38 cents for Song 1.

	C	D	E	F	G	H	I	J	K	L	M
3		Song 1	Song 2	Song 3	Song 4	Song 5	Song 6	Song 7	Song 8	Song 9	Song 10
4	1	\$0.38	\$0.94	\$1.32	\$1.85	\$0.57	\$1.16	\$1.39	\$2.23	\$1.95	\$1.88
5	2	\$1.02	\$1.83	\$2.23	\$0.56	\$0.30	\$2.12	\$2.09	\$1.63	\$1.65	\$1.50
6	3	\$0.60	\$1.44	\$1.91	\$1.67	\$0.82	\$2.41	\$0.64	\$1.00	\$1.80	\$1.46
7	4	\$0.31	\$2.10	\$1.12	\$0.84	\$1.30	\$0.61	\$0.84	\$1.95	\$0.88	\$0.51
8	5	\$1.71	\$1.22	\$1.76	\$2.35	\$0.99	\$1.96	\$2.04	\$0.65	\$2.33	\$1.56
9	6	\$1.44	\$2.35	\$2.27	\$0.94	\$1.20	\$0.43	\$0.51	\$1.97	\$1.85	\$1.51
10	7	\$0.22	\$1.04	\$1.77	\$1.51	\$0.91	\$2.11	\$1.33	\$2.10	\$2.14	\$0.61
11	8	\$1.27	\$0.93	\$2.40	\$1.57	\$1.91	\$0.35	\$1.14	\$1.24	\$1.70	\$0.63
12	9	\$1.85	\$0.36	\$0.28	\$0.61	\$1.75	\$1.59	\$1.55	\$1.00	\$0.43	\$1.37
13	10	\$0.85	\$2.12	\$1.29	\$2.41	\$1.02	\$0.42	\$1.18	\$2.32	\$1.27	\$0.58
14	11	\$2.07	\$2.24	\$1.26	\$1.49	\$0.25	\$0.42	\$0.75	\$2.09	\$1.65	\$1.14

Figure 6-7: Data for Problem 4

- %4 If you charge a single price for each song, what price can maximize your revenue?
- %4 If you use a two-part tariff pricing scheme (for example \$3 to download any songs + 40 cents per song), how can you maximize revenue?
- %4 How much more money do you make with the two-part tariff?

NOTE For the second part, you must realize that a person will buy either nothing, their top-valued song, their top 2 songs, their top 3 songs,...their top 9 songs or top 10 songs.

- %4 Show that for the demand curve $q = 40 - 2p$ the consumer places a value of exactly \$256 on the first 16 units.
- %4 (Requires calculus) Suppose the demand for a computer by a leading corporate client (for $p \geq 100$) is given by $200000 / p^2$. Assume the cost of producing a computer is \$100.

If the same price is charged for each computer, what price maximizes profit?

If you charge HP for the first CUT computers and LP for remaining computers purchased, how can you maximize profit?

%4 Part II: Pricing

Suppose Verizon has only three cell phone customers. The demand curve for each customer's monthly demand (in hours) is shown here:

Customer	Demand curve
1	$Q = 60 - 20p$
2	$Q = 70 - 30p$
3	$Q = 50 - 8p$

Here p = price in dollars charged for each hour of cell phone usage. It costs Verizon 25 cents to provide each hour of cell phone usage.

If Verizon charges the same price for each hour of cell phone usage, what price should they charge?

Find the profit maximizing a two-part tariff for Verizon. How much does the best two-part tariff increase the profit over the profit maximizing the single price?

%4 For the power company example construct a standard quantity discount pricing strategy that leaves the customer with 25 percent of the consumer surplus.

7

Price Skimming and Sales

To solve many interesting pricing problems, you can use the assumption that consumers choose the option that gives them the maximum consumer surplus (if it is non-negative) as discussed in Chapter 5, “Price Bundling,” and Chapter 6, “Nonlinear Pricing.” This chapter uses this assumption and the Evolutionary Solver to explain two pricing strategies that are observed often:

- %4 Why do prices of high tech products usually drop over time?
- %4 Why do many stores have sales or price promotions?

Dropping Prices Over Time

The **prices of most high-tech products as well as** other products such as new fashion styles tend to drop over time. You might be too young to remember that when VCRs were first introduced, they sold for more than \$1,000. Then the price of VCRs quickly dropped. There are a few reasons for this behavior in prices, and this section examines three of them.

Learning Curve

The most common reason that prices of products—high-tech items in particular—drop over time is due to a learning or experience curve. As first observed by T.P. Wright in 1936 during his study of the costs of producing airplanes, it is often the case that the unit cost to produce a product follows a *learning curve*. Suppose $y =$ unit cost to produce the x^{th} unit of a product. In many situations $y = ax^{-b}$ where $a > 0$ and $b > 0$. If the unit cost of a product follows this equation, the unit cost of production follows a learning curve. If it can easily be shown (see Exercise 3) that if a product’s cost follows a learning curve, then whenever cumulative product doubles, unit cost drops by the same percentage $(1 - 2^{-b})$, see Exercise 1). In most observed cases, costs drop between 10 percent and 30 percent when cumulative production doubles.

%4 Part II: Pricing

If unit costs follow a learning curve, costs drop as more units are sold. Passing this drop in costs on to the consumer results in prices dropping over time. The learning curve also gives an incentive to drop prices and increase capacity so your company sells more, increases your cost advantage over competitors, and perhaps puts them out of business. This strategy was popularized in the 1970s by Bruce Henderson of the Boston Consulting Group. Texas Instruments followed this strategy with pocket calculators during the 1980s.

Competition

Aside from a learning curve causing these types of reductions, prices also drop over time as competitors enter the market; this increases supply that puts a downward pressure on prices.

Price Skimming

%4 third reason why prices of products drop over time is *price skimming*. When a new product comes out, everyone in the market places a different value on the product. If a company starts with a low price, it foregoes the opportunity to have the high-valuation customers pay their perceived product value. As time passes, the high-valuation customers leave the market, and the company must lower the price to sell to the remaining lower valuation customers.

The following example shows how price skimming works. Suppose there are 100 people who might be interested in buying a product. One person values the product at \$1; one person values the product at \$2; one person values the product at \$100. At the beginning of each of the next 10 years, you set a price for the product. Anyone who values the product at an amount at least what you charge buys the product. The work in the Skim.xls file shows the pricing strategy that maximizes your total revenue over the next 10 years. The following steps walk you through the maximization.

Enter trial prices for each year in C5:C14.

Copy the formula =C5-1 from D5 to D6:D14 to keep track of the highest price valuation person that is left after a given year. For example, after Year 1 people valuing the product at \$91 or less are left.

Copy the formula =D4-D5 from E5 to E6:E14 to track the unit sale for each year. For example, in Year 1 all people valuing the product at \$92 or more (9 people) buy the product.

Copy the formula =E5*C5 from F5 to F6:F14 to compute each year's revenue.

In cell F15 compute the total revenue with the formula =SUM(F5:F14)

%4 Use the Solver to find revenue-maximizing prices. The Solver window is shown in Figure 7-1. Choose each year's price to be an integer between \$1 and \$100 with a goal to maximize revenue. Figure 7-2 shows the sequence of prices shown in C5:C14.

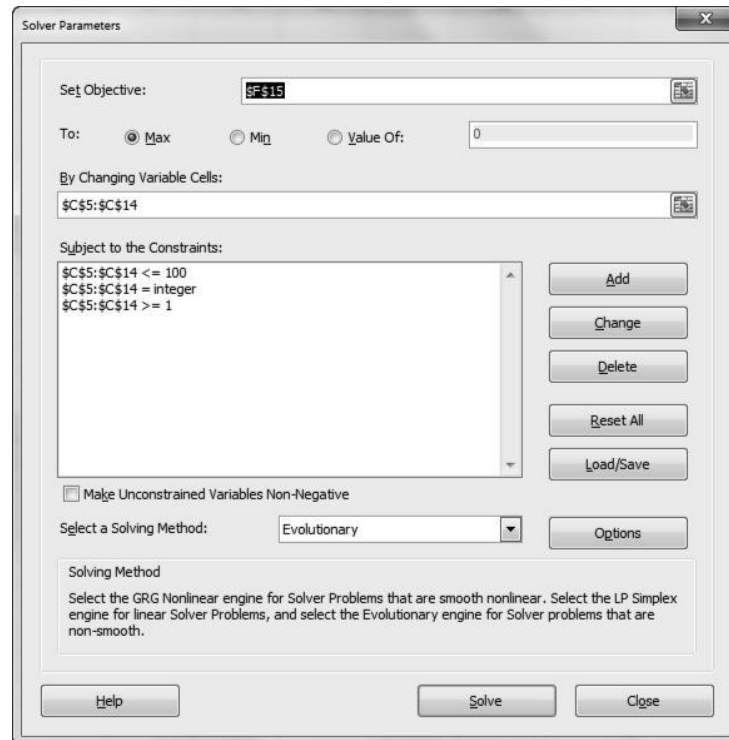


Figure 7-1 Solver window for price skimming model

	A	B	C	D	E	F
1		Skimming				
2						
3		Year	Price	Left<=	Unit sales	Revenue
4		0		100		
5		1	\$ 92.00	91	9	828
6		2	\$ 83.00	82	9	747
7		3	\$ 74.00	73	9	666
8		4	\$ 65.00	64	9	585
9		5	\$ 56.00	55	9	504
10		6	\$ 46.00	45	10	460
11		7	\$ 37.00	36	9	333
12		8	\$ 27.00	26	10	270
13		9	\$ 18.00	17	9	162
14		10	\$ 9.00	8	9	81
15				Total		4636

Figure 7-2: Price skimming model

%4 Part II: Pricing

You should discover that prices decline as you “skim” the high-valuation customers out of the market early. You could model prior purchasers coming into market after, say, three years, if the product wears out. You just need to track the status of the market (that is, how many people with each valuation are currently in the market) at each point in time and then the Solver adjusts the skimming strategy.

Companies often have other reasons for engaging in price skimming. For instance, a firm may engage in price skimming because it wants its products to be perceived as high quality (e.g. Apple’s iPhone and iPad devices). Alternatively, the firm may engage in price skimming to artificially manipulate demand and product interest (e.g. Nintendo’s Wii console.)

There are, however, downsides to a price skimming strategy:

- %4 The early high prices lead to high profit margins that may encourage competitors to enter the market.
- %4 Early high prices make it difficult to take advantage of the learning curve.
- %4 Early high prices reduce the speed at which the product diffuses through the market. This is because with fewer earlier purchasers there are fewer adopters to spread the word about the product. The modeling of the diffusion of a product is discussed in Chapter 27, “The Bass Diffusion Model.”

Why Have Sales?

The **main idea behind retailers** having sales of certain products is that different people in the market place different values on the same products. For a durable good such as an electric razor, at different points in time, there will be a different mix of people wanting to buy. When people with low product valuations predominate in the market, you should charge a low price. When people with high product valuations predominate in the market, you should charge a higher price. The following example develops a simple spreadsheet model which illustrates this idea.

Assume that all customers value an electric razor at \$30, \$40, or \$50. Currently, there are 210 customers in the market, and an equal number value the razor at \$30, \$40, or \$50. Each year 20 new customers with each valuation enter the market. A razor is equally likely to last for one or two years. The work in the sales worksheet of the Sales.xls workbook (see Figure 7-3) determines a pricing policy that maximizes your profit over the next 20 years. These steps are outlined in the following list:

- %4 Enter a code and the valuation associated with each code in D4:E6. The code for each year is the changing cell and determines the price charged that year. For instance, a code of 2 for one year means a price of \$40 will be charged.

%4 Enter a trial set of codes in B8:B27.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				New	20								
2			Why Have Sales?				Buyerst+1=Nonbuyerst+(buyerst-salest)+.5*salest+.5new						
3				Code	Price		Nonbuyerst+1=.5New+.5Salest						
4				1	30								totalrevenue
5				2	40								401662.964
6				3	100								
7	Period	Code	Price	High buyers	High nonbuyers	Medium buyers	Medium nonbuyers	Low buyers	Low nonbuyers	High sales	Medium sales	Low sales	Revenue
8	1	3	100	45	45	45	45	45	45	45	0	0	4500
9	2	1	30	77.5	32.5	100	10	100	10	77.5	100	100	8325
10	3	3	100	81.25	48.75	70	60	70	60	81.3	0	0	8125
11	4	1	30	99.375	50.625	140	10	140	10	99.4	140	140	11381.25
12	5	3	100	110.313	59.6875	90	80	90	80	110	0	0	11031.25
13	6	1	30	124.844	65.15625	180	10	180	10	125	180	180	14545.3125
14	7	3	100	137.578	72.42188	110	100	110	100	138	0	0	13757.8125
15	8	1	30	151.211	78.78906	220	10	220	10	151	220	220	17736.3281
16	9	3	100	164.395	85.60547	130	120	130	120	164	0	0	16439.4531
17	10	1	30	177.803	92.19727	260	10	260	10	178	260	260	20934.082
18	11	3	100	191.099	98.90137	150	140	150	140	191	0	0	19109.8633
19	12	1	30	204.451	105.5493	300	10	300	10	204	300	300	24133.5205
20	13	3	100	217.775	112.2253	170	160	170	160	218	0	0	21777.4658
21	14	1	30	231.113	118.8873	340	10	340	10	231	340	340	27333.3801
22	15	3	100	244.444	125.5563	190	180	190	180	244	0	0	24444.3665
23	16	1	30	257.778	132.2218	380	10	380	10	258	380	380	30533.345
24	17	3	100	271.111	138.8891	210	200	210	200	271	0	0	27111.0916
25	18	1	30	284.445	145.5555	420	10	420	10	284	420	420	33733.3363
26	19	3	100	297.778	152.2223	230	220	230	220	298	0	0	29777.7729
27	20	1	30	311.111	158.8889	460	10	460	10	311	460	460	36933.3341

Figure 7-3 Why Have Sales? model

%4 Compute the price charged each year in the cell range C8:C27. The Year 1 price is computed in C8 with the formula =VLOOKUP(B8,lookup2)

%4 Copy this formula to the cell range C8:C27 to compute the price charged each year.

%4 Enter the number of people valuing a razor at \$30, \$40, and \$50 (classified as buyers and nonbuyers) in Year 1 in D8:I8. Half of all new people are classified as buyers and half as nonbuyers. Also 50 percent of holdovers are buyers and 50 percent are nonbuyers.

%4 The key to the model is to accurately track each period for the number of buyers and nonbuyers of each valuation. The key relationships include the following two equations:

$$\begin{aligned}\text{Buyers}(t+1) &= \text{Non buyers}(t) + (\text{buyers}(t) - \text{Sales}(t)) \\ &\quad .5 * \text{Sales}(t) + .5 * \text{New} \\ \text{Nonbuyers}(t+1) &= .5 * \text{New} + .5 * \text{Sales}(t)\end{aligned}$$

%4 Copy the formula =E8+(D8-J8)+0.5*J8+0.5*New from D9 to D10:D28 to use Equation 1 to compute the number of high-valuation buyers in periods 2–20.

%4 Copy the formula 0.5*New+0.5*J8 from E9 to E10:E28 to use Equation 2 to compute the number of high-valuation nonbuyers in periods 2–20.

%4 In J8:L8 determine the number of people of each type (High, Medium, and Low valuation) who purchase during Year 1. Essentially, all members of High, Medium, or Low in the market purchase if the price does not exceed their valuation. Thus in K8 the formula `=IF(C8<=E$5,F8,0)` determines the number of Medium valuation people purchasing during Year 1. The Solver window for the sales example is shown in Figure 7-4.



Figure 7-4 Solver window for Why Have Sales?

In the Solver model the target cell is meant to maximize the total profit (cell M4). The changing cells are the codes (B8:B27). Constrain the codes to be integers between 1 and 3. You use codes as changing cells rather than prices because prices need to be constrained as integers between 30 and 50. This would cause the Solver to consider silly options such as charging \$38. During most years a price of \$30 is optimal, but during some years a price of \$40 is optimal. The price of \$40 is optimal only during years in which the number of \$40 customers in the market exceeds the number of \$30 customers in the market. The \$50 people get off easy! They never

pay what the product is worth. Basically, the razors are on sale 50 percent of the time. (Ten of the 20 years' price is \$30!)

It is interesting to see how changing the parameters of this example can change the optimal pricing policy. Suppose the high-valuation customers value the product at \$100. Then the Solver finds the optimal solution, as shown in Figure 7-5 (see the high price worksheet in the Sales.xls file). You cycle between an expensive price and a fire sale!

Of course stores have sales for other reasons as well:

- %4 Drugstores and supermarkets often put soda on sale as a loss leader to draw people into the store in the hope that the customer will buy other products on their current trip to the store or return to the store in the future.
- %4 Stores often have sales to clear out excess inventory to make room for products that have better sales potential. This is particularly true when a new version of a product comes out (like a new PC or phone.)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				New		20							
2				Why Have Sales?			Buyerst+1=Nonbuyerst+(buyerst-salest)*.5*salest+.5new						
3				Code	Price		Nonbuyerst+1=.5New+.5Salest						totalrevenue
4				1	30								401662.964
5				2	40								
6				3	100					100	40	30	
7	Period	Code	Price	High buyers	High nonbuyers	Medium buyers	Medium nonbuyers	Low buyers	Low nonbuyers	High sales	Medium sales	Low sales	Revenue
8	1	3	100	45	45	45	45	45	45	45	0	0	4500
9	2	1	30	77.5	32.5	100	10	100	10	77.5	100	100	8325
10	3	3	100	81.25	48.75	70	60	70	60	81.3	0	0	8125
11	4	1	30	99.375	50.625	140	10	140	10	99.4	140	140	11381.25
12	5	3	100	110.313	59.6875	90	80	90	80	110	0	0	11031.25
13	6	1	30	124.844	65.15625	180	10	180	10	125	180	180	14545.3125
14	7	3	100	137.578	72.42188	110	100	110	100	138	0	0	13757.8125
15	8	1	30	151.211	78.78906	220	10	220	10	151	220	220	17736.3281
16	9	3	100	164.395	85.60547	130	120	130	120	164	0	0	16439.4531
17	10	1	30	177.803	92.19727	260	10	260	10	178	260	260	20934.082
18	11	3	100	191.099	98.90137	150	140	150	140	191	0	0	19109.8633
19	12	1	30	204.451	105.5493	300	10	300	10	204	300	300	24133.5205
20	13	3	100	217.775	112.2253	170	160	170	160	218	0	0	21777.4658
21	14	1	30	231.113	118.8873	340	10	340	10	231	340	340	27333.3801
22	15	3	100	244.444	125.5563	190	180	190	180	244	0	0	24444.3665
23	16	1	30	257.778	132.2218	380	10	380	10	258	380	380	30533.345
24	17	3	100	271.111	138.8891	210	200	210	200	271	0	0	27111.0916
25	18	1	30	284.445	145.5555	420	10	420	10	284	420	420	33733.3363
26	19	3	100	297.778	152.2223	230	220	230	220	298	0	0	29777.7729
27	20	1	30	311.111	158.8889	460	10	460	10	311	460	460	36933.3341

Figure 7-5 Sales solution when high valuation = \$100

- %4 Stores often have a sale on a new product to get customers to sample the product in the hopes of maximizing long-term profits from sales on the product. The importance of long-term profits in marketing analytics will be discussed in Chapters 19-22, which cover the concept of lifetime customer value.

Summary

In this chapter you learned the following:

- %4 Prices of high-tech products often drop over time because when the product first comes out, companies want to charge a high price to customers who value the product highly. After these customers buy the product, the price must be lowered to appeal to customers who have not yet purchased the product.
- %4 Because customers have heterogeneous valuations for a product, if you do not have sales, you never can sell your product to customers with low valuations for the product.

Exercises

- %4 **Joseph A. Bank** often has a deal in which men can buy one suit and get two free. The file `Banks.xlsx` contains the valuations of 50 representative customers for one, two, or three suits. Suppose it costs Joseph A. Banks \$150 to ready a suit for sale. What strategy maximizes Joseph A. Banks profit: charging a single price for each suit, charging a single price for two suits, or charging a single price for three suits?
- %4 The file `Coupons.xlsx` gives the value that a set of customers associates with a Lean Cuisine entree. The file also gives the “cost” each person associates with clipping and redeeming a coupon. Assume it costs Lean Cuisine \$1.50 to produce a dinner and 10 cents to redeem a coupon. Also the supermarket sells the entrée for twice the price it pays for the entrée. Without coupons what price should Lean Cuisine charge the store? How can Lean Cuisine use coupons to increase its profit?
- %4 Show that if the unit cost to produce the x^{th} unit of a product is given by ax^{-b} , then doubling the cumulative cost to produce a unit always drops by the same fraction.
- %4 If unit cost drops by 20 percent, when cumulative production doubles, then costs follow an 80 percent learning curve. What value of b corresponds to an 80 percent learning curve?
- %4 (Requires calculus). Suppose the first computer of a new model produced by Lenovo costs \$800 to produce. Suppose previous models follow an 85 percent learning curve and it has produced 4,000 computers. Estimate the cost of producing the next 1,000 computers.

8

Revenue Management

Often you purchase an item that seems identical and the price changes. Here are some examples:

- %4 When I fly from Indianapolis to Los Angeles to visit my daughter, the price I have paid for a roundtrip ticket has varied between \$200 and \$900.
- %4 When I stay in LA at the downtown Marriott, I usually pay \$300 per night for a weekday and under \$200 a night for a weekend.
- %4 When I rent a car from Avis on a weekend I pay much less than the weekday rate.
- %4 The day after Easter I can buy Easter candy much cheaper than the day before!
- %4 My local steakhouse offers a second entrée for half price on Monday–Wednesday.
- %4 It is much cheaper to rent a house in Key West, Florida in the summer than during the winter.
- %4 The Indiana Pacers charge much more for a ticket when they play the Heat.
- %4 Movies at my local theater cost \$5 before 7 p.m. and \$10 after 7 p.m.

All these examples illustrate the use of *revenue management*. Revenue management (often referred to as *yield management*) is used to describe pricing policies used by organizations that sell goods whose value is time-sensitive and usually perishable. For example, after a plane takes off, a seat on the plane has no value. After Easter, Easter candy has reduced value. On April 2, an April 1 motel room has no value.

Revenue management has increased the bottom line for many companies. For example:

- %4 American Airlines credits revenue management with an annual revenue increase of \$500 million.
- %4 In 2003, Marriott reported revenue management increased profit by \$6.7 million.
- %4 Revenue management is credited with saving National Rental Car from bankruptcy.

%4 Part II: Pricing

This chapter explains the basic ideas behind revenue management. The reader who wants to become an expert in revenue management should read Kalyan Talluri and Garrett Ryzin's treatise (*The Theory and Practice of Revenue Management* (Springer-Verlag, 2004)).

The main revenue management concepts discussed in this chapter are the following:

%4 **Making people pay an amount close to their actual valuation for a product.** For example, business travelers are usually willing to pay more for a plane ticket, so revenue management should somehow charge most business travelers a higher price than leisure travelers for a plane ticket.

%4 **Understanding how to manage the uncertainty about usage of the perishable good.** For example, given that some passengers do not show up for a flight, an airline needs to sell more tickets than seats on the plane, or else the plane leaves with some empty seats.

%4 **Matching a variable demand to a fixed supply.** For example, power companies will often charge more for power during a hot summer day in an attempt to shift some of the high power demand during the afternoon to the cooler evening or morning hours.

Estimating Demand for the Bates Motel and Segmenting Customers

In many industries customers can be segmented based on their differing willingness to pay for a product. For example, business travelers (who do not stay in their destination on Saturday nights) are usually willing to pay more for a plane ticket than non-business travelers (who usually stay in their destination on Saturday.) Similarly business travelers (who often reserve a room near the reservation date) often are willing to pay more for a hotel room than a non-business traveler (who usually reserves the room further in advance.) This section uses an example from the hotel industry to show how a business can increase revenue and profit by utilizing the differing valuations of different market segments.

Estimating the Demand Curve

Suppose the 300-room Bates Motel wants to maximize its revenue for June 15, 2016. They asked 2 percent (10 people) of the potential market (business has dropped at the Bates Motel since an incident in 1960) their valuation for a stay at the Motel on

June 15, 2016, and found the results shown in Figure 8-1. This work is in the demand curve worksheet of the file Batesmotel.xlsx

	D	E	F	G
3			2% of market	
4	Customer	Value		
5	1	323		
6	2	151		
7	3	534		
8	4	378		
9	5	358		
10	6	284		
11	7	50		
12	8	113		
13	9	225		
14	10	456		
15				
16				
17		Price	Demand	Price
18		323	5	323
19		151	8	151
20		534	1	534
21		378	3	378
22		358	4	358
23		284	6	284
24		50	10	50
25		113	9	113
26		225	7	225
27		456	2	456

Figure 8-1: Estimating Bates Motel demand curve

You can estimate a demand curve for June 15, 2016, after you have 10 points on the demand curve. For example, 5 of the 10 people have valuations of at least \$323, so if you charge \$323 for a room, 2 percent of your demand equals 5. To compute

4 points on the demand curve, simply list the 10 given valuations in E18:E27. Then copy the formula =COUNTIF(\$E\$5:\$E\$14,">="&E18) from F18 to F19:F27. In cell F18, for example, this formula counts how many people have valuations at least as large as \$323.

You can now use the Excel Trend Curve to find the straight-line demand curve that best fits the data. To do so, follow these steps:

Enter the prices again in the cell range G18:G27 so that you can fit a demand curve with quantity on the x-axis and price on the y-axis.

Select the range F18:G27 and then on the ribbon navigate to Insert > Charts > Scatter. Choose the first option (just dots) to obtain a plot of demand against price.

%4 Right-click the series; select Add Trendline... and choose the Linear option. Then check Display Equation on chart. This yields the chart shown in Figure 8-2, which shows the best-fitting linear demand curve $p = 564.87 - 50.49 * q$

%4 Round this off to simplify calculations; you can assume the demand curve is $p = 565 - 50q$

5. Solve for q in terms of p ; you find: $q = \frac{565 - p}{50}$

%4 Because this is 2 percent of your demand, multiply the previous demand estimate by 50. This cancels out the 50 in the denominator and you find your entire demand is given by $q = 565 - p$

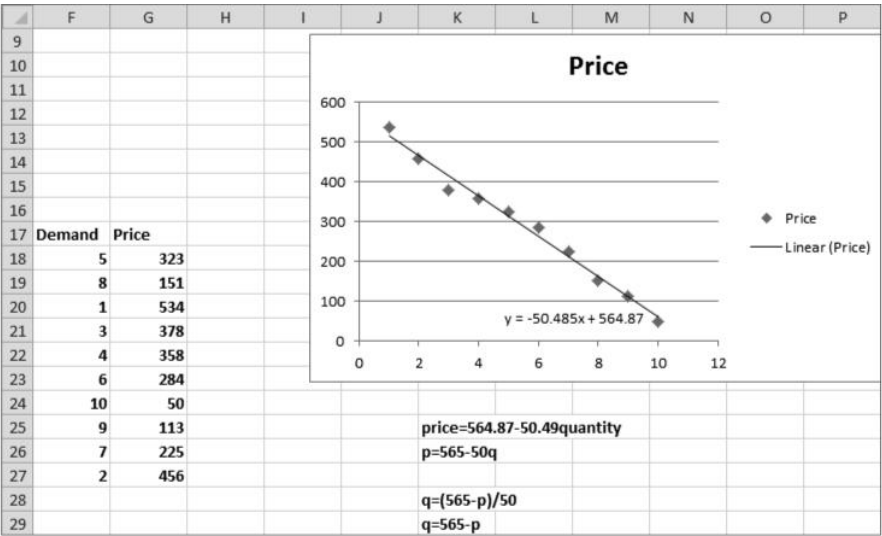


Figure 8-2: Chart of Bates Motel demand curve

Optimal Single Price

To show how revenue management increases profits, you need to determine profits when revenue management is not applied; that is, when each customer is charged the same price.

To simplify your approach, temporarily ignore the 300-room capacity limitation. Under this assumption, if the Bates Motel wants to maximize profit, it should simply choose a price to maximize: profit = price * (565-price). The approach described in Chapter 4, “Estimating Demand Curves and Using Solver to Optimize Price” is now used to determine a single profit maximizing price. The work is in the single price worksheet of the Batesmotel.xlsx file as shown in Figures 8-3 and the Solver



window in Figure 8-4. Maximum revenue of \$79,806.25 is obtained by charging a price of \$282.50.

	F	G	H
5	$q=565-p$		
6			
7			
8	price	\$282.50	
9	demand	282.5	
10	revenue	\$79,806.25	

Figure 8-3: Computing the best single price



Figure 8-4: Solver window to find best single price

Using Two Prices to Segment Customers

To maximize revenue the Bates Motel would love to charge each customer an individual valuation. However, this is illegal; each customer must be charged the same

price. Yield management, however, provides a legal way to price discriminate and increase profits.

If there is a variable highly correlated with a customer's willingness to pay, then Bates can approximate the individual pricing strategy. Suppose that low-valuation customers are willing to purchase early, and high-value customers usually purchase near the date of the reservation. Bates can approximate the individual pricing strategy by charging a price *low* for advance purchases and a price *high* for last minute purchases. Assuming that all customers with high valuations arrive at the last minute this would result in a demand $565 - \text{high}$ for last-minute sales and a demand $(565 - \text{low}) - (565 - \text{high})$ for advance sales. The Solver model to determine the revenue maximizing high and low prices is in the worksheet (see Figure 8-5.) You proceed as follows:

Segmentation

	E	F	G	H	I	J	K
1							
2							
3							
4							
5							
6	q=565-p			For			
7				all customers >=high charge high			
8				rest charge low			
9							
10	high	\$376.67					
11	low	\$188.33					
12	highdemand	188.33366					
13	lowdemand	188.333449					
14	highrevenue	\$70,938.95					
15	Lowrevenue	\$35,469.38					
16	Totalrevenue	\$106,408.33					

Figure 8-5: Model for computing prices with segmentation

%4 In cell F12 compute the demand for last-minute reservations with the formula
 =565-high

%4 In cell F13 compute the advance demand with the formula
 $\text{=(565-low)-(565-high)}$

%4 In cell F14 compute the revenue from last-minute reservations with the
 formula =high*highdemand

%4 In cell F15 compute the revenue from advance reservations with the formula
 =lowdemand*low

%4 In cell F16 compute the total revenue with the formula =SUM(F14:F15)

%4 As shown in the Solver window (refer to Figure 8-6), you choose non-negative high and low prices to maximize total revenue.



Figure 8-6: Solver window with segmentation

You'll find that a high price of \$376.67 should be charged for last-minute reservations and a low price of \$188.33 for advance reservations. Bates' revenue has now increased by 33 percent to \$106,408.33. Of course, this revenue increase would be realized only if there were a perfect correlation between willingness to pay and a customer wanting to buy in advance at the last minute. If this correlation existed, it provides a legal way for Bates to charge higher prices to high-valuation customers. In this situation, Bates can add a qualifying restriction (say, reserve rooms at least 2 weeks in advance) or a *fence* that separates high-valuation and low-valuation customers. Airlines often use similar tactics by using the qualifying restriction of "staying over a Saturday night" as an imperfect fence to separate leisure and business travelers.

This solution, however, resulted in more people showing up than the hotel has rooms. You can resolve this issue by adding constraints on received reservations.

%4 Part II: Pricing

Segmentation with Capacity Constraints

In Figure 8-7, you can see how to ensure that the number of rooms reserved will not exceed capacity. After all you don't want Norman to stay in his mother's room. To avoid this problem, copy the Segmentation worksheet to a new worksheet named Segmentation with capacity. Then compute total rooms reserved in cell H16 and add a constraint that $H16 \leq J16$. As shown in Figure 8-7, \$415 should be charged for last-minute reservations and \$265 for advance reservations. One hundred and fifty reservations of each type will be purchased and the revenue drops by approximately 4 percent to \$102,000. Note that the new prices result in the number of reservations equaling capacity.

	E	F	G	H	I	J	K
6	$q=565-p$						
7				All customers $\geq$ high charge high			
8				rest charge low			
9							
10	high	\$415.00					
11	low	\$265.00					
12	advancedemand	150.00					
13	lastmindemand	150.00					
14	advancerevenue	\$62,250.00					
15	lastminrevenue	\$39,750.00		Usage		Capacity	
16	Totalrevenue	\$102,000.00		300.00 $\leq$		300	

Figure 8-7: Optimal prices with capacity restriction

Handling Uncertainty

The analysis of the Bates Motel in the previous section implicitly assumed that when prices were set Bates knew *exactly* how many people would reserve in advance and at the last minute. Of course, this is not the case. In this section you learn how Bates should deal with this uncertainty. When developing revenue management systems, airlines also deal with uncertainty, such as the number of people that will fail to show up for a flight.

Determining a Booking Limit

To illustrate the role of uncertainty in revenue management, assume all advance reservations arrive before all last-minute reservations. Bates charges \$105 for an advance reservation and \$159 for a last-minute reservation. Also assume that sufficient customers want to reserve in advance to fill all the rooms. Because Bates does not know how many last-minute reservations will

occur, it “protects” a certain number of rooms for last-minute reservations.

%4 *protection limit* is the number of rooms *not* sold to advance reservations because late arrivals are willing to pay more for a room. Alternatively, a *booking limit* is the maximum number of rooms reserved for advance purchases. For example, a booking limit of 200 means Bates will allow at most 200 rooms to be reserved at the \$105 price. Of course a booking limit of 200 rooms is equivalent to a protection limit of $300 - 200 = 100$ rooms.

Assume the number of last-minute reservations that Bates receives is unknown and follows a normal random variable with a mean of 100 and a standard deviation of 20. This implies there is a 68 percent chance that between 80 and 120 last-minute reservations will be received, and a 95 percent chance that between 60 and 140 last-minute reservations will be received.

Now you can determine the protection level Q that maximizes Bates’ expected profit using the powerful concept of *marginal analysis*. In marginal analysis you try and determine an optimal value of a decision variable by comparing the benefits of a unit change in the variable to the cost incurred from a unit change in a variable. To apply marginal analysis to the determination of the optimal protection level you check if for a given value of Q , Bates can benefit by reducing the protection level from $Q + 1$ to Q . To do so perform these steps:

Define $F(Q)$ the probability number of last-minute reservations as less than or equal to Q , or $F(Q) = P(\text{last minute reservations} \leq Q)$

Because you assumed all rooms could be filled at the discounted price, reducing the protection level by 1 would surely gain Bates \$105.

With probability $1 - F(Q)$ Bates would sell the $Q + 1$ protected room at a \$159 price, so on average Bates would lose $(1 - F(Q)) * 159 + F(Q) * 0$

$= (1 - F(Q)) * 159$ if it reduces the protection level by 1. Therefore Bates should reduce the protection level from $Q + 1$ to Q if and only if

$$105 \geq (1 - F(Q)) * 159 \text{ or } F(Q) \geq 54/159 = .339$$

%4 You can find the 33.9 percentile of last-minute reservations using the Excel NORMINV function. NORMINV(probability, mean, standard_dev) gives the value Q for a normal random variable with a given mean and standard deviation such that $F(Q) = p$. Enter =NORMINV(0.339,100,20) into Excel. This yields 91.70. Therefore $F(91) < .339$ and $F(92) > .339$, so Bates Motel should protect 92 rooms.

Overbooking Models

Airlines usually have several fare classes, so they must determine more than one booking limit. As the time of the flight approaches, airlines update the booking

%4 Part II: Pricing

limits based on the number of reservations received. This updating requires lots of past data on similar flights. In most cases, revenue management requires a large investment in information technology and data analysis, so the decision to institute a revenue management program should not be taken lightly.

Airlines always deal with the fact that passengers who have a ticket may not show up. If airlines do not “overbook” the flight by selling more tickets than seats, most flights leave with empty seats that could have been filled. Of course, if they sell too many tickets, they must give “bumped” passengers compensation, so airlines must trade off the risk of empty seats against the risk of overbooking. Marginal analysis can also be used to analyze this trade-off problem.

To illustrate the idea, suppose the price for a New York to Indianapolis flight on Fly by Night (FBN) airlines is \$200. The plane seats 100 people. To protect against no-shows, FBN tries to sell more than 100 tickets. Federal Law requires that any ticketed customer who cannot board the plane is entitled to \$100 compensation. Past data indicates the number of no-shows for this flight follows a normal random variable with mean 20 and standard deviation 5. To maximize expected revenue less compensation costs for the flight, how many tickets should FBN try to sell for each flight? Assuming that unused tickets are refundable, follow these steps:

Let Q equal number of tickets FBN will try to sell and NS equal number of no shows. You can model NS as a continuous random variable, so NS can assume a fractional value. Therefore assume, for example, that if $Q - NS$ is between 99.5 and 100.5, then 100 passengers will show up.

For a given value of Q , consider whether you should reduce Q from $Q + 1$ to Q . If $Q - NS \geq 99.5$, then you save \$100 by reducing ticket limit from $Q + 1$ to Q . This is because one less person will be overbooked. On the other hand, if $Q - NS < 99.5$, then reducing the ticket limit from $Q + 1$ to Q results in one less ticket sale, which reduces revenue by \$200. If $F(x)$ = Probability number of no-shows is less than or equal to x , then

With Probability $F(Q - 99.5)$ reducing the ticket limit from $Q + 1$ to Q saves \$100.

With probability $1 - F(Q - 99.5)$ reducing the ticket limit from $Q + 1$ to Q costs \$200.

Therefore reducing the ticket limit from $Q + 1$ to Q benefits you if:

$$100F(Q - 99.5) - 200(1 - F(Q - 99.5)) \geq 0$$

or
200

$$F(Q - 99.5) \geq \frac{200 - 100}{200 + 100} = .667$$

%4 Now $\text{NORMINV}(0.667, 20, 5) = 22.15$. This implied that if and only if $Q - 99.5 \geq 22.15$ or $Q \geq 121.65$ you should reduce Q from $Q + 1$ to Q . Therefore you should reduce tickets sold from 123 to 122 and stop there. In short, to maximize expected revenue less compensation costs, FBN should cut ticket sales off at 122 tickets.

The problem faced by the airlines is much more complicated than this simple overbooking model. At every instant the airline must update their view of how many tickets will be sold for the flight and use this information to determine optimal decisions on variables such as booking limits.

Markdown Pricing

Many **retailers practice revenue management** by reducing a product's price based on season or timing. For example, bathing suits are discounted at the end of the summer. Also Easter candy and Christmas cards are discounted after the holiday. The now defunct Filene's Basement of Boston for years used the following markdown policy:

- %4 Twelve days after putting an item on sale, the price was reduced by 25%.
- %4 Six selling days later, the price is cut by 50%.
- %4 After an additional 6 selling days, the items were offered at 75% off the original price.
- %4 After 6 more selling days, the item was given to charity.

The idea behind markdown pricing is that as time goes on, the value of an item to customers often falls. This logic especially applies to seasonal and perishable goods. For example, a bathing suit purchased in April in Indiana has much more value than

%4 bathing suit purchased in September because you can wear the April-purchased suit for many more upcoming summer days. Likewise, to keep perishable items from going bad before they are sold supermarkets markdown prices when a perishable item gets near its expiration date. Because customers only buy products if perceived value exceeds cost, a reduction in perceived value necessitates a reduction in price if you still want to maintain a reasonable sales level. The following example further illustrates mark-down pricing.

Consider a store that sells a product (such as swimsuits) over a three-month period. The product is in demand most when it first hits the stores. In the workbook *Markdownpricing.xlsx* you will determine how pricing strategy maximizes profit. Figure 8-8 shows the markdown pricing spreadsheet model.

%4 Part II: Pricing

Suppose you have 400 swimsuits to sell and the methods of Chapter 4 have been used to estimate the following demand curves:

%4 Month 1 Demand = $300 - \text{price}$

%4 Month 2 Demand = $300 - 1.3\text{price}$

%4 Month 3 Demand = $300 - 1.8\text{price}$

	C	D	E	F	G	H	I	J	K
1									
2		Markdown Pricing							
3		Sell our 400 swimsuits							
4									
5		Month	1	2	3				
6		intercept	300	300	300				
7		slope	-1	-1.3	-1.8		total sold		available
8		price	\$162.20	\$127.58	\$95.53		400 <=		400
9		demand	137.80488	134.1463408	128.0487804				
10		revenue	\$22,351.28	\$17,114.35	\$12,232.30				
11									
12			total revenue						
13			\$51,697.94						

Figure 8-8: Markdown pricing model

To determine the sequence of prices that maximizes your revenue, proceed as follows (see the order 400 worksheet):

%4 Solver requires you to start with numbers in the changing cells, so in E8:G8 enter trial values for each week's price.

%4 Copy the formula =E6+E7*E8 from E9 to E9:G9 to generate the actual demand each month.

%4 Copy the formula =E8*E9 from E10 to F10:G10 to compute each month's revenue.

%4 The total revenue is computed in cell E13 with the formula =SUM(E10:G10)

You can maximize revenue (E13) by changing prices (E8:G8) and constrain units sold (I8) to equal 400. The Solver window is shown in Figure 8-9.

A maximum revenue of \$51,697.94 is obtained by charging \$162.20 during Month 1, \$127.58 during Month 2, and \$95.53 during Month 3. Of course, the prices are set to ensure that all 400 swimsuits are sold.

Your model would be more realistic if the store realizes that it should also try to optimize the number of swimsuits it buys at the beginning of the season. Assume the store must pay \$100 per purchased swimsuit. In the how many worksheet (see Figure 8-10) you can use Solver to determine prices that maximize profit.

In E10:G10 you can revise the profit formulas to include the purchase cost by copying the formula $= (E8 - \text{cost}) * E9$ from E10 to F10:G10. After deleting the constraint $I8 = 400$, the Solver window is identical to Figure 8-9. Profit is maximized at a value of \$17,557.69 by buying 245 swimsuits and charging \$200 in Month 1, \$165.38 in Month 2, and \$133.33 in Month 3.



Figure 8-9: Solver window for markdown pricing

	C	D	E	F	G	H	I
1	Maximize Profit						
2	Dynamic Pricing						
3		cost	\$100.00				
4							
5		Month	1	2	3		
6		intercept	300	300	300		
7		slope	-1	-1.3	-1.8		total
8		price	\$200.00	\$165.38	\$133.33		245
9		demand	100	85.00000009	60		
10		profit	\$10,000.00	\$5,557.69	\$2,000.00		
11							
12			total profit				
13			\$17,557.69				

Figure 8-10: Markdown pricing and purchase decision

%4 Part II: Pricing

Summary

In this chapter you learned the following:

- %4 Revenue management enables organizations including airlines, hotels, rental car agencies, restaurants, and sports teams to increase profits by reducing the unused amount of perishable inventory (seats, hotel rooms, and so on). Revenue management also enables organizations to better match the price charged to customers with what they are willing to pay.
- %4 Revenue management has a much greater chance of succeeding if “fences” (such as staying over Saturday night for airlines) exist to separate high-valuation customers from low-valuation customers.
- %4 To handle the fact that organizations do not know how many high-value customers will demand a product, organizations often set booking limits to constrain low-price sales so that more capacity is reserved for late arriving high-valuation customers.
- %4 To adjust for the possibility of no shows, organizations need to sell more capacity than is available.
- %4 Marginal analysis is helpful to solve booking limits and overbooking problems.
- %4 Often the valuation customers have for products drops over time. This requires that retailers lower or mark down their prices over time.

Exercises

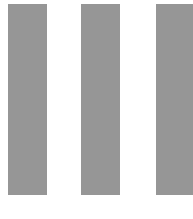
- %4 **Redo the analysis** in the first section, “Estimating Demand for the Bates Motel and Segmenting Customers,” assuming demand follows a constant elasticity demand curve. Use the Power Curve option on the Excel Trendline to fit the demand curve.
- %4 How can TV networks use revenue management?
- %4 How can Broadway plays use revenue management?
- %4 A flight from New York to Atlanta has 146 seats. Advance tickets purchased cost \$74. Last-minute tickets cost \$114. Demand for full-fare tickets is normally distributed with a mean of 92 and standard deviation of 30. What booking limit maximizes expected revenues? Assume there are no no-shows and always enough advanced purchasers to fill the flight.
- %4 Suppose a Marriott offers a \$159 discount rate for a midweek stay. Its regular rate is \$225. The hotel has 118 rooms. Suppose it is April 1 and the Marriott wants to maximize profit from May 29 bookings. The Marriott knows it can

fill all rooms at the discounted price, but to maximize profit it must reserve or protect some rooms at the high price. Because business travelers book late, the hotel decides to protect or reserve rooms for late-booking business customers. The question is how many rooms to protect. The number of business travelers who will reserve a room is unknown, and you can assume it is normal with the mean = 27.3 and a standard deviation of 6. Determine the Protection Limit that maximizes the expected profit. Again assume that there are always enough leisure travelers to pay the discount rate for the unsold rooms.

%4 The Atlanta to Dallas FBN flight has 210 seats and the fare is \$105. Any overbooked passenger costs \$300. The number of no-shows is normal with a mean of 20 and standard deviation of 5. All tickets are nonrefundable. How many reservations should FBN accept on this flight?

%4 The pre-Christmas demand for Christmas cards at a local Hallmark stores is given by $q = 2000 - 300p$. The demand for Christmas cards after Christmas is given by $q = 1000 - 400p$. If the store pays \$1 per card, how can they maximize profits from Christmas cards? Assume they want all inventory sold.





Forecasting

Chapter 9: Simple Linear Regression
and Correlation

Chapter 10: Using Multiple Regression to Forecast
Sales

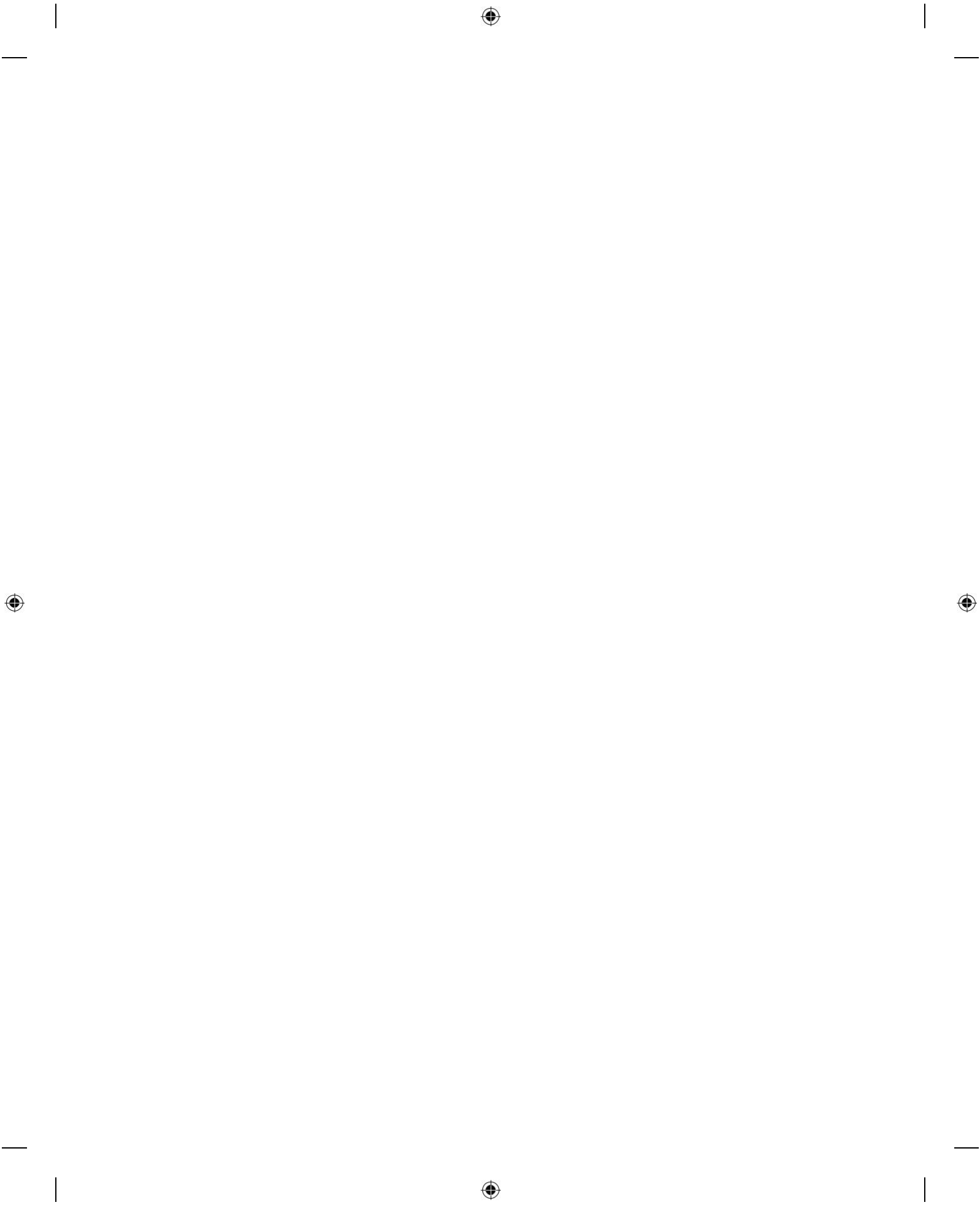
Chapter 11: Forecasting in the Presence of
Special Events

Chapter 12: Modeling Trend and Seasonality

Chapter 13: Ratio to Moving Average Forecast
Method

Chapter 14: Winter's Method

Chapter 15: Using Neural Networks to Forecast
Sales



9

Simple Linear Regression and Correlation

Often the marketing analyst needs to determine how variables are related, and much of the rest of this book is devoted to determining the nature of relationships between variables of interest. Some commonly important marketing questions that require analyzing the relationships between two variables of interest include:

- %4 How does price affect demand?
- %4 How does advertising affect sales?
- %4 How does shelf space devoted to a product affect product sales?

This chapter introduces the simplest tools you can use to model relationships between variables. It first covers finding the line that best fits the hypothesized causal relationship between two variables. You then learn to use correlations to analyze the nature of non-causal relationships between two or more variables.

Simple Linear Regression

Every business analyst should **have the ability** to estimate the relationship between important business variables. In Microsoft Office Excel, the Trendline feature can help you determine the relationship between two variables. The variable you want to predict is the *dependent variable*. The variable used for prediction is the *independent variable*. Table 9-1 shows some examples of business relationships you might want to estimate.

Table 9-1: Examples of Relationships

Independent Variable	Dependent Variable
Units produced by a plant in 1 month	Monthly cost of operating a plant
Dollars spent on advertising in 1 month	Monthly sales
Number of employees	Annual travel expenses
Daily sales of cereal	Daily sales of bananas
Shelf space devoted to chocolate	Sales of chocolate
Price of bananas sold	Pounds of bananas sold

The first step to determine how two variables are related is to graph the data points so that the independent variable is on the x-axis and the dependent variable is on the y-axis. You can do this by using the Scatter Chart option in Microsoft Excel and performing the following steps:

- %4 With the Scatter Chart option selected, click a data point (displayed in blue) and click Trendline in the Analysis group on the Chart Tools Layout tab.
- %4 Next click More Trendline Options..., or right-click and select Add Trendline... You'll see the Format Trendline dialog box, which is shown in Figure 9-1.
- %4 If your graph indicates that a straight line can be drawn that provides a reasonable fit (a reasonable fit will be discussed in the "Defining R^2 " section of this chapter) to the points, choose the Linear option. Nonlinear relationships are discussed in the "Modeling Nonlinearities and Interactions" section of Chapter 10, "Using Multiple Regression to Forecast Sales."

Analyzing Sales at Mao's Palace Restaurant

To illustrate how to model a linear relationship between two variables, take a look at the daily sales of products at Mao's Palace, a local Chinese restaurant (see Figure 9-2). Mao's main product is bowls filled with rice, vegetables, and meat made to the customer's order. The file *Maospalace.xlsx* gives daily unit sales of bowl price, bowls, soda, and beer.

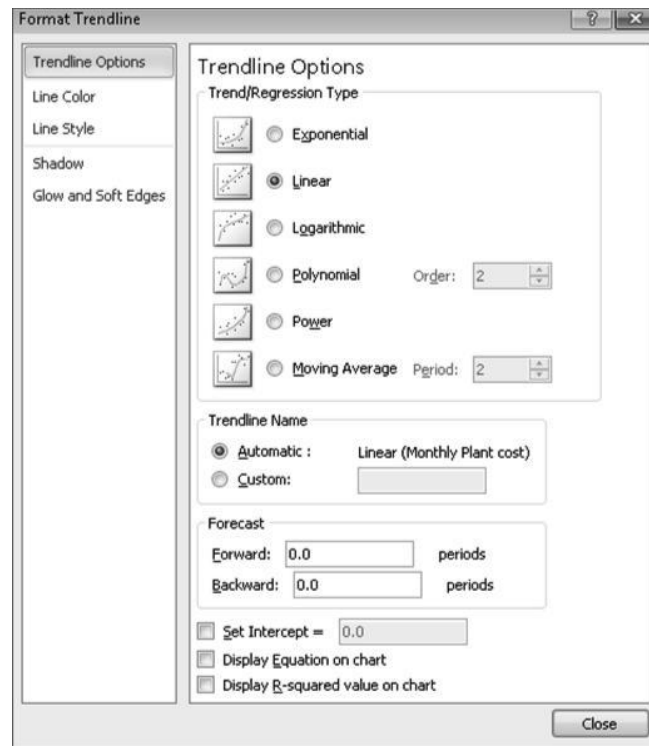


Figure 9-1: Trendline dialog box

Now suppose you want to determine how the price of the bowls affects daily sales. To do this you create an XY chart (or a scatter plot) that displays the independent variable (price) on the x-axis and the dependent variable (bowl sales) on the y-axis. The column of data that you want to display on the x-axis must be located to the left of the column of data you want to display on the y-axis. To create the graph, you perform two steps:

- %4 Select the data in the range E4:F190 (including the labels in cells E4 and F4).
- %4 Click Scatter in the Charts group on the Insert tab of the Ribbon, and select the first option (Scatter with only Markers) as the chart type. Figure 9-3 shows the graph.

Part III: Forecasting

	E	F	G	H
1				
2				
3				
4	Bowl Price	Bowls	Soda	Beer
5	\$9.30	391	313	90
6	\$9.10	418	326	100
7	\$8.50	459	358	115
8	\$9.50	424	331	81
9	\$8.70	447	380	89
10	\$9.70	383	291	92
11	\$9.80	399	307	96
12	\$8.80	440	361	66
13	\$8.60	436	344	74
14	\$9.60	413	351	62
15	\$8.20	428	338	64
16	\$8.00	479	374	101
17	\$8.10	462	388	69
18	\$9.80	387	325	77
19	\$8.90	454	341	114
20	\$9.40	418	314	88
21	\$8.30	447	375	107
22	\$9.60	442	376	102
23	\$9.90	381	312	95
24	\$9.30	401	301	68
25	\$8.10	468	370	70
26	\$8.70	428	321	64
27	\$8.10	480	374	115

Figure 9-2: Sales at Mao's Palace

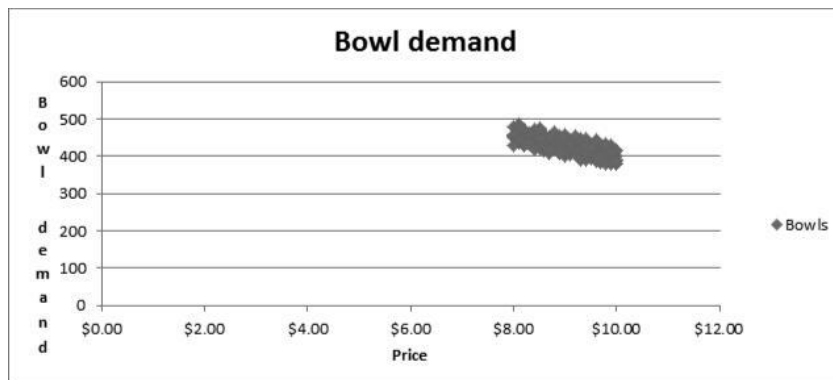


Figure 9-3: Scatterplot of Bowl demand versus Price

If you want to modify this chart, you can click anywhere inside the chart to display the Chart Tools contextual tab. Using the commands on the Chart Tools Design tab, you can do the following:

- %4 Change the chart type.
- %4 Change the source data.
- %4 Change the style of the chart.
- %4 Move the chart.

Using the commands on the Chart Tools Layout tab, you can do the following:

- %4 Add a chart title.
- %4 Add axis labels.
- %4 Add labels to each point that gives the x and y coordinate of each point.
- %4 Add gridlines to the chart.

Looking at the scatter plot, it seems reasonable that there is a straight line (or linear relationship) between the price and bowl sales. You can see the straight line that “best fits” the points by adding a trend line to the chart. To do so, perform the following steps:

- %4 Click within the chart to select it, and then click a data point. All the data points display in blue with an X covering each point.
- %4 Right-click and then click Add Trendline...
- %4 In the Format Trendline dialog box, select the Linear option, and then check the Display Equation on chart and the Display R-squared value on chart boxes, as shown in Figure 9-4. The R-Squared Value on the chart is defined in the “Defining R^2 ” section of this chapter.
- %4 Click Close to see the results shown in Figure 9-5. To add a title to the chart and labels for the x-and y-axes, select Chart Tools, click Chart Title, and then click Axis Titles in the Labels group on the Layout tab.
- %4 To add more decimal points to the equation, select the trend-line equation; after selecting Layout from Chart Tools, choose Format Selection.
- %4 Select Number and choose the number of decimal places to display.

%4 Part III: Forecasting

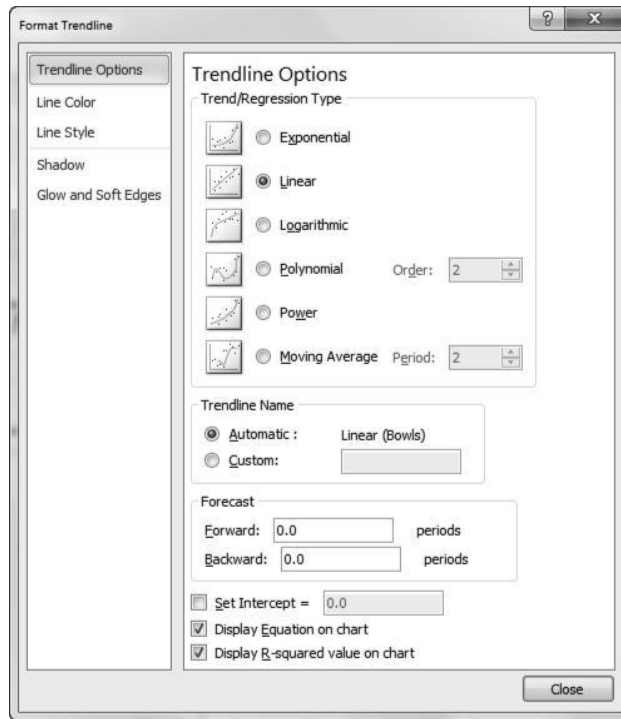


Figure 9-4: Trendline settings for Bowl demand

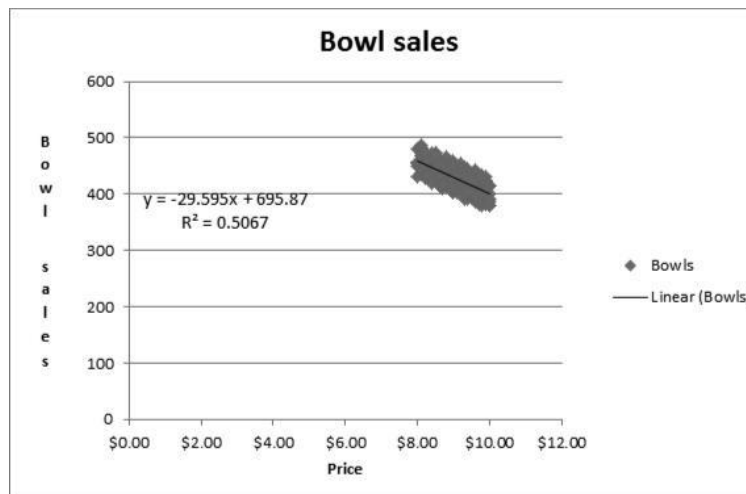


Figure 9-5: Trendline for Bowl demand

How Excel Determines the Best-Fitting Line

When you create a scatter chart and plot a trend line using the Trendline feature, it chooses the line that minimizes (over all lines that could be drawn) the sum of the squared vertical distance from each point to the line. The vertical distance from each point to the line is an *error*, or *residual*. The line created by Excel is called the *least squares line*. You minimize the sum of squared errors rather than the sum of the errors because in simply summing the errors, positive and negative errors can cancel each other out. For example, a point 100 units above the line and a point 100 units below the line cancel each other if you add errors. If you square errors, however, the fact that your predictions for each point are wrong will be used by Excel to find the best-fitting line. Another way to see that minimizing the sum of squared errors is reasonable is to look at a situation in which all points lie on one line. Then minimizing the least squares line would yield this line and a sum of squared errors equal to 0.

Thus, Excel calculates that the best-fitting straight line for predicting daily bowl sales from the price by using the equation $\text{Daily Bowl Sales} = -29.595 \times \text{Price} + 695.87$. The -29.595 slope of this line indicates that the best guess is that a \$1 increase in the price of a bowl reduces demand by 29.595 bowls.

WARNING You should not use a least-squares line to predict values of an independent variable that lies outside the range for which you have data. Your line should be used only to predict daily bowl sales for days in which the bowl price is between \$8 and \$10.

Computing Errors or Residuals

Referring back to the Mao's Palace example, you can compute predicted bowl sales for each day by copying the formula $= -29.595 \times E5 + 695.87$ from C5 to C6:C190. Then copy the formula $= F5 - C5$ from D5 to D6:D190. This computes the errors (or residuals). These errors are shown in Figure 9-6. For each data point, you can define the error by the amount by which the point varies from the least-squares line. For each day, the error equals the observed demand minus the predicted demand. A positive error indicates a point is above the least-squares line, and a negative error indicates that the point is below the least-squares line. In cell D2, the sum of the errors is computed, which obtained 1.54. In reality, for any least-squares line, the sum of the errors should equal 0. 1.54 is obtained because the equation is rounded

%4 Part III: Forecasting

to three decimal points.) The fact that errors sum to 0 implies that the least-squares line has the intuitively satisfying property of splitting the points in half.

B	C	D	E	F	G	H
		sum errors				
		1.537				
	Predicted					
	Bowl Sales	Error	Bowl Price	Bowls	Soda	Beer
	420.6365	-29.6365	\$9.30	391	313	90
	426.5555	-8.5555	\$9.10	418	326	100
	444.3125	14.6875	\$8.50	459	358	115
	414.7175	9.2825	\$9.50	424	331	81
	438.3935	8.6065	\$8.70	447	380	89
	408.7985	-25.7985	\$9.70	383	291	92
	405.839	-6.839	\$9.80	399	307	96
	435.434	4.566	\$8.80	440	361	66
	441.353	-5.353	\$8.60	436	344	74
	411.758	1.242	\$9.60	413	351	62
	453.191	-25.191	\$8.20	428	338	64
	459.11	19.89	\$8.00	479	374	101

Figure 9-6: Errors in predicting Bowl demand

Defining R^2

As you can see in the Mao's Palace example, each day both the bowl price and bowl sales vary. Therefore it is reasonable to ask what percentage of the monthly variation in sales is explained by the daily variation in price. In general the percentage of the variation in the dependent variable explained by the least squares line is known as R^2 . For this regression the R^2 value is 0.51, which is shown in Figure 9-5. You can state that the linear relationship explains 51 percent of the variation in monthly operating costs.

Once you determine the R^2 value, your next question might be what causes the other 49 percent of the variation in daily bowl sales costs. This value is explained by various other factors. For example, the day of the week and month of the year might affect bowl sales. Chapter 10, "Using Multiple Regression to Forecast Sales" explains how to use *multiple regression* to determine other factors that influence operating costs. In most cases, finding factors that increase R^2 increases prediction accuracy. If a factor only results in a slight increase in R^2 , however, using that factor to predict the dependent variable can actually decrease forecast accuracy. (See Chapter 10 for further discussion of this idea.)

Another question that comes up a lot in reference to R^2 values is what is a good R^2 value? There is no definitive answer to this question. As shown in Exercise 5 toward the end of the chapter, a high R^2 can occur even when

a trend line is not a good predictor of y . With one independent variable, of course, a larger R^2 value indicates a better fit of the data than a smaller R^2 value.

%4 better measure of the accuracy of your predictions is the *standard error of the regression*, described in the next section.

Accuracy of Predictions from a Trend Line

When you fit a line to points, you obtain a standard error of the regression that measures the *spread* of the points around the least-squares line. You can compute the standard error associated with a least-squares line with the STEYX function. The syntax of this function is STEYX(known_y's, known_x's) , where *yrange* contains the values of the dependent variable, and *xrange* contains the values of the independent variable. To use this function, select the range E4:F190 and use FORMULAS CREATE FROM SELECTION to name your price data Bowl_Price and your sales data Bowls . Then in cell K1, compute the standard error of your cost estimate line with the formula =STEYX(Bowls,Bowl_Price) . Figure 9-7 shows the result.

	J	K	L
1			
2	STD ERROR	17.41867	
3	SLOPE	-29.5945	
4	INTERCEPT	695.8741	
5	RSQ	0.506748	

Figure 9-7: Computing standard error of the regression

Approximately 68 percent of your points should be within one standard error of regression (SER) of the least-squares line, and approximately 95 percent of your points should be within two SER of the least-squares line. These measures are reminiscent of the descriptive statistics rule of thumb described in Chapter 2, “Using Excel Charts to Summarize Marketing Data.” In your example, the absolute value of approximately 68 percent of the errors should be 17.42 or smaller, and the absolute value of approximately 95 percent of the errors should be 34.84, or $2 * 17.42$, or smaller. You can find that 57 percent of your points are within one SER of the least-squares line, and all (100 percent) of the points are within two standard SER of the least-squares line. Any point that is more than two SER from the least-squares line is called an *outlier*.

Looking for causes of outliers can often help you to improve the operation of your business. For example, a day in which actual demand was 34.84 higher than anticipated would be a demand outlier on the high side. If you ascertain the cause of this high sales outlier and make it recur, you would clearly improve profitability.

Similarly, consider a month in which actual sales are over 34.84 less than expected. If you can ascertain the cause of this low demand outlier and ensure it occurred less often, you would improve profitability. Chapters 10 and 11 explain how to use outliers to improve forecasting.

The Excel Slope, Intercept, and RSQ Functions

You have learned how to use the Trendline feature to find the line that best fits a linear relationship and to compute the associated R^2 value. Sometimes it is more convenient to use Excel functions to compute these quantities. In this section, you learn how to use the Excel SLOPE and INTERCEPT functions to find the line that best fits a set of data. You also see how to use the RSQ function to determine the associated R^2 value.

The Excel SLOPE(*known_y's*, *known_x's*) and INTERCEPT(*known_y's*, *known_x's*) functions return the slope and intercept, respectively, of the least-squares line. Thus, if you enter the formula SLOPE(Bowls, Bowl_Price) in cell K3 (see Figure 9-7) it returns the slope (–29.59) of the least-squares line. Entering the formula INTERCEPT(Bowls, Bowl_Price) in cell K4 returns the intercept (695.87) of the

least-squares line. By the way, the

RSQ(*known_y's*, *known_x's*)

function returns the

Bowl_Price)

R^2 value associated with a least-squares line. So, entering the formula RSQ(Bowls, in cell K5 returns the R^2 value of 0.507 for your least-squares line. Of course this R^2 value is identical to the RSQ value obtained from the Trendline.

Using Correlations to Summarize Linear Relationships

Trendlines are a great way to understand how two variables are related. Often, however, you need to understand how more than two variables are related. Looking at the correlation between any pair of variables can provide insights into how multiple variables move up and down in value together. Correlation measures linear association, not causation.

The correlation (usually denoted by r) between two variables (call them x and y) is a unit-free measure of the strength of the linear relationship between x and y . The correlation between any two variables is always between -1 and $+1$. Although the exact formula used to compute the correlation between two variables isn't very important, interpreting the correlation between the variables is.

A correlation near $+1$ means that x and y have a strong positive linear relationship. That is, when x is larger than average, y is almost always larger than average, and when

x is smaller than average, y is almost always smaller than average. For example, for the data shown in Figure 9-8, (x = units produced and y = monthly production cost) x and y have a correlation of $+0.95$. You can see that in Figure 9-8 the least squares line fits the points very well and has a positive slope which is consistent with large values of x usually occurring with large values of y

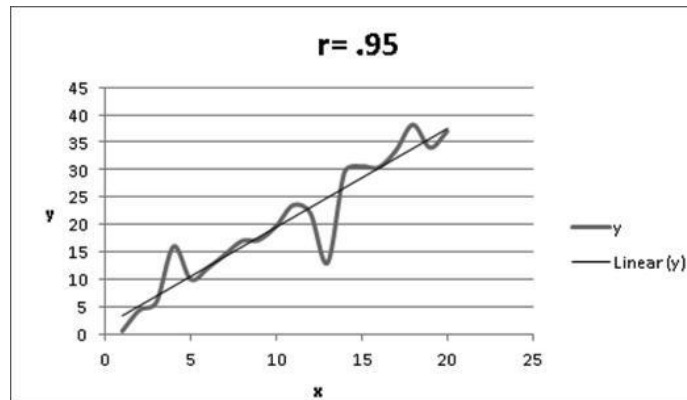


Figure 9-8: Correlation = $+0.95$

If x and y have a correlation near -1 , this means that there is a strong negative linear association between x and y . That is, when x is larger than average, y is usually smaller than average, and when x is smaller than average, y is usually larger than average. For example, for the data shown in Figure 9-9, x and y have a correlation of -0.90 . You can see that in Figure 9-9 the least squares line fits the points very well and has a negative slope which is consistent with large values of x usually occurring with small values of y

%4 correlation near 0 means that x and y have a weak linear association. That is, knowing whether x is larger or smaller than its mean tells you little about whether y will be larger or smaller than its mean. Figure 9-10 shows a graph of the dependence of unit sales (y) on years of sales experience (x). Years of experience and unit sales have a correlation of 0.003. In the data set, the average experience is 10 years. You can see that when a person has more than 10 years of sales experience, sales can be either low or high. You also see that when a person has fewer than 10 years of sales experience, sales can be low or high. Although experience and sales have little or no linear relationship, there is a strong nonlinear relationship (see the fitted curve in Figure 9-10) between years of experience and sales. Correlation does not measure the strength of nonlinear associations.

%4 Part III: Forecasting

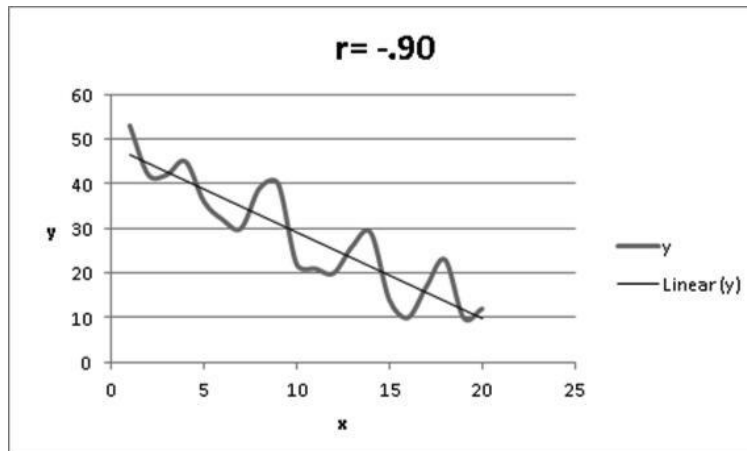


Figure 9-9: Correlation = -0.90

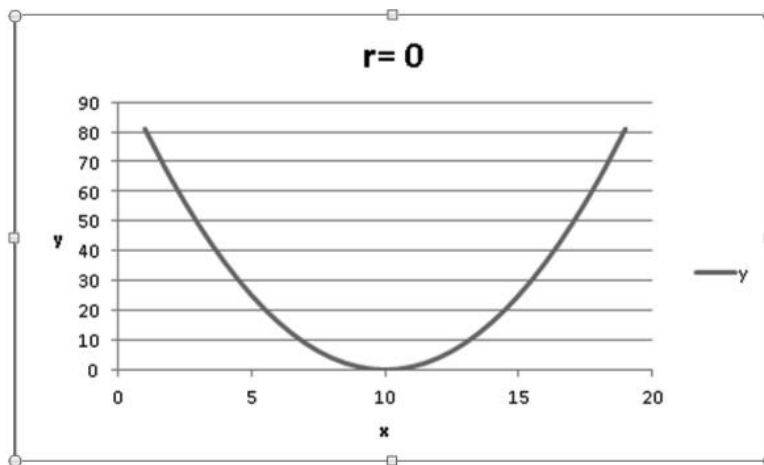


Figure 9-10: Correlation near 0

Finding a Correlation with the Data Analysis Add-In

You will now learn how Excel's Data Analysis Add-in and the Excel Correlation function can be used to compute correlations. The Data Analysis Add-In makes it easy to find correlations between many variables. To install the Data Analysis Add-in, perform the following steps:

- %4 Click the File tab and select Options.
- %4 In the Manage box click Excel Add-Ins, and choose Go.
- %4 In the Add-Ins dialog box, select Analysis ToolPak and then click OK.

Now you can access the Analysis ToolPak functions by clicking Data Analysis in the Analysis group on the Data tab.

You can use this functionality to find the correlations between each pair of variables in the Mao's Palace data set. To begin select the Data Analysis Add-In, and choose Correlation. Then fill in the dialog box, as shown in Figure 9-11.

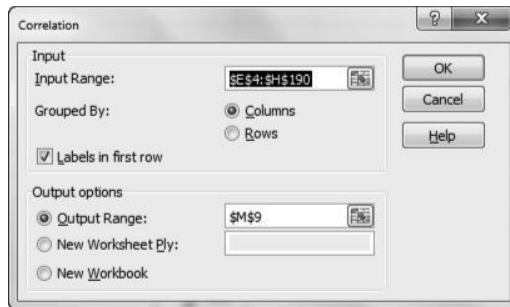


Figure 9-11: Correlation dialog box

To compute correlations with the Data Analysis Add-in proceed as follows:

- %4 Select the range which contains the relevant data and data labels. The easiest way to accomplish this is to select the upper-left cell of the data range (E5) and then press Ctrl+Shift+Right Arrow, followed by Ctrl+Shift+Down Arrow.
- %4 Check the Labels In First Row option because the first row of the input range contains labels. Enter cell M9 as the upper-left cell of the output range.
- %4 After clicking OK, you see the results, as shown in Figure 9-12.

	M	N	O	P	Q
3					
4					
5		Price Bowl Correlation			
6		-0.71186			
7					
8					
9		Bowl Price	Bowls	Soda	Beer
10	Bowl Price	1			
11	Bowls	-0.71186	1		
12	Soda	-0.58095	0.831008	1	
13	Beer	-0.19367	0.338691	0.246803	1

Figure 9-12: Correlation matrix

From Figure 9-12, you find there is a -0.71 correlation between Bowl Price and Bowl Sales, indicating a strong negative linear association. The $.083$ correlation

%4 Part III: Forecasting

between Soda Sales and Bowl Sales indicates a strong positive linear association. The +0.25 correlation between beer and soda sales indicates a slight positive linear association between beer and soda sales.

Using the CORREL Function

As an alternative to using the Correlation option of the Analysis Toolpak, you can use the CORREL function. For example, enter the formula

=CORREL(Bowl_

Price,F5:F190) in cell N6 and you can confirm that the correlation between price and bowl sales is -0.71.

Relationship Between Correlation and R^2

The correlation between two sets of data is simply $-\sqrt{R^2}$ for the trend line, where you choose the sign for the square root to be the same as the sign of the slope of the trend line. Thus the correlation between bowl price and bowl sales is $-\sqrt{.507} = -0.711$.

Correlation and Regression Toward the Mean

You have probably heard the phrase “regression toward the mean.” Essentially, this means that the predicted value of a dependent variable will be in some sense closer to its average value than the independent variable. More precisely, suppose you try to predict a dependent variable y from an independent variable x . If x is k standard deviations above average, then your prediction for y will be $r \times k$ standard deviations above average. (Here, r = correlation between x and y .) Because r is between -1 and $+1$, this means that y is fewer standard deviations away from the mean than x . This is the real definition of “regression toward the mean.” See Exercise 9 for an interesting application of the concept of regression toward the mean.

Summary

Here is a summary of what you have learned in this chapter:

- %4 The Excel Trendline can be used to find the line that best fits data.
- %4 The R^2 value is the fraction of variation in the dependent variable explained by variation in the independent variable.
- %4 Approximately 95 percent of the forecasts from a least-squares line are accurate within two standard errors of the regression.
- %4 Given two variables x and y , the correlation r (always between -1 and $+1$) between x and y is a measure of the strength of the linear association between x and y .

- %4 Correlation may be computed with the Analysis ToolPak or the CORREL function.
- %4 If x is k standard deviations above the mean, you can predict y to be rk standard deviations above the mean.

Exercises

- %4 The file `Delldata.xlsx` (available on the companion website) contains monthly returns for the Standard & Poor's stock index and for Dell stock. The *beta* of a stock is defined as the slope of the least-squares line used to predict the monthly return for a stock from the monthly return for the market. Use this file to perform the following exercises:

Estimate the beta of Dell.

Interpret the meaning of Dell's beta.

If you believe a recession is coming, would you rather invest in a high-beta or low-beta stock?

During a month in which the market goes up 5 percent, you are 95 percent sure that Dell's stock price will increase between which range of values?

- %4 The file `Housedata.xlsx` (available on the companion website) gives the square footage and sales prices for several houses in Bellevue, Washington. Use this file to answer the following questions:

You plan to build a 500-square-foot addition to your house. How much do you think your home value will increase as a result?

What percentage of the variation in home value is explained by the variation in the house size?

A 3,000-square-foot house is listed for \$500,000. Is this price out of line with typical real estate values in Bellevue? What might cause this discrepancy?

- %4 You know that 32 degrees Fahrenheit is equivalent to 0 degrees Celsius, and that 212 degrees Fahrenheit is equivalent to 100 degrees Celsius. Use the trend curve to determine the relationship between Fahrenheit and Celsius temperatures. When you create your initial chart, before clicking Finish, you must indicate (using Switch Rows and Columns from the Design Tab on Chart Tools) that data is in columns and not rows because with only two data points, Excel assumes different variables are in different rows.

176 Part III: Forecasting

- %4** The file `Electiondata.xlsx` (available on the companion website) contains, for several elections, the percentage of votes Republicans gained from voting machines (counted on election day) and the percentage Republicans gained from absentee ballots (counted after election day). Suppose that during an election, Republicans obtained 49 percent of the votes on election day and 62 percent of the absentee ballot votes. The Democratic candidate cried “Fraud.” What do you think?
- %4** The file `GNP.xls` (available on the companion website) contains quarterly GNP data for the United States in the years 1970–2012. Try to predict next quarter’s GNP from last quarter’s GNP. What is the R^2 ? Does this mean you are good at predicting next quarter’s GNP?
- %4** Find the trend line to predict soda sales from daily bowl sales.
- %4** The file `Parking.xlsx` contains the number of cars parked each day both in the outdoor lot and in the parking garage near the Indiana University Kelley School of Business. Find and interpret the correlation between the number of cars parked in the outdoor lot and in the parking garage.
- %4** The file `Printers.xlsx` contains daily sales volume (in dollars) of laser printers, printer cartridges, and school supplies. Find and interpret the correlations between these quantities.
- %4** NFL teams play 16 games during the regular season. Suppose the standard deviation of the number of games won by all teams is 2, and the correlation between the number of games a team wins in two consecutive seasons is 0.5. If a team goes 12 and 4 during a season, what is your best prediction for how many games they will win next season?

10

Using Multiple Regression to Forecast Sales

A common need in marketing analytics is forecasting the sales of a product. This chapter continues the discussion of *causal forecasting* as it pertains to this need. In causal forecasting, you try and predict a dependent variable (usually called Y) from one or more independent variables (usually referred to as $X_1, X_2, \dots, X_n$). In this chapter the dependent variable Y usually equals the sales of a product during a given time period.

Due to its simplicity, univariate regression (as discussed in Chapter 9, “Simple Linear Regression and Correlation”) may not explain all or even most of the variance in Y . Therefore, to gain better and more accurate insights about the often complex relationships between a variable of interest and its predictors, as well as to better forecast, one needs to move towards multiple regression in which more than one independent variable is used to forecast Y . Utilizing multiple regression may lead to improved forecasting accuracy along with a better understanding of the variables that actually cause Y .

For example, a multiple regression model can tell you how a price cut increases sales or how a reduction in advertising decreases sales. This chapter uses multiple regression in the following situations:

- %4 Setting sales quotas for computer sales in Europe
- %4 Predicting quarterly U.S. auto sales
- %4 Understanding how predicting sales from price and advertising requires knowledge of nonlinearities and interaction
- %4 Understanding how to test whether the assumptions needed for multiple regression are satisfied
- %4 How multicollinearity and/or autocorrelation can disturb a regression model

Introducing Multiple Linear Regression

In a **multiple linear regression model**, you can try to predict a dependent variable Y from independent variables $X_1 X_2 \dots X_n$. The assumed model is as follows:

$$(1) Y = B_0 + B_1 X_1 + B_2 X_2 + \dots B_n X_n + \text{error term}$$

In Equation 1:

%4 B_0 is called the *intercept* or *constant term*

%4 B_i is called the *regression coefficient* for the independent variable X_i

The *error term* is a random variable that captures the fact that regression models typically do not fit the data perfectly; rather they approximate the relationships in the data. A positive value of the error term occurs if the actual value of the dependent variable exceeds your predicted value ($B_0 + B_1 X_1 + B_2 X_2 + \dots B_n X_n$). A negative value of the error term occurs when the actual value of the dependent variable is less than the predicted value.

The error term is required to satisfy the following assumptions:

%4 The error term is normally distributed.

%4 The variability or spread of the error term is assumed not to depend on the value of the dependent variable.

%4 For time series data successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.

In the “Testing Validity of Multiple Regression Assumptions,” section of this chapter you will learn how to determine if the assumptions of regression analysis are satisfied, and what to do if the assumptions are not satisfied.

To best illustrate how to use multiple regression, the remainder of the chapter presents examples of its use based on a fictional computer sales company, HAL Computer. HAL sets sales quotas for all salespeople based on their territory. To set fair quotas, HAL needs a way to accurately forecast computer sales in each person’s territory. From the 2011 *Pocket World in Figures* by *The Economist*, you can obtain the following data from 2007 (as shown in Figure 10-1 and file Europe.xlsx) for European countries:

%4 Population (in millions)

%4 Computer sales (in millions of U.S. dollars)

- %4 Sales per capita (in U.S. dollars)
- %4 GNP per head
- %4 Average Unemployment Rate 2002–2007
- %4 Percentage of GNP spent on education

	F	G	H	I	J	K	L
3				Source Economist Pocket World in Figures 2011			
4	Country	Pop (millions)	Computer Sales	Sales/Capita	GNP per head	Unemployment rate	%age spend on education
5	Austria	8.4	941.2	\$112.05	\$49,600	4.2	5.8
6	Belgium	10.5	1681.9	\$160.18	\$47,090	8.1	5.9
7	Bulgaria	7.6	154	\$20.26	\$6,550	13.5	3.5
8	Czech Rep.	10.2	1028.7	\$100.85	\$20,670	6.6	4.4
9	Denmark	5.5	935.4	\$170.07	\$62,120	5.2	8.4
10	Finland	5.3	1971	\$371.89	\$51,320	9.9	6.3
11	France	61.9	5928.9	\$95.78	\$44,510	10	5.7
12	Germany	82.5	6824.3	\$82.72	\$44,450	9.1	4.6
13	Greece	11.2	813	\$72.59	\$31,670	9.9	3.9
14	Hungary	10	449	\$44.90	\$15,410	7.3	5.1
15	Ireland	4.4	576.9	\$131.11	\$60,460	6.3	4.3
16	Italy	58.9	3858.2	\$65.50	\$38,490	9.3	5
17	Netherlands	16.5	2168.5	\$131.42	\$52,960	4.4	5
18	Poland	38	2847	\$74.92	\$13,850	14.4	5.6
19	Portugal	10.7	728.6	\$68.09	\$22,920	6.3	5.9
20	Romania	21.3	687.2	\$32.26	\$9,300	7	3.3
21	Spain	44.8	4745.8	\$105.93	\$35,220	14.2	4.4
22	Switzerland	7.5	1130.4	\$150.72	\$64,430	3.6	5.6
23	Sweden	9.2	2113.4	\$229.72	\$51,950	6.3	7.6
24	Turkey	75.8	2879	\$37.98	\$9,940	8.6	3.7

Figure 10-1: HAL computer data

This data is *cross sectional data* because the same dependent variable is measured in different locations at the same point in time. In *time series data*, the same dependent variable is measured at different times.

In order to apply the multiple linear regression model to the example, Y = Per Capital Computer spending, $n = 3$, X_1 = Per Capita GNP, X_2 = Unemployment Rate, and X_3 = Percentage of GNP spent on education.

Running a Regression with the Data Analysis Add-In

You **can use** the Excel Data Analysis Add-In to determine the best-fitting multiple linear regression equation to a given set of data. See Chapter 9 for a refresher on installation instructions for the Data Analysis Add-In.

%4 Part III: Forecasting

To run a regression, select Data Analysis in the Analysis Group on the Data tab, and then select Regression. When the Regression dialog box appears, fill it in, as shown in Figure 10-2.

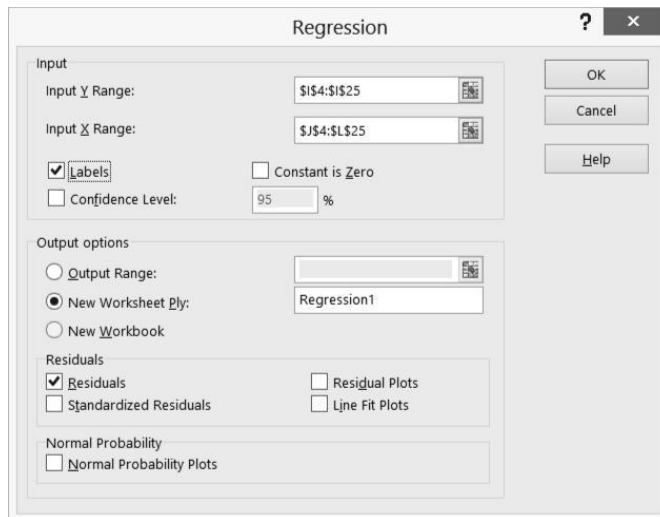
The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$I\$4:\$I\$25' and 'Input X Range' set to '\$J\$4:\$L\$25'. The 'Labels' checkbox is checked, and 'Confidence Level' is set to '95 %'. The 'Constant is Zero' checkbox is unchecked. The 'Output options' section has 'New Worksheet Ply:' selected, with the worksheet name 'Regression1' entered. The 'Residuals' section has 'Residuals' checked, while 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' are unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. On the right side of the dialog box are 'OK', 'Cancel', and 'Help' buttons.

Figure 10-2: Regression dialog box

- %4 The Y Range (I4:I25) includes the data you want to predict (computer per capita sales), including the column label.
- %4 The X Range (J4:L25) includes those values of the independent variables for each country, including the column label.
- %4 Check the Labels box because your X range and Y range include labels. If you do not include labels in the X and Y range, then Excel will use generic labels like Y, X₁, X₂, ..., X_n which are hard to interpret.
- %4 The worksheet name Regression1 is the location where the output is placed.
- %4 By checking the Residuals box, you can ensure Excel will generate the error (for each observation error = actual value of Y – predicted value for Y).

After selecting OK, Excel generates the output shown in Figures 10-3 and 10-4. For Figure 10-4, the highlighted text indicates data that is thrown out later in the chapter.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	<i>Regression Statistics</i>								
4	Multiple R	0.731106465							
5	R Square	0.534516664							
6	Adjusted R Square	0.452372545							
7	Standard Error	58.42625704							
8	Observations	21							
9									
10	ANOVA								
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
12	Regression	3	66638.03186	22212.68	6.507059	0.003940222			
13	Residual	17	58031.66769	3413.628					
14	Total	20	124669.6996						
15									
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
17	Intercept	-114.8351503	78.28996449	-1.46679	0.160688	-280.012537	50.34224	-280.0125369	50.3422364
18	GNP per head	0.002297712	0.00095193	2.413741	0.027355	0.000289316	0.004306	0.000289316	0.004306108
19	Unemployment rate	4.219524573	4.840005896	0.871802	0.395463	-5.99199526	14.43104	-5.991995264	14.43104441
20	%age spend on education	21.4226983	12.73611957	1.682043	0.110837	-5.44816518	48.29356	-5.448165177	48.29356178

Figure 10-3: First multiple regression output

	A	B	C	D
24	RESIDUAL OUTPUT			
25				
26	<i>Observation</i>	<i>Predicted Sales/Capita</i>	<i>Residuals</i>	
27	1	141.105011	-29.0574	
28	2	153.9361699	6.244782	
29	3	32.15788817	-11.8947	
30	4	54.76728846	46.08565	
31	5	229.7909036	-59.7182	
32	6	179.8197146	192.0671	Finland
33	7	151.7406304	-55.9587	
34	8	124.2402274	-41.5214	
35	9	83.25520078	-10.6659	
36	10	60.63088012	-15.7309	
37	11	142.7851159	-11.6715	
38	12	119.958849	-54.4546	
39	13	132.5310691	-1.10683	
40	14	97.71642325	-22.7954	
41	15	90.80533021	-22.7119	
42	16	6.765146383	25.49776	
43	17	120.2673827	-14.3343	
44	18	168.3638234	-17.6438	
45	19	193.9264924	35.7909	
46	20	23.55600061	14.42553	
47	21	112.9313707	49.15388	

Figure 10-4: Residuals from first regression

Interpreting the Regression Out put

After you run a regression, you next must interpret the output. To do this you must analyze a variety of elements listed in the output. Each element of the output affects the output in a unique manner. The following sections explain how to interpret the important elements of the regression output.

Coefficients

The Coefficients column of the output (cells B17:B20) gives the best fitting estimate of the multiple regression equation. Excel returns the following equation:

$$\%4 \text{ Predicted Computer Sales / Capita} = -114.84 + .002298 * (\text{Per Capita GNP}) \\ + 4.22 * (\text{Unemployment Rate}) + 21.42(\text{Percentage Spent on Education})$$

Excel found this equation by considering all values of B_0 , B_1 , B_2 , and B_3 and choosing the values that minimize the sum over all observations of $(\text{Actual Dependent Variable} - \text{Predicted Value})^2$. The coefficients are called the *least squares estimates* of B_0 , B_1 , ..., B_n . You square the errors so positive and negative values do not cancel. Note that if the equation perfectly fits each observation, then the sum of squared errors is equal to 0.

F Test for Hypothesis of No Linear Regression

Just because you throw an independent variable into a regression does not mean it is a helpful predictor. If you used the number of games each country's national soccer team won during 2007 as an independent variable, it would probably be irrelevant and have no effect on computer sales. The ANOVA section of the regression output (shown in Figure 10-3) in cells A10:F14 enables you to test the following hypotheses:

%4 **Null Hypothesis:** The *Hypothesis of No Linear Regression*: Together all the independent variables are not useful (or significant) in predicting Y

%4 **Alternative Hypothesis:** Together all the independent variables are useful (or significant).

To decide between these hypotheses, you must examine the *Significance F Value* in cell F12. The Significance F value of .004 tells you that the data indicates that there are only 4 chances in 1000 that your independent variables are not useful in predicting Y , so you would reject the null hypothesis. Most statisticians agree that a Significance F (often called *p value*) of .05 or less should cause rejection of the Null Hypothesis.

Accuracy and Goodness of Fit of Regression Forecasts

After you conclude that the independent variables together are significant, a natural question is, how well does your regression equation fit the data? The R^2 value in B5 and *Standard Error* in B7 (see Figure 10-3) answer this question.

%4 The R^2 value of .53 indicates that 53 percent of the variation in Y is explained by Equation 1. Therefore, 47 percent of the variation in Y is unexplained by the multiple linear regression model.

%4 The Standard Error of 58.43 indicates that approximately 68 percent of the predictions for Y made from Equation 2 are accurate within one standard error (\$58.43) and 95 percent of your predictions for Y made from Equation 2 are accurate within two standard errors (\$116.86.)

Determining the Significant Independent Variables

Because you concluded that together your independent variables are useful in predicting Y , you now must determine which independent variables are useful. To do this look at the p values in E17:E20. A p -value of .05 or less for an independent variable indicates that the independent variable is (after including the effects of all other independent variables in the equation) a significant predictor for Y . It appears that only GNP per head (p -value .027) is a significant predictor. At this point you want to see if there are any *outliers* or unusual data points. Outliers in regression are data points where the absolute value of the error (actual value of y – predicted value of y) exceeds two standard errors. Outliers can have a drastic effect on regression coefficients, and the analyst must decide whether to rerun the regression without the outliers.

The Residual Output and Outliers

For each data point or observation, the Residual portion of the regression output, as shown in Figure 10-4, gives you two pieces of information.

%4 The Predicted Value of Y from Equation 2. For example, Austria predicted per capita expenditures are given by the following:

$$(\$116.86) + (0.00229) * (49,600) + (4.22) * (4.2) + 21.52 = \$141.10$$

%4 The Residuals section of the output gives for each observation the error Actual value of Y – Predicted Value of Y . For Austria you find the residual is $\$112.05 - \$141.10 = \$-29.05$. The regression equation found by least squares

%4 Part III: Forecasting

has the intuitively pleasing property that the sum of the residuals equals 0. This implies that overestimates and underestimates of Y cancel each other out.

Dealing with Insignificant Independent Variables

In the last section you learned that GNP per head was the only significant independent variable and the other two independent variables were insignificant. When an independent variable is insignificant (has a p-value greater than .05) you can usually drop it and run the regression again. Before doing this though, you must decide what to do with your outlier(s). Because the standard error of the regression is 58.4, any error exceeding 116.8 in absolute value is an outlier. Refer to Figure 10-4 and you can see that Finland (which is highlighted) is a huge outlier. Finland's spending on computers is more than three standard errors greater than expected. When you delete Finland as an outlier, and then rerun the analysis, the result is in the worksheet Regression2 of file Europe.xlsx, as shown in Figure 10-5.

Checking the residuals you find that Switzerland is an outlier. (You under predict expenditures by slightly more than two standard errors.) Because Switzerland is not an outrageous outlier, you can choose to leave it in the data set in this instance. Unemployment Rate is insignificant (p-value of .84 > .05) so you can delete it from the model and run the regression again. The resulting regression is in worksheet Regression 3, of file Europe.xlsx as shown in Figure 10-6.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.860805637					
5	R Square	0.740986344					
6	Adjusted R Square	0.692421283					
7	Standard Error	29.9835813					
8	Observations	20					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	41150.44477	13716.81	15.2576	5.93265E-05	
13	Residual	16	14384.24236	899.0151			
14	Total	19	55534.68713				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-32.20876114	41.89082135	-0.76887	0.453169	-121.0133353	56.595813
18	GNP per head	0.001678416	0.000496537	3.380244	0.003816	0.000625805	0.002731
19	Unemployment rate	-0.527867146	2.575579641	-0.20495	0.840195	-5.987852075	4.9321178
20	%age spend on education	15.22764461	6.596202921	2.308547	0.034658	1.244319079	29.21097

Figure 10-5: Regression results: Finland outlier removed

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3		<i>Regression Statistics</i>					
4	Multiple R	0.860410573					
5	R Square	0.740306355					
6	Adjusted R Square	0.709754161					
7	Standard Error	29.12650434					
8	Observations	20					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	41112.68179	20556.34	24.23087	1.0542E-05	
13	Residual	17	14422.00534	848.3533			
14	Total	19	55534.68713				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	-38.48026121	27.79129164	-1.38462	0.184076	-97.1147612	20.154239
18	GNP per head	0.001723168	0.000433202	3.977751	0.000973	0.000809193	0.0026371
19	%age spend on education	15.30973984	6.395825812	2.393708	0.028487	1.815726905	28.803753

Figure 10-6: Regression output: unemployment rate removed

Both independent variables are significant, so use the following equation to predict Per Capita Computer Spending:

$$\%4 \quad -38.48 + 0.001723 * (\text{GNP Per Capita}) + 15.30974 * (\text{Percentage GNP Spent on Education})$$

Because $R^2 = 0.74$, the equation explains 74 percent of the variation in Computer Spending. Because the Standard error is 29.13, you can expect 95 percent of your forecasts to be accurate within \$58.26. From the Residuals portion of the output, you can see that Switzerland (error of \$62.32) is the only outlier.

Interpreting Regression Coefficients

The regression coefficient of a variable estimates the effect (after adjusting for all other independent variables used to estimate the regression equation) of a unit increase in the independent variable. Therefore Equation 3 may be interpreted as follows:

%4 After adjusting for a fraction of GNP spent on education, a \$1,000 increase in Per Capita GNP yields a \$1.72 increase in Per Capital Computer spending.

%4 After adjusting for Per Capita GNP, a 1 percent increase in the fraction of GNP spent on education yields a \$15.31 increase in Per Capita Computer spending.



Setting Sales Quotas

Often part of a salesperson's compensation is a commission based on whether a salesperson's sales quota is met. For commission payments to be fair, the company needs to ensure that a salesperson with a "good" territory has a higher quota than a salesperson with a "bad" territory. You'll now see how to use the multiple regression model to set fair sales quotas. Using the multiple regression, a reasonable annual sales quota for a territory equals the population * company market share * regression prediction for per capita spending.

Assume that a province in France has a per capita GNP of \$50,000 and spends 10 percent of its GNP on education. If your company has a 30 percent market share, then a reasonable per capita annual quota for your sales force would be the following:

$$0.30(-38.48 + 0.001723 * (50,000) + 15.30974 * (10)) = \$60.23$$

Therefore, a reasonable sales quota would be \$60.23 per capita.

Beware of Blind Extrapolation

While you can use regressions to portray a lot of valuable information, you must be wary of using them to predict values of the independent variables that differ greatly from the values of the independent variables that fit the regression equation. For example, the Ivory Coast has a Per Capita GNP of \$1,140, which is far less than any country in your European data set, so you could not expect Equation 3 to give a reasonable prediction for Per Capita Computer spending in the Ivory Coast.

Using Qualitative Independent Variables in Regression

In the **previous example** of multiple regression, you forecasted Per Capita Computer sales using Per Capita GNP and Fraction of GNP spent on education. Independent variables can also be quantified with an exact numerical value and are referred to as *quantitative independent variables*. In many situations, however, independent variables can't be easily quantified. This section looks at ways to incorporate a qualitative factor, such as seasonality, into a multiple regression analysis.

Suppose you want to predict quarterly U.S. auto sales to determine whether the quarter of the year impacts auto sales. Use the data in the file *Autos.xlsx*, as shown in Figure 10-7. Sales are listed in thousands of cars, and GNP is in billions of dollars.

You might be tempted to define an independent variable that equals 1 during the first quarter, 2 during the second quarter, and so on. Unfortunately, this approach

would force the fourth quarter to have four times the effect of the first quarter, which might not be true. The quarter of the year is a qualitative independent variable. To model a qualitative independent variable, create an independent variable (called a *dummy variable*) for all but one of the qualitative variable's possible values. (It is arbitrary which value you leave out. This example omits Quarter 4.) The dummy variables tell you which value of the qualitative variable occurs. Thus, you have a dummy variable for Quarter 1, Quarter 2, and Quarter 3 with the following properties:

- %4 Quarter 1 dummy variable equals 1 if the quarter is Quarter 1 and 0 if otherwise.
- %4 Quarter 2 dummy variable equals 1 if the quarter is Quarter 2 and 0 if otherwise.
- %4 Quarter 3 dummy variable equals 1 if the quarter is Quarter 3 and 0 if otherwise.

	A	B	C	D	E	F
9	Historical data					
10	Year	Quarter	Sales	GNP	Unemp	Int
11	79	1	2541	2541	5.9	9.4
12	79	2	2910	2640	5.7	9.4
13	79	3	2562	2595	5.9	9.7
14	79	4	2385	2701	6	12
15	80	1	2520	2785	6.2	13
16	80	2	2142	2509	7.3	9.6
17	80	3	2130	2570	7.7	9.2
18	80	4	2190	2667	7.4	14
19	81	1	2370	2878	7.4	14
20	81	2	2208	2835	7.4	15
21	81	3	2196	2897	7.4	15
22	81	4	1758	2744	8.3	12
23	82	1	1944	2582	8.8	13
24	82	2	2094	2613	9.4	12
25	82	3	1911	2529	10	9.3
26	82	4	2031	2544	10.7	7.9
27	83	1	2046	2633	10.4	7.8
28	83	2	2502	2878	10.1	8.4
29	83	3	2238	3051	9.4	9.1
30	83	4	2394	3274	8.5	8.8
31	84	1	2586	3594	7.9	9.2
32	84	2	2898	3774	7.5	9.8
33	84	3	2448	3861	7.5	10
34	84	4	2460	3919	7.2	8.8
35	85	1	2646	4040	7.4	8.2
36	85	2	2988	4133	7.3	7.5
37	85	3	2967	4303	7.1	7.1
38	85	4	2439	4393	7	7.2
39	86	1	2598	4560	7.1	8.9
40	86	2	3045	4587	7.1	7.7
41	86	3	3213	4716	6.9	7.4
42	86	4	2685	4796	6.8	7.4

Figure 10-7: Auto sales data

Part III: Forecasting

A Quarter 4 observation can be identified because the dummy variables for Quarter 1 through Quarter 3 equal 0. It turns out you don't need a dummy variable for Quarter 4. In fact, if you include a dummy variable for Quarter 4 as an independent variable in your regression, Microsoft Office Excel returns an error message. The reason you get an error is because if an exact linear relationship exists between any set of independent variables, Excel must perform the mathematical equivalent of dividing by 0 (an impossibility) when running a multiple regression. In this situation, if you include a Quarter 4 dummy variable, every data point satisfies the following exact linear relationship:

$$\begin{aligned} &(\text{Quarter 1 Dummy}) + (\text{Quarter 2 Dummy}) + (\text{Quarter 3 Dummy}) \\ &+ (\text{Quarter 4 Dummy}) = 1 \end{aligned}$$

NOTE An exact linear relationship occurs if there exists constants $c_0, c_1, \dots, c_N$, such that for each data point $c_0 + c_1x_1 + c_2x_2 + \dots + c_Nx_N = 0$. Here $x_1, \dots, x_N$ are the values of the independent variables.

You can interpret the “omitted” dummy variable as a “baseline” scenario; this is reflected in the “regular” intercept. Therefore, you can think of dummies as changes in the intercept.

To create your dummy variable for Quarter 1, copy the formula `IF(B12=1,1,0)` from G12 to G13:G42. This formula places a 1 in column G whenever a quarter is the first quarter, and places a 0 in column G whenever the quarter is not the first quarter. In a similar fashion, you can create dummy variables for Quarter 2 (in H12:H42) and Quarter 3 (in I12:I42). Figure 10-8 shows the results of the formulas.

In addition to seasonality, you'd like to use macroeconomic variables such as gross national product (GNP, in billions of 1986 dollars), interest rates, and unemployment rates to predict car sales. Suppose, for example, that you want to estimate sales for the second quarter of 1979. Because values for GNP, interest rate, and unemployment rate aren't known at the beginning of the second quarter 1979, you can't use the second quarter 1979 GNP, interest rate, and unemployment rate to predict Quarter 2 1979 auto sales. Instead, you use the values for the GNP, interest rate, and unemployment rate lagged one quarter to forecast auto sales. By copying the formula `=D11` from J12 to J12:L42, you can create the lagged value for GNP, the first of your macroeconomic-independent variables. For example, the range J12:L12 contains GNP, unemployment rate, and interest rate for the first quarter of 1979.

You can now run your multiple regression by clicking Data Analysis on the Data tab and then selecting Regression in the Data Analysis dialog box. Use C11:C42 as the Input Y Range and G11:L42 as the Input X Range; check the Labels box (row 11 contains labels), and also check the Residuals box. After clicking OK, you can

obtain the output, which you can see in the Regression worksheet of the file Autos.xlsx and in Figure 10-9.

	G	H	I	J	K	L
9						
10	Q1	Q2	Q3	LagGNP	LagUnemp	LagInt
11	Q1	Q2	Q3	LagGNP	LagUnemp	LagInt
12	0	1	0	2541	5.9	9.4
13	0	0	1	2640	5.7	9.4
14	0	0	0	2595	5.9	9.7
15	1	0	0	2701	6	11.9
16	0	1	0	2785	6.2	13.4
17	0	0	1	2509	7.3	9.6
18	0	0	0	2570	7.7	9.2
19	1	0	0	2667	7.4	13.6
20	0	1	0	2878	7.4	14.4
21	0	0	1	2835	7.4	15.3
22	0	0	0	2897	7.4	15.1
23	1	0	0	2744	8.3	11.8
24	0	1	0	2582	8.8	12.8
25	0	0	1	2613	9.4	12.4
26	0	0	0	2529	10	9.3
27	1	0	0	2544	10.7	7.9
28	0	1	0	2633	10.4	7.8
29	0	0	1	2878	10.1	8.4
30	0	0	0	3051	9.4	9.1
31	1	0	0	3274	8.5	8.8
32	0	1	0	3594	7.9	9.2
33	0	0	1	3774	7.5	9.8
34	0	0	0	3861	7.5	10.3
35	1	0	0	3919	7.2	8.8
36	0	1	0	4040	7.4	8.2
37	0	0	1	4133	7.3	7.5
38	0	0	0	4303	7.1	7.1
39	1	0	0	4393	7	7.2
40	0	1	0	4560	7.1	8.9
41	0	0	1	4587	7.1	7.7
42	0	0	0	4716	6.9	7.4

Figure 10-8: Dummy and lagged variables

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.884139126					
5	R Square	0.781701994					
6	Adjusted R Square	0.727127492					
7	Standard Error	190.5240756					
8	Observations	31					
9	ANOVA						
10		df	SS	MS	F	Significance F	
11	Regression	6	3119625.193	519937.5322	14.32357552	6.79746E-07	
12	Residual	24	871186.1616	36299.4234			
13	Total	30	3990811.355				
14							
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	3154.700285	462.6530922	6.818716525	4.7214E-07	2199.83143	4109.56914
18	Q1	156.833091	98.87110703	1.586237838	0.125774521	-47.22680256	360.8929846
19	Q2	379.7835116	96.08921514	3.95240518	0.000594196	181.4651595	578.1018637
20	Q3	203.035501	95.40891864	2.128055783	0.043800625	6.12121161	399.9497905
21	LagGNP	0.174156906	0.05842	2.981117865	0.006490201	0.053583977	0.294729835
22	LagUnemp	-93.83233214	28.32328716	-3.312904029	0.002918487	-152.2887117	-35.37595254
23	LagInt	-73.9167147	17.78851573	-4.155305356	0.000355622	-110.6303992	-37.20303022

Figure 10-9: Summary regression output for auto example

%4 Part III: Forecasting

In Figure 10-9, you can see that Equation 1 is used to predict quarterly auto sales as follows:

Predicted quarterly sales = $3154.7 + 156.833Q1 + 379.784Q2 + 203.036Q3 + 174(\text{LAGGNP in billions}) - 93.83(\text{LAGUNEMP}) - 73.91(\text{LAGINT})$

Also in Figure 10-9, you see that each independent variable except Q1 has a p-value less than or equal to 0.05. The previous discussion would indicate that you should drop the Q1 variable and rerun the regression. Because Q2 and Q3 are significant, you know there is significant seasonality, so leave Q1 as an independent variable because this treats the seasonality indicator variables as a “package deal.” You can therefore conclude that all independent variables have a significant effect on quarterly auto sales. You interpret all coefficients in your regression equation *ceteris paribus* (which means that each coefficient gives the effect of the independent variable after adjusting for the effects of all other variables in the regression).

Each regression coefficient is interpreted as follows:

- %4 A \$1 billion increase in last quarter’s GNP increases quarterly car sales by 174.
- %4 An increase of 1 percent in last quarter’s unemployment rate decreases quarterly car sales by 93,832.
- %4 An increase of 1 percent in last quarter’s interest rate decreases quarterly car sales by 73,917.

To interpret the coefficients of the dummy variables, you must realize that they tell you the effect of seasonality relative to the value left out of the qualitative variables. Therefore

- %4 In Quarter 1, car sales exceed Quarter 4 car sales by 156,833.
- %4 In Quarter 2, car sales exceed Quarter 4 car sales by 379,784.
- %4 In Quarter 3, car sales exceed Quarter 4 car sales by 203,036.

Car sales are highest during the second quarter (April through June; tax refunds and summer are coming) and lowest during the third quarter. (October through December; why buy a new car when winter salting will ruin it?)

You should note that each regression coefficient is computed after adjusting for all other independent variables in the equation (this is often referred to as *ceteris paribus*, or all other things held equal).

From the Summary output shown in Figure 10-9, you can learn the following:

- %4 The variation in your independent variables (macroeconomic factors and seasonality) explains 78 percent of the variation in your dependent variable (quarterly car sales).

%4 The standard error of your regression is 190,524 cars. You can expect approximately 68 percent of your forecasts to be accurate within 190,524 cars and about 95 percent of your forecasts to be accurate within 381,048 cars ($2 * 190,524$).

%4 There are 31 observations used to fit the regression.

The only quantity of interest in the ANOVA portion of Figure 10-9 is the significance (0.00000068). This measure implies that there are only 6.8 chances in 10,000,000, that when taken together, all your independent variables are useless in forecasting car sales. Thus, you can be quite sure that your independent variables are useful in predicting quarterly auto sales.

Figure 10-10 shows for each observation the predicted sales and residual. For example, for the second quarter of 1979 (observation 1), predicted sales from Equation 1 are 2728.6 thousand, and your residual is 181,400 cars ($2910 - 2728.6$). Note that no residual exceeds 381,000 in absolute value, so you have no outliers.

	A	B	C
27	RESIDUAL OUTPUT		
28			
29	Observation	Predicted Sales	Residuals
30	1	2728.588616	181.4113836
31	2	2587.848606	-25.84860587
32	3	2336.034563	48.96543676
33	4	2339.328281	180.6717193
34	5	2447.266343	-305.2663429
35	6	2400.118977	-270.1189769
36	7	2199.7408	-9.740800106
37	8	2076.383266	293.6167341
38	9	2276.947422	-68.94742189
39	10	2026.185621	169.8143789
40	11	1848.731191	-90.73119119
41	12	2138.394335	-194.3943352
42	13	2212.298456	-118.2984563
43	14	2014.216596	-103.2165964
44	15	1969.394332	61.60566841
45	16	2166.640544	-120.6405443
46	17	2440.632301	61.3676994
47	18	2290.352403	-52.35240272
48	19	2131.386979	262.6130214
49	20	2433.681173	152.3188271
50	21	2739.094517	158.9054833
51	22	2586.877653	-138.8776532
52	23	2362.035446	97.96455439
53	24	2667.994409	-21.99440885
54	25	2937.601377	50.39862257
55	26	2838.174893	128.8251074
56	27	2713.079218	-274.0792178
57	28	2887.577992	-289.5779921
58	29	3004.570968	40.42903226
59	30	2921.225251	291.7747488
60	31	2781.597472	-96.59747188

Figure 10-10: Residual output for Auto example

Part III: Forecasting

Modeling Interactions and Nonlinearities

Equation 1 assumes that each independent variable affects Y in a **linear fashion**. This means, for example, that a unit increase in X^1 will increase Y by B_1 for any values of

$X_2, \dots, X_n$. In many marketing situations this assumption of linearity is unrealistic. In this section, you learn how to model situations in which an independent variable can interact with or influence Y in a nonlinear fashion.

Nonlinear Relationship

An independent variable can often influence a dependent variable through a nonlinear relationship. For example, if you try to predict product sales using an equation such as the following, price influences sales linearly.

$$\text{Sales} = 500 - 10 * \text{Price}$$

This equation indicates that a unit increase in price can (at any price level) reduce sales by 10 units. If the relationship between sales and price were governed by an equation such as the following, price and sales would be related nonlinearly.

$$\text{Sales} = 500 + 4 * \text{Price} - .40 * \text{Price}^2$$

As shown in Figure 10-11, larger increases in price result in larger decreases in demand. In short, if the change in the dependent variable caused by a unit change in the independent variable is not constant, there is a nonlinear relationship between the independent and dependent variables.

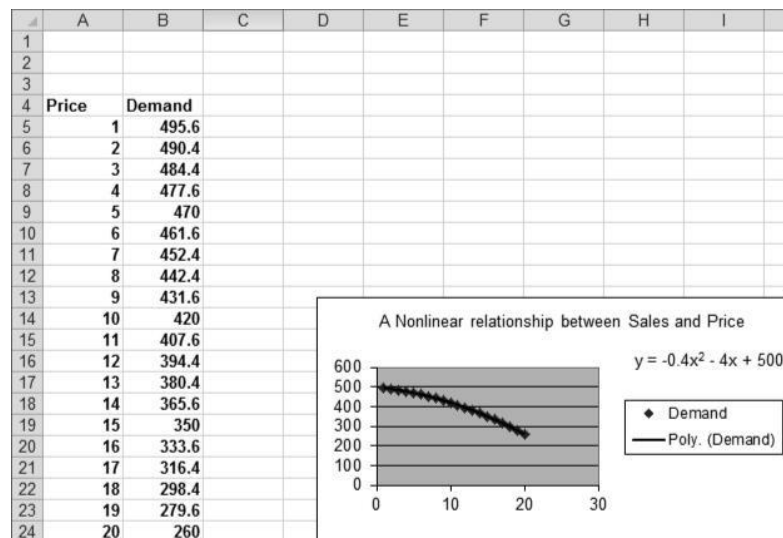


Figure 10-11: Nonlinear relationship between Sales and Price

Interaction

If the effect of one independent variable on a dependent variable depends on the value of another independent variable, you can say that the two independent variables exhibit *interaction*. For example, suppose you try to predict sales using the price and the amount spent on advertising. If the effect to change the level of advertising dollars is large when the price is low and small when the price is high, price and advertising exhibit interaction. If the effect to change the level of advertising dollars is the same for any price level, sales and price do not exhibit any interaction. You will encounter interactions again in Chapter 41, “Analysis of Variance: Two-way ANOVA.”

Testing for Nonlinearities and Interactions

To see whether an independent variable has a nonlinear effect on a dependent variable, simply add an independent variable to the regression that equals the square of the independent variable. If the squared term has a low p-value (less than 0.05), you have evidence of a nonlinear relationship.

To check whether two independent variables exhibit interaction, simply add a term to the regression that equals the product of the independent variables. If the term has a low p-value (less than 0.05), you have evidence of interaction. The file *Priceandads.xlsx* illustrates this procedure. In worksheet *data* from this file (see Figure 10-12), you have the weekly unit sales of a product, weekly price, and weekly advertising expenditures (in thousands of dollars).

With this example, you’ll want to predict weekly sales from the price and advertising. To determine whether the relationship is nonlinear or exhibits any interactions, perform the following steps:

- %4 Add in Column H *Advertising*Price*, in Column I *Price²*, and in Column J *Ad²*
- %4 Next, run a regression with *Y* Range E4:E169 and *X* Range F4:J169. You can obtain the regression output, as shown in the worksheet *nonlinear* and Figure 10-13.
- %4 All independent variables except for *Price²* have significant p-values (less than .05). Therefore, drop *Price²* as an independent variable and rerun the regression. The result is in Figure 10-14 and the worksheet *final*

%4 Part III: Forecasting

	E	F	G	H	I	J
1						
2						
3						
4	Sales	Price	Ad	A*P	Price^2	Ad^2
5	22845	8	1	8	64	1
6	20417	9	8	72	81	64
7	23761	5	3	15	25	9
8	22674	4	12	48	16	144
9	22782	7	5	35	49	25
10	23807	5	3	15	25	9
11	18924	10	9	90	100	81
12	21855	9	5	45	81	25
13	21749	10	4	40	100	16
14	22683	4	12	48	16	144
15	20968	6	11	66	36	121
16	22202	10	2	20	100	4
17	23241	6	5	30	36	25
18	19004	10	9	90	100	81
19	23978	4	1	4	16	1
20	20497	8	9	72	64	81
21	22322	9	3	27	81	9
22	22628	8	4	32	64	16
23	21051	10	6	60	100	36
24	24515	3	3	9	9	9
25	22126	10	2	20	100	4
26	22141	5	11	55	25	121
27	21151	9	7	63	81	49
28	22558	5	10	50	25	100

Figure 10-12: Nonlinearity and interaction data

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.996924531							
5	R Square	0.99385852							
6	Adjusted R Square	0.993665392	Price^2 has high p value so delete it and rerun analysis						
7	Standard Error	135.2764087							
8	Observations	165							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	5	470861057.7	94172212	5146.105	7.9582E-174			
13	Residual	159	2909653.375	18299.71					
14	Total	164	473770711.1						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	24005.74767	111.4951345	215.3076	5.5E-198	23785.54521	24225.95	23785.55	24225.95
18	Price	-135.6678621	32.18950019	-4.21466	4.18E-05	-199.242002	-72.0937	-199.242	-72.0937
19	Ad	660.0035108	16.15110952	40.86428	3.07E-86	628.1051313	691.9019	628.1051	691.9019
20	A*P	-74.12725368	1.425595543	-51.9974	1E-101	-76.94279942	-71.3117	-76.9428	-71.3117
21	Price^2	-0.178202781	2.349205511	-0.07586	0.939629	-4.817874684	4.461469	-4.81787	4.461469
22	Ad^2	-37.37381917	1.019418942	-36.6619	1.84E-79	-39.38716769	-35.3605	-39.3872	-35.3605

Figure 10-13: First regression output for Nonlinearity and Interaction example

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.996924419							
5	R Square	0.993858298							
6	Adjusted R Squ	0.993704755							
7	Standard Error	134.8554475							
8	Observations	165							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	4	470860952.4	117715238.1	6472.852287	9.3099E-176			
13	Residual	160	2909758.675	18185.99172					
14	Total	164	473770711.1						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	24012.24758	71.11479957	337.6547179	3.3475E-230	23871.80286	24152.69231	23871.80286	24152.69231
18	Price	-137.997013	9.633696108	-14.32441001	1.78044E-30	-157.0226141	-118.9714118	-157.0226141	-118.9714118
19	Ad	660.0418883	16.09294845	41.01435424	8.35145E-87	628.2598999	691.8238767	628.2598999	691.8238767
20	A*P	-74.12897476	1.420979292	-52.16752641	2.3559E-102	-76.93526893	-71.32268059	-76.93526893	-71.32268059
21	Ad^2	-37.37288222	1.016172056	-36.77810466	5.94521E-80	-39.37972197	-35.36604248	-39.37972197	-35.36604248

Figure 10-14: Final regression output for Nonlinearity and Interaction example

The Significance F Value is small, so the regression model has significant predictive values. All independent variables have extremely small p-values, so you can predict the weekly unit sales with the equation

$$\text{Predicted Unit Sales} = 24,012 - 138 * \text{Price} + 660.04 * \text{Ad} - 74.13 * \text{Ad} * \text{P} \\ \%4 \quad 37.33\text{Ad}^2$$

The -37.33Ad^2 term implies that each additional \$1,000 in advertising can generate fewer sales (diminishing returns). The $-74.13 * \text{Ad} * \text{P}$ term implies that at higher prices additional advertising has a smaller effect on sales.

The R^2 value of 99.4 percent implies your model explains 99.4 percent of the variation in weekly sales. The Standard Error of 134.86 implies that roughly 95 percent of your forecasts should be accurate within 269.71. Interactions and non-linear effects are likely to cause multicollinearity, which is covered in the section “Multicollinearity” later in this chapter.

Testing Validity of Regression Assumptions

Recall earlier in the chapter you learned the regression assumptions that should be satisfied by the error term in a multiple linear regression. For ease of presentation, these assumptions are repeated here:

- %4 The error term is normally distributed.
- %4 The variability or spread of the error term is assumed not to depend on the value of the dependent variable.

%4 Part III: Forecasting

%4 For time series data, successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.

This section further discusses how to determine if these assumptions are satisfied, the consequences of violating the assumptions, and how to resolve violation of these assumptions.

Normally Distributed Error Term

You can infer the nature of an unknown error term through examination of the residuals. If the residuals come from a normal random variable, the normal random variable should have a symmetric density. Then the skewness (as measured by Excel SKEW function described in Chapter 2) should be near 0.

Kurtosis, which may sound like a disease but isn't, can also help you identify if the residuals are likely to have come from a normal random variable. Kurtosis near 0 means a data set exhibits "peakedness" close to the normal. Positive kurtosis means that a data set is more peaked than a normal random variable, whereas negative kurtosis means that data is less peaked than a normal random variable. The kurtosis of a data set may be computed with the Excel KURT function.

For different size data sets, Figure 10-15 gives 95 percent confidence intervals for the skewness and kurtosis of data drawn from a normal random variable.

	Kurtosis		Skewness	
Sample Size	2.5	97.5	2.5	97.5
10	-1.74	3.41	-1.37	1.36
20	-1.27	2.46	-1.02	1.03
30	-1.09	2.06	-0.86	0.85
40	-0.99	1.77	-0.73	0.75
50	-0.91	1.62	-0.66	0.67
60	-0.85	1.49	-0.61	0.62
70	-0.80	1.36	-0.57	0.57
80	-0.77	1.27	-0.53	0.54
90	-0.73	1.20	-0.51	0.51
100	-0.71	1.13	-0.48	0.48

Figure 10-15: 95 percent confidence interval for skewness and kurtosis for sample from a normal distribution

For example, it is 95 percent certain that in a sample of size 50 from a normal random variable, kurtosis is between -0.91 and 1.62 . It is also 95 percent certain that

in a sample of size 50 from a normal random variable, skewness is between -0.66 and 0.67 . If your residuals yield a skewness or kurtosis outside the range shown in Figure 10-15, then you have reason to doubt the normality assumption.

In the computer spending example for European countries, you obtained a skewness of 0.83 and a kurtosis of 0.18 . Both these numbers are inside the ranges specified in Figure 10-15, so you have no reason to doubt the normality of the residuals.

Non-normality of the residuals invalidates the p-values that you used to determine significance of independent variables or the entire regression. The most common solution to the problem of non-normal random variables is to transform

$$\sqrt{y}, \text{ or } -$$

normality of the errors.

Heteroscedasticity: A Nonconstant Variance Error Term

If larger values of an independent variable lead to a larger variance in the errors, you have violated the constant variance of the error term assumption, and *heteroscedasticity* is present. Heteroscedasticity, like non-normal residuals, invalidates the p-values used earlier in the chapter to test for significance. In most cases you can identify heteroscedasticity by graphing the predicted value on the x-axis and the absolute value of the residual on the y-axis. To see an illustration of this, look at the file *Heteroscedasticity.xlsx*. A sample of the data is shown in Figure 10-16.

In this file, you are using the data in *Heteroscedasticity.xlsx* and trying to predict the amount a family spends annually on food from their annual income. After running a regression, you can graph the absolute value of the residuals against predicted food spending. Figure 10-17 shows the resulting graph.

The upward slope of the line that best fits the graph indicates that your forecast accuracy decreases for families with more income, and heteroscedasticity is clearly present. Usually heteroscedasticity is resolved by replacing the dependent variable Y by $\ln Y$ or $\sqrt{Y}$. The reason why these transformations often resolve heteroscedasticity is that these transformations reduce the spread in the dependent variable. For example, if three data points have $Y = 1$, $Y = 10,000$ and $Y = 1,000,000$ then after using the $\sqrt{Y}$ transformation the three points now have a dependent variable with values 1, 100, and 1000 respectively.

%4 Part III: Forecasting

	I	J
4	Income	Food spending
5	\$74,201.00	\$9,646.13
6	\$41,659.00	\$8,331.80
7	\$44,085.00	\$9,698.70
8	\$63,529.00	\$10,799.93
9	\$48,436.00	\$9,202.84
10	\$82,481.00	\$13,196.96
11	\$35,243.00	\$4,934.02
12	\$57,563.00	\$9,210.08
13	\$39,589.00	\$5,938.35
14	\$53,826.00	\$10,226.94
15	\$78,861.00	\$14,194.98
16	\$87,406.00	\$11,362.78
17	\$74,020.00	\$15,544.20
18	\$82,290.00	\$9,874.80
19	\$38,921.00	\$4,670.52
20	\$80,960.00	\$17,001.60
21	\$37,107.00	\$8,163.54
22	\$80,531.00	\$14,495.58
23	\$79,760.00	\$13,559.20
24	\$57,427.00	\$12,633.94
25	\$67,657.00	\$9,471.98
26	\$75,449.00	\$14,335.31
27	\$71,390.00	\$10,708.50

Figure 10-16: Heteroscedasticity data

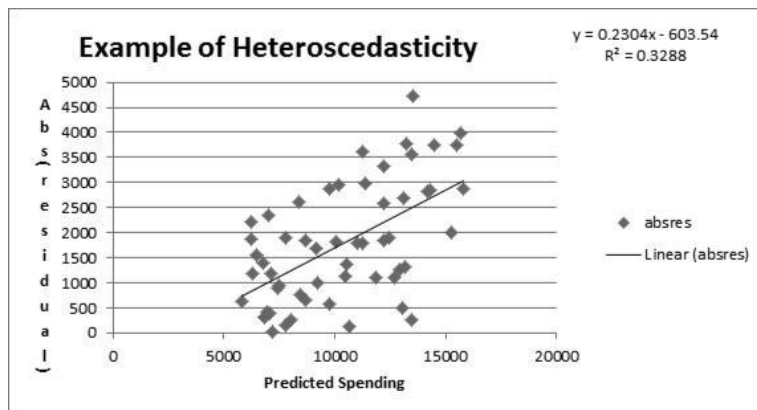


Figure 10-17: Example of Heteroscedasticity

Autocorrelation: The Nonindependence of Errors

Suppose your data is times series data. This implies the data is listed in chronological order. The auto data is a good example. The p-values used to test the hypothesis

of no linear regression and the significance of an independent variable are not valid if your error terms appear to be dependent (nonindependent). Also, if your error terms are nonindependent, you can say that *autocorrelation* is present. If autocorrelation is present, you can no longer be sure that 95 percent of your forecasts will be accurate within two standard errors. Probably fewer than 95 percent of your forecasts will be accurate within two standard errors. This means that in the presence of autocorrelation, your forecasts can give a false sense of security. Because the residuals mirror the theoretical value of the error terms in Equation 1, the easiest way to see if autocorrelation is present is to look at a plot of residuals in chronological order. Recall the residuals sum to 0, so approximately half are positive and half are negative. If your residuals are independent, you would expect sequences of the form ++, +-, -+, and -- to be equally likely. Here + is a positive residual and - is a negative residual.

Graphical Interpretation of Autocorrelation

You can use a simple time series plot of residuals to determine if the error terms exhibit autocorrelation, and if so, the type of autocorrelation that is present.

Figure 10-18 shows an illustration of independent residuals exhibiting no autocorrelation.

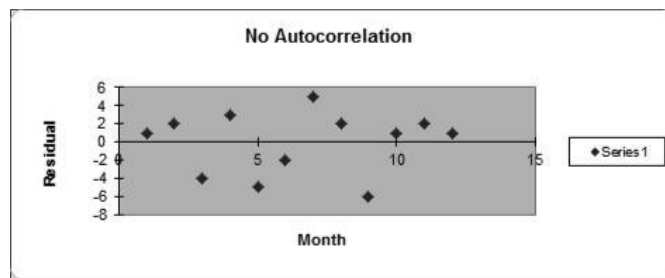


Figure 10-18: Residuals indicate no autocorrelation

Here you can see 6 changes in sign out of 11 possible changes.

Figure 10-19, however, is indicative of *positive autocorrelation*. Figure 10-19 shows only one sign change out of 11 possible changes. Positive residuals are followed by positive residuals, and negative residuals are followed by negative residuals. Thus, successive residuals are positively correlated. When residuals exhibit few sign changes (relative to half the possible number of sign changes), positive autocorrelation is suspected. Unfortunately, positive autocorrelation is common in business and economic data.

%4 Part III: Forecasting

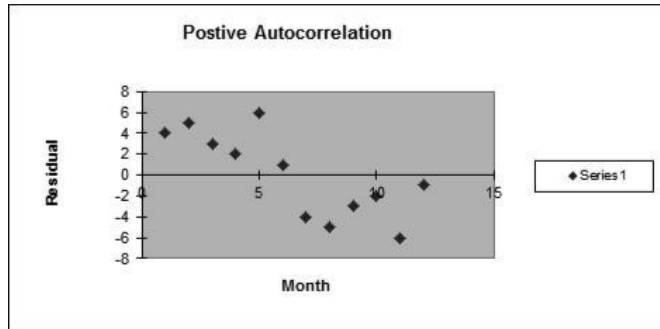


Figure 10-19: Residuals indicate positive autocorrelation

Figure 10-20 is indicative of *negative autocorrelation*. Figure 10-20 shows 11 sign changes out of a possible 11. This indicates that a small residual tends to be followed by a large residual, and a large residual tends to be followed a small residual. Thus, successive residuals are negatively correlated. This shows that many sign changes (relative to half the number of possible sign changes) are indicative of negative autocorrelation.

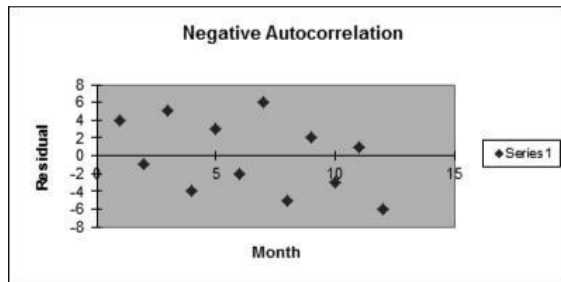


Figure 10-20: Residuals indicate negative autocorrelation

To help clarify these three different types of graphical interpretation, suppose you have n observations. If your residuals exhibit no correlation, then the chance

of seeing either less than $-\sqrt{\frac{n-1}{2}}$ or more than $\sqrt{\frac{n-1}{2}}$ sign changes is approximately 5 percent. Thus you can conclude the following:

- %4 If you observe less than or equal to $\sqrt{\frac{n-1}{2}}$ sign changes, conclude that positive autocorrelation is present.
- %4 If you observe at least $\sqrt{\frac{n-1}{2}}$ sign changes, conclude that negative autocorrelation is present.
- %4 Otherwise you can conclude that no autocorrelation is present.

Detecting and Correcting for Autocorrelation

The simplest method to correct for autocorrelation is presented in the following steps. To simplify the presentation, assume there is only one independent variable (Call it X):

- %4 Determine the correlation between the following two time series: your residuals and your residuals lagged one period. Call this correlation p
- %4 Run a regression with the dependent variable for time t being $Y_t - pY_{t-1}$ and independent variable $X_t - pX_{t-1}$
- %4 Check the number of sign changes in the new regression's residuals. Usually, autocorrelation is no longer a problem, and you can rearrange your equation to predict Y_t from Y_{t-1} , X_t , and X_{t-1}

To illustrate this procedure, you can try and predict consumer spending (in billions of \$) during a year as a function of the money supply (in billions of \$). Twenty years of data are given in Figure 10-21 and are available for download from the file `autocorr.xls`

Now complete the following steps:

- %4 Run a regression with X Range B1:B21 and Y Range A1:A21, and check the Labels and Residuals box. Figure 10-22 shows the residuals.
- %4 Observe that a sign change in the residuals occurs if, and only if, the product of two successive residuals is <0 . Therefore, copying the formula `=IF(I27*I26<0,1,0)` from J27 to J28:J45 counts the number of sign changes. Compute the total number of sign changes (4) in cell J24 with the formula `=SUM(J27:J45)`

	A	B
1	Exp	Money stock
2	214.6	159.3
3	217.7	161.2
4	219.6	162.8
5	227.2	164.6
6	230.9	165.9
7	233.3	167.9
8	234.1	168.3
9	232.3	169.7
10	233.7	170.5
11	236.5	171.6
12	238.7	173.9
13	243.2	176.1
14	249.4	178
15	254.3	179.1
16	260.9	180.2
17	263.3	181.2
18	265.6	181.6
19	268.2	182.5
20	270.4	183.3
21	275.6	184.3

Figure 10-21: Data for Autocorrelation example

%4 Part III: Forecasting

	E	F	G	H	I	J	K
23			RESIDUAL OUTPUT			sign changes	
24						4	
25			Observation	Predicted Exp	Residuals	Residual(t-1)	
26	correlation		1	211.7298848	2.87012	Sign change	
27	Residual(t)		2	216.1005891	1.59941	0	2.87012
28	and Residual(t-1)		3	219.7811822	-0.18118	1	1.59941
29	0.8227		4	223.9218494	3.27815	1	-0.18118
30			5	226.9123312	3.98767	0	3.27815
31			6	231.5130725	1.78693	0	3.98767
32			7	232.4332208	1.66678	0	1.78693
33			8	235.6537397	-3.35374	1	1.66678
34			9	237.4940363	-3.79404	0	-3.35374
35			10	240.024444	-3.52444	0	-3.79404
36			11	245.3152965	-6.6153	0	-3.52444
37			12	250.376112	-7.17611	0	-6.6153
38			13	254.7468163	-5.34682	0	-7.17611
39			14	257.277224	-2.97722	0	-5.34682
40			15	259.8076317	1.09237	1	-2.97722
41			16	262.1080024	1.192	0	1.09237
42			17	263.0281506	2.57185	0	1.192
43			18	265.0984842	3.10152	0	2.57185
44			19	266.9387808	3.46122	0	3.10152
45			20	269.2391514	6.36085	0	3.46122

Figure 10-22: Residuals for Autocorrelation example

%4 In cell J22 compute the “cutoff” for the number of sign changes that indicates the presence of positive autocorrelation. If the number of sign changes is < 5.41 , then you can suspect the positive autocorrelation is present:

$$=9.5 - \text{SQRT}(19)$$

%4 Because you have only four sign changes, you can conclude that positive autocorrelation is present.

%4 To correct for autocorrelation, find the correlation between the residuals and lagged residuals. Create the lagged residuals in K27:K45 by copying the formula =I26 from K27 to K28:K45.

%4 Find the correlation between the residuals and lagged residuals (0.82) in cell L26 using the formula =CORREL(I27:I45, K27:K45)

%4 To correct for autocorrelation run a regression with dependent variable $\text{Expenditures}_t - .82 \text{ Expenditures}_{t-1}$ and independent variable $\text{Money Supply}_t - .82 \text{ Money Supply}_{t-1}$. See Figure 10-23.

	A	B	C	D
1	Exp	Money stock	Exp(t)-.82Exp(t-1)	MS(t)-.82(MS(t-1))
2	214.6	159.3		
3	217.7	161.2	41.728	30.574
4	219.6	162.8	41.086	30.616
5	227.2	164.6	47.128	31.104
6	230.9	165.9	44.596	30.928
7	233.3	167.9	43.962	31.862
8	234.1	168.3	42.794	30.622
9	232.3	169.7	40.338	31.694
10	233.7	170.5	43.214	31.346
11	236.5	171.6	44.866	31.79
12	238.7	173.9	44.77	33.188
13	243.2	176.1	47.466	33.502
14	249.4	178	49.976	33.598
15	254.3	179.1	49.792	33.14
16	260.9	180.2	52.374	33.338
17	263.3	181.2	49.362	33.436
18	265.6	181.6	49.694	33.016
19	268.2	182.5	50.408	33.588
20	270.4	183.3	50.476	33.65
21	275.6	184.3	53.872	33.994

Figure 10-23: Transformed data to correct for autocorrelation

%4 In Column C create your transformed dependent variable by copying the formula =A3-0.82*A2 from C3 to C4:C21.

%4 Copy this same formula from D3 to D4:D21 to create the transformed independent variable $Money\ Supply_t - .82Money\ Supply_{t-1}$

%4 Now run a regression with the Y Range as C3:C21 and X Range as D3:D21. Figure 10-24 shows the results.

Because the p-value for your independent variable is less than .15, you can conclude that your transformed independent variable is useful for predicting your transformed independent variable. You can find the residuals from your new regression change sign seven times. This exceeds the positive autocorrelation cutoff of 4.37 sign changes. Therefore you can conclude that you have successfully removed the positive autocorrelation. You can predict period t expenditures with the following equation:

$$\text{Period } t \text{ expenditures} = -41.97 + 2.74(\text{Period}(t) - .82\text{Period}(t-1)) \\ \text{Money Supply} - .82\text{Period}(t-1) \text{ Money Supply})$$

%4 Part III: Forecasting

	H	I	J	K	L	M	N
1							
2	SUMMARY OUTPUT						
3			$\text{Exp}(t) = .82\text{Exp}(t-1) - 41.98 + 2.74 * (\text{MS}(t) - .82\text{MS}(t-1))$				
4	Regression Statistics		or				
5	Multiple R	0.82491	$\text{Exp}(t) = .82\text{Exp}(t-1) - 41.98 + 2.74 * (\text{MS}(t) - .82\text{MS}(t-1))$				
6	R Square	0.68048					
7	Adjusted R Square	0.66051					
8	Standard Error	2.30236					
9	Observations	18					
10							
11	ANOVA						
12		df	SS	MS	F	Significance F	
13	Regression	1	180.6254254	180.625	34.0748	2.5233E-05	
14	Residual	16	84.81364059	5.30085			
15	Total	17	265.439066				
16							
17		Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
18	Intercept	-41.9766	15.25392438	-2.75186	0.01418	-74.313504	-9.6397687
19	Var 1	2.74079	0.469526345	5.83736	2.5E-05	1.74544379	3.7361461

Figure 10-24: Regression output for transformed data

You can rewrite this equation as the following:

$$\text{Period } t \text{ expenditures} = .82\text{Period}(t-1) \text{ Expenditures} - 41.97 + 2.74(\text{Period}(t) \text{ Money Supply} - .82\text{Period}(t-1) \text{ Money Supply})$$

Because everything on the right hand side of the last equation is known at Period t , you can use this equation to predict Period t expenditures.

Multicollinearity

If two **or more independent variables** in a regression analysis are highly correlated,

%4 regression analysis may yield strange results. Whenever two or more independent variables are highly correlated and the regression coefficients do not make sense, you can say that *multicollinearity* exists.

Figure 10-25 (see file `housing.xls`) gives the following data for the years 1963–1985: the number of housing starts (in thousands), U.S. population (in millions), and mortgage rate. You can use this data to develop an equation that can forecast housing starts by performing the following steps:

It seems logical that housing starts should increase over time, so include the year as an independent variable to account for an upward trend. The more people in the United States, the more housing starts you would expect, so include Housing Starts as an independent variable. Clearly, an increase in mortgage rates decreases housing starts, so include the mortgage rate as an independent variable.

	A	B	C	D
1	Housing data			
2	thousand millions			
3	Starts	Pop	Mort Rate	Year
4	1635	189	5.89	1963
5	1561	192	5.82	1964
6	1510	194	5.81	1965
7	1196	197	6.25	1966
8	1322	199	6.46	1967
9	1545	201	6.97	1968
10	1500	203	7.8	1969
11	1434	205	8.45	1970
12	2085	208	7.74	1971
13	2379	210	7.6	1972
14	2057	212	7.96	1973
15	1353	214	8.92	1974
16	1171	216	9	1975
17	1548	218	9	1976
18	2002	220	9.02	1977
19	2036	223	9.56	1978
20	1760	225	10.78	1979
21	1312	228	12.66	1980
22	1100	230	14.7	1981
23	1072	232	15.14	1982
24	1712	234	12.57	1983
25	1756	236	12.38	1984
26	1745	238	11.55	1985

Figure 10-25: Multicollinearity data

%4 Now run a multiple regression with the Y range being A3:A26 and the X Range being B3:D26 to obtain the results shown in Figure 10-26.

%4 Observe that neither POP nor YEAR is significant. (They have p-values of .59 and .74, respectively.) Also, the negative coefficient of YEAR indicates that there is a downward trend in housing starts. This doesn't make sense though. The problem is that POP and YEAR are highly correlated. To see this, use the DATA ANALYSIS TOOLS CORRELATION command to find the correlations between the independent variables.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.660983152					
5	R Square	0.436898728					
6	Adjusted R Squ	0.347988001					
7	Standard Error	279.5855911					
8	Observations	23					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	1152331.526	384110.5087	4.913903437	0.010796446	
13	Residual	19	1485193.952	78168.10274			
14	Total	22	2637525.478				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	221794.5255	661959.9507	0.335057318	0.741252754	-1163704.005	1607293.056
18	Pop	90.39255757	162.9146358	0.554846145	0.585473195	-250.5917998	431.376915
19	Mort Rate	-206.8964396	55.49611616	-3.728124668	0.00142535	-323.0511818	-90.74169748
20	Year	-120.3847361	352.8922741	-0.341137352	0.736743719	-858.9969839	618.2275117

Figure 10-26: First regression output: Multicoillinearity example

4 Part III: Forecasting

Select Input Range B3:D26.

Check the labels box.

Put the output on the new sheet Correlation.

You should obtain the output in Figure 10-27.

	A	B	C	D
1		Pop	Mort Rate	Year
2	Pop	1		
3	Mort Rate	0.913679	1	
4	Year	0.999655	0.90995	1

Figure 10-27: Correlation matrix for Multicollinearity example

The .999 correlation between POP and YEAR occurs because both POP and YEAR increase linearly over time. Also note that the correlation between Mort Rate and the other two independent variables exceeds .9. Due to this, multicollinearity exists. What has happened is that the high correlation between the independent variables has confused the computer about which independent variables are important. The solution to this problem is to drop one or more of the highly correlated independent variables and hope that the independent variables remaining in the regression will be significant. If you decide to drop YEAR, change your X Range to B3:C26 to obtain the output shown in Figure 10-28. If you have access to a statistical package, such as SAS or SPSS, you can identify the presence of multicollinearity by looking at the Variance Inflation Factor (VIF) of each independent variable. A general rule of thumb is that any independent variable with a variance inflation factor exceeding 5 is evidence of multicollinearity.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.658369001					
5	R Square	0.433449742					
6	Adjusted R Square	0.376794716					
7	Standard Error	273.3396002					
8	Observations	23					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	1143234.737	571617.3685	7.650684739	0.003407096	
13	Residual	20	1494290.741	74714.53705			
14	Total	22	2637525.478				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-4024.025797	1627.113433	-2.473107108	0.022486161	-7418.123366	-629.928276
18	Pop	34.91659242	9.564839902	3.650515092	0.001590215	14.96469527	54.86848956
19	Mort Rate	-200.8475581	51.41238103	-3.906599034	0.000875172	-308.0918558	-93.60326032

Figure 10-28: Final regression output for Multicollinearity example

POP is now highly significant (p-value = .001). Also, by dropping YEAR you actually decreased s from 280 to 273. This decrease is because dropping YEAR reduced the

confusion the computer had due to the strong correlation between POP and YEAR. The final predictive equation is as follows:

$$\text{Housing Starts} = -4024.03 + 34.92\text{POP} - 200.85\text{MORT RAT}$$

The interpretation of this equation is that after adjusting for interest rates, an increase in U.S. population of one million people results in \$34,920 in housing starts. After adjusting for Population, an increase in interest rates of 1 percent can reduce housing starts by \$200,850. This is valuable information that could be used to forecast the future cash flows of construction-related industries.

NOTE After correcting for multicollinearity, the independent variables now have signs that agree with common sense. This is a common by-product of correcting for multicollinearity.

Validation of a Regression

The **ultimate goal of regression analysis** is for the estimated models to be used for accurate forecasting. When using a regression equation to make forecasts for the future, you must avoid over fitting a set of data. For example, if you had seven data points and only one independent variable, you could obtain an $R^2 = 1$ by fitting a sixth degree polynomial to the data. Unfortunately, such an equation would probably work poorly in fitting future data. Whenever you have a reasonable amount of data, you should hold back approximately 20 percent of your data (called the *Validation Set*) to validate your forecasts. To do this, simply fit regression to 80 percent of your data (called the *Test Set*). Compute the standard deviation of the errors for this data. Now use the equation generated from the Test Set to compute forecasts and the standard deviation of the errors for the Validation Set. Hopefully, the standard deviation for the Validation Set will be fairly close to the standard deviation for the Test Set. If this is the case, you can use the regression equation for future forecasts and be fairly confident that the accuracy of future forecasts will be approximated by the s for the Test Set. You can illustrate the important

idea of validation with the data from your housing example.

Using the years 1963–1980 as your Test Set and the years 1981–1985 as the Validation Set, you can determine the suitability of the regression with independent variables *POP* and *MORT RAT* for future forecasting using the powerful TREND function. The syntax of the TREND function is TREND(known_y's,[known_x's],[new_x's],[const]) . This function fits a multiple regression using the known y 's and known x 's and then uses this regression to make forecasts for the dependent variable using

%4 Part III: Forecasting

the new x 's data. [Constant] is an optional argument. Setting [Constant]=False causes Excel to fit the regression with the constant term set equal to 0. Setting [Constant]=True or omitting [Constant] causes Excel to fit a regression in the normal fashion.

The TREND function is an array function (see Chapter 2) so you need to select the cell range populated by the TREND function and finally press Ctrl+Shift+Enter to enable TREND to calculate the desired results. As shown in Figure 10-29 and worksheet Data , you will now use the TREND function to compare the accuracy of regression predictions for the 1981-1985 validation period to the accuracy of regression predictions for the fitted data using the following steps.

	A	B	C	D	E	F	G	H
1	Housing data						Std Dev	
2	thousand millions						1963-1980	285.701
3	Starts	Pop	Mort Rate	Year	Predictions	Error	1981-1985	255.886
4	1635	189	5.89	1963	1329.1723	305.828		
5	1561	192	5.82	1964	1490.6514	70.3486		
6	1510	194	5.81	1965	1587.9926	-77.9926		
7	1196	197	6.25	1966	1606.047	-410.047		
8	1322	199	6.46	1967	1641.5187	-319.519		
9	1545	201	6.97	1968	1592.6229	-47.6229		
10	1500	203	7.8	1969	1453.7352	46.2648		
11	1434	205	8.45	1970	1365.468	68.532		
12	2085	208	7.74	1971	1706.931	378.069		
13	2379	210	7.6	1972	1840.8314	538.169		
14	2057	212	7.96	1973	1834.1193	222.881		
15	1353	214	8.92	1974	1658.6724	-305.672		
16	1171	216	9	1975	1730.7033	-559.703		
17	1548	218	9	1976	1825.2322	-277.232		
18	2002	220	9.02	1977	1914.1366	87.8634		
19	2036	223	9.56	1978	1904.0686	131.931		
20	1760	225	10.78	1979	1655.5031	104.497		
21	1312	228	12.66	1980	1268.5938	43.4062		
22	1100	230	14.7	1981	789.42395	310.576		
23	1072	232	15.14	1982	760.21392	311.786		
24	1712	234	12.57	1983	1577.4907	134.509		
25	1756	236	12.38	1984	1725.4524	30.5476		
26	1745	238	11.55	1985	2053.3979	-308.398		

Figure 10-29: Use of Trend function to validate regression

%4 To generate forecasts for the years 1963–1985 using the 1963–1980 data, simply select the range E4:E26 and enter in E4 the array formula =TREND(A4:A21, B4:C21,B4:C26) (refer to Figure 10-29). Rows 4-21 contain the data for the years 1963-1980 and Rows 4-26 contain the data for the years 1963-1985.

%4 Compute the error for each year's forecast in Column F. The error for 1963 is computed in F4 with the formula =A4-F4

%4 Copy this formula down to row 26 to compute the errors for the years 1964–1985.

%4 In cell H2 compute the standard deviation (285.70) of the errors for the years 1963–1980 with the formula =STDEV(F4:F21)

%4 In cell H3 compute the standard deviation (255.89) of the forecast errors for the years 1981–1985 with the formula =AVERAGE(F22:F26)

The forecasts are actually more accurate for the Validation Set! This is unusual, but it gives you confidence that 95 percent of all future forecasts should be accurate within $2s = 546,700$ housing starts.

Summary

In this chapter you learned the following:

%4 The multiple linear regression model models a dependent variable Y as $B_0 + B_1X_1 + B_2X_2 + \dots B_nX_n + \text{error term}$

%4 The error term is required to satisfy the following assumptions:

The error term is normally distributed.

The variability or spread of the error term is assumed not to depend on the value of the dependent variable.

For time series data, successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.

%4 Violation of these assumptions can invalidate the p-values in the Excel output.

%4 You can run a regression analysis using the Data Analysis Tool.

%4 The Coefficients portion of the output gives the least squares estimates of $B_0, B_1, \dots, B_n$

%4 A Significance F in the ANOVA section of the output less than .05 causes you to reject the hypothesis of no linear regression and conclude that your independent variables have significant predictive value.

%4 Independent variables with p-value greater than .05 should be deleted, and the regression should be rerun until all independent variables have p-values of .05 or less.

%4 Approximately 68 percent of predictions from a regression should be accurate within one standard error and approximately 95 percent of predictions from a regression should be accurate within two standard errors.

%4 Qualitative independent variables are modeled using indicator variables.

%4 By adding the square of an independent variable as a new independent variable, you can test whether the independent variable has a nonlinear effect on Y

%4 Part III: Forecasting

By adding the product of two independent variables (say X_1 and X_2) as a new independent variable, you can test whether X_1 and X_2 interact in their effect on Y .

You can check for the presence of autocorrelation in a regression based on time series data by examining the number of sign changes in the residuals; too few sign changes indicate positive autocorrelation and too many sign changes indicate negative autocorrelation.

If independent variables are highly correlated, then their coefficients in a regression may be misleading. This is known as multicollinearity.

Exercises

%4 Fizzy Drugs wants to optimize the yield from an important chemical process. The company thinks that the number of pounds produced each time the process runs depends on the size of the container used, the pressure, and the temperature. The scientists involved believe the effect to change one variable might depend on the values of other variables. The size of the process container must be between 1.3 and 1.5 cubic meters; pressure must be between 4 and 4.5 mm; and temperature must be between 22 and 30 degrees Celsius. The scientists patiently set up experiments at the lower and upper levels of the three control variables and obtain the data shown in the file Fizzy.xlsx

Determine the relationship between yield, size, temperature, and pressure.

Discuss the interactions between pressure, size, and temperature.

What settings for temperature, size, and pressure would you recommend?

%4 For 12 straight weeks, you have observed the sales (in number of cases) of canned tomatoes at Mr. D's Supermarket. (See the file Grocery.xlsx.) Each week, you keep track of the following:

Was a promotional notice for canned tomatoes placed in all shopping carts?

Was a coupon for canned tomatoes given to each customer?

Was a price reduction (none, 1, or 2 cents off) given?

Use this data to determine how the preceding factors influence sales. Predict sales of canned tomatoes during a week in which you use a shopping cart notice, a coupon, and reduce price by 1 cent.

%4 The file `Countryregion.xlsx` contains the following data for several under-developed countries:

- Infant mortality rate
- Adult literacy rate
- Percentage of students finishing primary school
- Per capita GNP

Use this data to develop an equation that can be used to predict infant mortality. Are there any outliers in this set of data? Interpret the coefficients in your equation. Within what value should 95 percent of your predictions for infant mortality be accurate?

%4 The file `Baseball96.xlsx` gives runs scored, singles, doubles, triples, home runs, and bases stolen for each major league baseball team during the 1996 season. Use this data to determine the effects of singles, doubles, and other activities on run production.

%4 The file `Cardata.xlsx` provides the following information for 392 different car models:

- Cylinders
- Displacement
- Horsepower
- Weight
- Acceleration
- Miles per gallon (MPG)

Determine an equation that can predict MPG. Why do you think all the independent variables are not significant?

%4 Determine for your regression predicting computer sales whether the residuals exhibit non-normality or heteroscedasticity.

%4 The file `Oreos.xlsx` gives daily sales of Oreos at a supermarket and whether Oreos were placed 7" from the floor, 6" from the floor, or 5" from the floor. How does shelf position influence Oreo sales?

%4 The file `USmacrodata.xlsx` contains U.S. quarterly GNP, Inflation rates, and Unemployment rates. Use this file to perform the following exercises:

%4 Part III: Forecasting

Develop a regression to predict quarterly GNP growth from the last four quarters of growth. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.

Develop a regression to predict quarterly inflation rate from the last four quarters of inflation. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.

Develop a regression to predict quarterly unemployment rate from the unemployment rates of the last four quarters. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.

Does our regression model for predicting auto sales exhibit autocorrelation, non-normality of errors, or heteroscedasticity?

11

Events Forecasting in the Presence of Special

Often special factors such as seasonality and promotions affect demand for a product. The Excel Regression tool can handle 15 independent variables, but many times that isn't enough. This chapter shows how to use the Excel Solver to build forecasting models involving up to 200 changing cells. The discussion is based on a student project (admittedly from the 1990s) that attempted to forecast the number of customers visiting the Eastland Plaza Branch of the Indiana University (IU) Credit Union each day. You'll use this project to learn how to forecast in the face of special factors.

Building the Basic Model

In this section **you will learn how to build a model** to forecast daily customer count at the Indiana University Credit Union. The development of the model should convince you that careful examination of outliers can result in more accurate forecasting.

The data collected for this example is contained in the original worksheet in the Creditunion.xlsx file and is shown in Figure 11-1. It is important to note that this data is before direct deposit became a common method to deposit paychecks. For each day of the year the following information is available:

- %4 Month of the year
- %4 Day of the week
- %4 Whether the day was a faculty or staff payday
- %4 Whether the day before or the day after was a holiday

If you try to run a regression on this data by using dummy variables (as described in Chapter 10, "Using Multiple Regression to Forecast Sales,") the dependent variable would be the number of customers arriving each day (the data in column E). Nineteen independent variables are needed:

- %4 11 to account for the month (12 months minus 1)
- %4 4 to account for the day of the week (5 business days minus 1)

%4 Part III: Forecasting

2 to account for the types of paydays that occur each month

2 to account for whether a particular day follows or precedes a holiday

	B	C	D	E	F	G	H	I	J
1								RSQ	0.771
2									
3	MONTH	DAYMON	DAYWEEK	CUST	SPECIAL	SP	FAC	BH	AH
4	1	2	2	1825	SP,FAC,AE	1	1	0	1
5	1	3	3	1257		0	0	0	0
6	1	4	4	969		0	0	0	0
7	1	5	5	1672	SP		1	0	0
8	1	8	1	1098		0	0	0	0
9	1	9	2	691		0	0	0	0
10	1	10	3	672		0	0	0	0
11	1	11	4	754		0	0	0	0
12	1	12	5	972		0	0	0	0
13	1	15	1	816		0	0	0	0
14	1	16	2	717		0	0	0	0
15	1	17	3	728		0	0	0	0
16	1	18	4	711		0	0	0	0
17	1	19	5	1545	SP		1	0	0
18	1	22	1	873		0	0	0	0
19	1	23	2	713		0	0	0	0
20	1	24	3	626		0	0	0	0
21	1	25	4	653		0	0	0	0
22	1	26	5	1080		0	0	0	0
23	1	29	1	650		0	0	0	0
24	1	30	2	644		0	0	0	0
25	1	31	3	803		0	0	0	0

Figure 11-1: Data for Credit Union example

Microsoft Office Excel enables only 15 independent variables, so when a regression forecasting model requires more you can use the Excel Solver feature to estimate the coefficients of the independent variables. As you learned earlier, Excel's Solver can be used to optimize functions. The trick here is to apply Solver to minimize the sum of squared errors, which is the equivalent to running a regression. Because the Excel Solver allows up to 200 changing cells, you can use Solver in situations where the Excel Regression tool would be inadequate. You can also use Excel to compute the R-squared values between forecasts and actual customer traffic as well as the standard deviation for the forecast errors. To analyze this data, you create a forecasting equation by using a lookup table to "look up" the day of the week, the month, and other factors. Then you use Solver to choose the coefficients for each level of each factor that yields the minimum sum of squared errors. (Each day's error equals actual customers minus forecasted customers.) The following steps walk you through this process:

%4 First, create indicator variables (in columns G through J) for whether the day is a staff payday (SP), faculty payday (FAC), before a holiday (BH), or after a holiday (AH). (Refer to Figure 11-1). For example, cells G4, H4, and J4 use 1 to indicate that January 2 was a staff payday, faculty payday, and after a holiday. Cell I4 contains 0 to indicate that January 2 was not before a holiday.

%4 The forecast is defined by a constant (which helps to center the forecasts so that they will be more accurate), and effects for each day of the week, each month, a staff payday, a faculty payday, a day occurring before a holiday, and a day occurring after a holiday. Insert Trial values for all these parameters (the Solver changing cells) in the cell range O4:O26, as shown in Figure 11-2. Solver can then choose values that make the model best fit the data. For each day, the forecast of customer count will be generated by the following equation:

Predicted customer count=Constant+(Month effect)+(Day of week effect)+(Staff payday effect, if any)+(Faculty payday effect, if any)+(Before holiday effect, if any)+(After holiday effect, if any)

	I	J	K	L	M	N	O	P	Q	R
1	RSQ	0.771		stdevr	163.1772					
2			SSE	6736582						
3	BH	AH	Forecast	Sq Err	Error	Day of Week			average	
4	0	1	1766.78	3389.56	58.21993	1	103.357		dayweek	0
5	0	0	709.603	299643	547.3965	2	-139.19		month	-4E-09
6	0	0	745.698	49863.78	223.302	3	-150.34			
7	0	0	1557.22	13174.18	114.7788	4	-114.25			
8	0	0	963.303	18143.29	134.697	5	300.424			
9	0	0	720.753	885.2568	-29.7533	SP	396.851			
10	0	0	709.603	1414.02	-37.6035	FAC	394.894			
11	0	0	745.698	68.92317	8.301998	BH	205.293			
12	0	0	1160.37	35483.19	-188.37	AH	254.281			
13	0	0	963.303	21698.16	-147.303	Month				
14	0	0	720.753	14.08701	-3.75327	1	-110.69			
15	0	0	709.603	338.4326	18.39654	2	-75.715			
16	0	0	745.698	1203.951	-34.698	3	-40.341			
17	0	0	1557.22	149.3565	-12.2212	4	0.02839			
18	0	0	963.303	8154.623	-90.303	5	87.8157			
19	0	0	720.753	60.11313	-7.75327	6	133.341			
20	0	0	709.603	6989.539	-83.6035	7	115.803			
21	0	0	745.698	8592.92	-92.698	8	28.7743			
22	0	0	1160.37	6459.308	-80.3698	9	-87.563			
23	0	0	963.303	98158.74	-313.303	10	-53.002			
24	0	0	720.753	5891.064	-76.7533	11	-42.761			
25	0	0	709.603	8722.913	93.39654	12	44.3091			
26	0	0	1175.57	11328.02	106.4332	constant	970.635			
27	0	0	1592.2	203224.6	450.8044					

Figure 11-2: Changing cells for Credit Union example

%4 Using this model, compute a forecast for each day's customer count by copy-ing the following formula from K4 to K5:K257:

$\$O\$26 + \text{VLOOKUP}(B4, \$N\$14:\$O\$25, 2) + \text{VLOOKUP}(D4, \$N\$4:\$O\$8, 2) + G4 * \$O\$9 + H4 * \$O\$10 + I4 * \$O\$11 + J4 * \$O\$12.$

Cell O26 picks up the constant term. VLOOKUP(B4, \$N\$14:\$O\$25, 2) picks up the month coefficient for the current month, and VLOOKUP(D4, \$N\$4:\$O\$8, 2) picks up the day of the week coefficient for

%4 Part III: Forecasting

the current week. $=G4*\$O\$9+H4*\$O\$10+I4*\$O\$11+J4*\$O\$12$ picks up the effects (if any) when the current day is SP, FAC, BH, or AH.

%4 Copy the formula $=(E4-K4)^2$, from L4 to L5:L257 to compute the squared error for each day. Then, in cell L2, compute the sum of squared errors with the formula $=SUM(L4:L257)$

%4 In cell R4, average the day of the week changing cells with the formula $=AVERAGE(O4:O8)$, and in cell R5, average the month changing cells with the formula $=AVERAGE(O14:O25)$. Later in this section you will add constraints to your Solver model which constrain the average month and day of the week effects to equal 0. These constraints ensure that a month or day of the week with a positive effect has a higher than average customer count, and a month or day of the week with a negative effect has a lower than average customer count.

%4 Use the Solver settings shown in Figure 11-3 to choose the forecast parameters to minimize the sum of squared errors.



Figure 11-3: Solver settings for Credit Union example

The Solver model changes the coefficients for the month, day of the week, BH, AH, SP, FAC, and the constant to minimize the sum of square errors. It also constrains the average day of the week and month effect to equal 0. The Solver enables you to obtain the results shown in Figure 11-2. These show that Friday is the busiest day of the week and June is the busiest month. A staff payday raises the forecast (all else being equal—in the Latin, *ceteris paribus*) by 397 customers.

Evaluating Forecast Accuracy

To evaluate the accuracy of the forecast, you compute the R^2 value between the forecasts and the actual customer count in cell J1. You use the formula `=RSQ(E4:E257,K4:K257)` to do this. This formula computes the percentage of the actual variation in customer count that is explained by the forecasting model. The independent variables explain 77 percent of the daily variation in customer count.

You can compute the error for each day in column M by copying the formula `=E4-K4` from M4 to M5:M257. A close approximation to the standard error of the forecast is given by the standard deviation of the errors. This value is computed in cell M1 by using the formula `=STDEV(M4:M257)`. Thus, approximately 68 percent of the forecasts should be accurate within 163 customers, 95 percent accurate within 326 customers, and so on.

After you evaluate the accuracy of your forecasts and compute the error, you will want to try and spot any outliers. Recall that an observation is an outlier if the absolute value of the forecast error exceeds two times the standard error of the regression. To locate the outliers, perform the following steps:

- %4 Select the range M4:M257, and then click Conditional Formatting on the Home tab.
- %4 Select New Rule and in the New Formatting Rule dialog box, choose Use a Formula to Determine Which Cells to Format.
- %4 Fill in the rule description in the dialog box, as shown in Figure 11-4.

This procedure essentially copies the formula from M4 to M5:M257 and formats the cell in red if the formula is true. This ensures that all outliers are highlighted in red.

%4 Part III: Forecasting

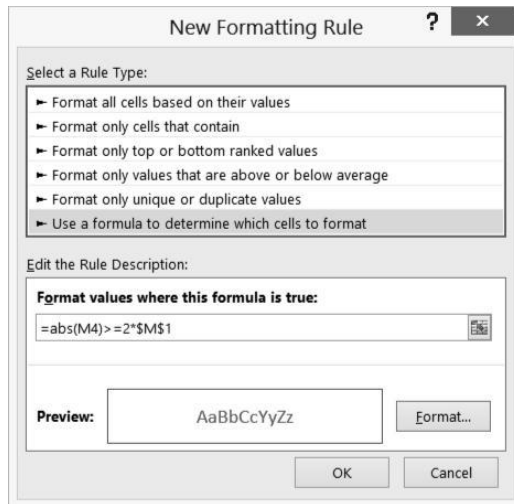


Figure 11-4: Highlighting outliers

After choosing a format with a red font, the conditional formatting settings display in red any error whose absolute value exceeds $2 \times$ (standard deviation of errors). Looking at the outliers, you see that the customer count for the first three days of the month is often under forecast. Also, during the second week in March (spring break), the data is over forecast, and the day before spring break, it is greatly under forecast.

Refining the Base Model

To remedy this problem, the 1st three days worksheet from the *Creditunion.xlsx* file shows additional changing cells for each of the first three days of the month and for spring break and the day before spring break. There are also additional trial values for these new effects in cells O26:O30. By copying the following formula from K4 to K5:K257 you can include the effects of the first three days of the month:

```
= $O$25+VLOOKUP(B4,$N$13:$O$24,2)+VLOOKUP(D4,$N$4:$O$8,2)+G4*$O$9+H4*  
$O$10+I4*$O$11+J4*$O$12+IF(C4=1,$O$26,IF(C4=2,$O$27,IF(C4=3,$O$28,0)))
```

NOTE The term `=IF(C4=1,$O$26,IF(C4=2,$O$27,IF(C4=3,$O$28,0)))` picks up the effect of the first three days of the month. For example, if Column C indicates that the day is the first day of the month, the First Day of the Month effect from O26 is added to the forecast.

You can now manually enter the spring break coefficients in cells K52:K57. For this example, you add +O29 to the formula in cell K52 and +O30 in cells K53:K57.

After including the new changing cells in the Solver dialog box, you can find the results shown in Figure 11-5. Notice that the first three days of the month greatly increase customer count (possibly because of government support and Social Security checks) and that spring break reduces customer count. Figure 11-5 also shows the improvement in forecasting accuracy. The R^2 value is improved to 87 percent and the standard error is reduced to 122 customers.

	H	I	J	K	L	M	N	O	P	Q	R	S
1		RSQ	0.871		stdeverr	122.285						
2				SSE	3783269							
3	FAC	BH	AH	Forecast	Sq Err	Error	Day of Week			average		
4	1	0	1	1879.63	2984.54	-54.631		1	107.706	dayweek	0	
5	0	0	0	995.4	68434.5	261.6		2	-138.93	month	2.5E-14	
6	0	0	0	722.934	60548.4	246.066		3	-153.32			
7	0	0	0	1554.45	13818.2	117.551		4	-115.08			
8	0	0	0	945.724	23188.1	152.276		5	299.624			
9	0	0	0	699.086	65.3826	-8.08595	SP		416.808			
10	0	0	0	684.701	161.326	-12.7014	FAC		96.6442			
11	0	0	0	722.934	965.091	31.0659	BH		196.457			
12	0	0	0	1137.64	27436.9	-165.641	AH		299.116			
13	0	0	0	945.724	16828.2	-129.724		1	-105.51			
14	0	0	0	699.086	320.913	17.9141		2	-81.763			
15	0	0	0	684.701	1874.77	43.2986		3	-27.856			
16	0	0	0	722.934	142.422	-11.9341		4	-7.2892			
17	0	0	0	1554.45	89.2902	-9.44935		5	83.8453			
18	0	0	0	945.724	5288.71	-72.7235		6	130.672			
19	0	0	0	699.086	193.601	13.9141		7	106.616			
20	0	0	0	684.701	3445.86	-58.7014		8	13.2601			
21	0	0	0	722.934	4890.78	-69.9341		9	-64.687			
22	0	0	0	1137.64	3322.49	-57.641		10	-68.305			
23	0	0	0	945.724	87452.4	-295.724		11	-33.753			
24	0	0	0	699.086	3034.46	-55.0859		12	54.7719			
25	0	0	0	684.701	13994.6	118.299	constant		943.528			
26	1	0	0	1396.77	13173.2	-114.775	d1		553.449			
27	0	0	0	1946.17	9375.37	96.8265	d2		367.977			
28	0	0	0	969.471	31162.5	176.529	d3		310.699			
29	0	0	0	722.834	294.688	17.1665	day before sp break		223.704			
30	0	0	0	708.449	109.181	-10.449	sp break		-57.035			

Figure 11-5: Credit Union model including Spring Break and First Three Days of Month factors

By looking at the forecast errors for the week 12/24 through 12/31 (see Figure 11-6), you see that you've greatly over forecasted the customer counts for the days in this week. You've also under forecasted customer counts for the week before Christmas. Further examination of the forecast errors (often called residuals) also shows the following:

- %4 Thanksgiving is different than a normal holiday in that the credit union is far less busy than expected the day after Thanksgiving.
- %4 The day before Good Friday is busy because people leave town for Easter.

%4 Part III: Forecasting

Tax day (April 16) is also busier than expected.

The week before Indiana University starts fall classes (last week in August) is not busy, possibly because many staff and faculty take a “summer fling vacation” before the hectic onrush of the fall semester.

A	B	C	D	E	F	G	H	I	J	K	L	M
								RSQ	0.871		stdev	122.285
										SSE	3783269	
	MONTH	DAY	MONDAY	WEE	CUST	SPECIAL	SP	FAC	BH	AH	Forecast	Sq Err
	12	18	2	1058	0	0	0	0	0	859.369	39454.4	198.631
	12	19	3	1104	0	0	0	0	0	844.984	67089.1	259.016
	12	20	4	1018	0	0	0	0	0	883.217	18166.5	134.783
	12	21	5	1955 SP		1	0	0	0	1714.73	57728.6	240.268
	12	24	1	941 BH		0	0	1	0	1302.46	130656	-361.463
	12	26	3	999 AH		0	0	0	1	1144.1	21054.1	-145.1
	12	27	4	619	0	0	0	0	0	883.217	69810.6	-264.217
	12	28	5	937	0	0	0	0	0	1297.92	130266	-360.924
	12	31	1	1146 BH		0	0	1	0	1302.46	24480.7	-156.463

Figure 11-6: Pre- and post-Christmas forecasts are way off!

In the Christmas week worksheet, these additional factors are included as changing cells in the forecast models. After adding the new parameters as changing cells, run Solver again. The results are shown in Figure 11-7. The R^2 is up to 92 percent and the standard error is down to 98.61 customers! Note that the post-Christmas week reduced the daily customer count by 359; the day before Thanksgiving added 607 customers; the day after Thanksgiving reduced customer count by 161, and so on.

Notice that you improve the forecasting model by using outliers. If your outliers have something in common (such as being the first three days of the month), include the common factor as an independent variable and your forecasting error drops.

The forecasting model can provide useful insights in a variety of situations. For instance, a similar analysis was performed to predict daily customer count for dinner at a major restaurant chain. The special factors corresponded to holidays. Super Bowl Sunday was the least busy day and Valentine’s Day and Mother’s Day were the busiest. Also, Saturday was the busiest day of the week for dinner, and Friday was the busiest day of the week for lunch. Using the model described in this section, after adjusting for all other factors the restaurant chain found the following:

%4 On Saturdays 192 more people than average ate dinner and on Mondays 112 fewer people than average ate dinner.

%4 On Super Bowl Sunday 212 fewer people than average ate dinner and on Mother’s Day and Valentine’s Day 350 more people ate dinner. Since an average

of 401 people ate dinner daily, the model shows that business almost doubles on Mother's Day and Valentine's Day, and business is cut in half on Super Bowl Sunday.

In contrast, for pizza delivery companies such as Domino's, Super Bowl Sunday is the busiest day of the year. Given daily counts of delivered pizzas, it would be easy to come up with an accurate estimate of the effect of Super Bowl Sunday on pizza deliveries.

	F	G	H	I	J	K	L	M	N	O	P	Q	R
1				RSQ	0.916		stderror	98.61					
2						SSE	2460390			cutoff	110.56262		
3	SPECIAL	SP	FAC	BH	AH	Forecast	Sq Err	Error	Day of Week	1	108	sign change	average
4	SP,FAC	1	1	0	1	1981	24363.99	-156		2	-155	1	dayweek
5		0	0	0	0	976	78914.97	280.9		3	-165	0	month
6		0	0	0	0	718	63163.6	251.3		4	-121	0	
7	SP	1	0	0	0	1539	17577.99	132.6		5	332	0	
8		0	0	0	0	947	22839.22	151.1			368	0	
9		0	0	0	0	684	46.55953	6.823	SP		97.1	1	
10		0	0	0	0	674	3.925658	-1.98	FAC		273	1	
11		0	0	0	0	718	1319.41	36.32	BH		478	1	
12		0	0	0	0	1171	39631.63	-199	AH		-111	0	
13		0	0	0	0	947	17127.87	-131		1	-82.1	1	
14		0	0	0	0	684	1077.379	32.82		2	-26.4	0	
15		0	0	0	0	674	2918.017	54.02		3	-34.8	1	
16		0	0	0	0	718	44.57311	-6.68		4	71	1	
17	SP	1	0	0	0	1539	31.15878	5.582		5	127	1	
18		0	0	0	0	947	5457.293	-73.9		6	94	1	
19		0	0	0	0	684	830.7915	28.82		7	60.9	1	
20		0	0	0	0	674	2302.208	-48		8	-75.3	0	
21		0	0	0	0	718	4183.025	-64.7		9	-67.9	0	
22		0	0	0	0	1171	8295.008	-91.1		10	-35.9	0	
23		0	0	0	0	947	88133.87	-297		11	80.3	0	
24		0	0	0	0	684	1614.155	-40.2		12	950	1	
25		0	0	0	0	674	16645.82	129	constant		544	1	
26	FAC	0	1	0	0	1388	11207.14	-106	d1		354	1	
27	SP	1	0	0	0	1922	14631.65	121	d2		302	0	
28		0	0	0	0	976	28935.87	170.1	d3		183	0	
29		0	0	0	0	713	718.3699	26.8	day before sp br		-55.2	1	
30		0	0	0	0	703	25.02358	-5	sp break		-359	0	
31		0	0	0	0	747	2672.615	-51.7	christmas week		183	0	
32		0	0	0	0	1200	1689.043	-41.1	before xmas week		607	0	
33		0	0	0	0	976	9004.371	-94.9	before thanks		-161	1	
34		0	0	0	0	713	3003.306	54.8	after thanks		320	1	
35		0	0	0	0	703	2401.231	-49	good thursday		-165	1	
36		0	0	0	0	747	12388.28	111.3	summerfling		244	0	
37	SP	1	0	0	0	1568	6171.827	78.56	tax day				

Figure 11-7: Final forecast model

Checking the Randomness of Forecast Errors

A good forecasting method should create forecast errors or residuals that are random. Random errors mean that the errors exhibit no discernible pattern. If forecast errors are random, the sign of your errors should change (from plus to minus or minus to plus) approximately half the time. Therefore, a commonly used test to evaluate the randomness of forecast errors is to look at the number of sign changes in the errors. If you have n observations, nonrandomness of the errors is indicated if you find either fewer than $\frac{n-1}{2} - \sqrt{n}$ or more than $\frac{n-1}{2} + \sqrt{n}$ n changes in sign. The Christmas

week worksheet, as shown in Figure 11-7, determines the number of sign changes in the residuals by copying the formula =IF(M5*M4<0,1,0) from cell P5 to P6:P257. A sign change in the residuals occurs if, and only if, the product of two consecutive

%4 Part III: Forecasting

residuals is negative. Therefore, the formula yields 1 whenever a change in the sign of the residuals occurs. In this worksheet example, there were 125 changes in sign.

Cell P1 computes $\frac{254-1}{\sqrt{254}} = 110.6$ changes in sign as the cutoff for nonrandom residuals. Therefore there are random residuals here.

Summary

In this chapter you learned the following:

- %4 The Excel Solver can be used to mimic regression analysis and work around the 15 independent variable limitations of Excel's Regression tool.
- %4 Using the Excel Solver you can often forecast daily demand with the model *Base Level + Day of Week Effect + Month Effect + Effect of Special Factors*
- %4 You can use outliers to spot omitted special factors.
- %4 If the signs of residuals change much less than half the time, then forecast errors are not random.

Exercises

- %4 How can you use the techniques outlined in this chapter to predict the daily sales of pens at Staples?
- %4 If you had several years of data, how would you incorporate a trend in the analysis?
- %4 The file *Dinner.xls* contains a model to predict daily dinner sales at a well-known chain restaurant. Column Q of the worksheet *outliers removed* contains the final forecast equation.

Explain in words the equation used to forecast daily dinner sales.

Explain how the day of the week affects dinner sales.

Explain how the time of year affects dinner sales.

Explain the special factors that affect dinner sales.

What other data might you want to collect to improve forecast accuracy?

- %4 The file *Promotiondata.xlsx* contains monthly sales (in pounds) of ice cream at a Kroger's supermarket for three years. The file also tells you when promotions occurred. Promotions are known to increase sales during the month of the promotion but decrease sales during the month after the promotion. Develop a model that can be used to predict monthly ice cream sales. Hint:

Add a term to your model involving the month number; the coefficient of this term will model the trend in ice cream sales.

- %4 What percentage of variation in ice cream sales is explained by your model?
- %4 Fill in the blank: 95 percent of forecasts for monthly sales will be accurate within _____.
- %4 What month appears to be the best for ice cream sales?
- %4 Describe the trend of ice cream sales.
- %4 Describe the effect of a promotion on ice cream sales.



12

Modeling Trend and Seasonality

Whether the marketing analyst works for a car manufacturer, airline, or consumer packaged goods company, she often must forecast sales of her company's product. Whatever the product, it is important to understand the trends (either upward or downward) and seasonal aspects of the product's sales. This chapter discusses how to determine the trends and seasonality of product sales. Using monthly data on U.S. air passenger miles (2003–2012) you will learn how to do the following:

- %4 Use moving averages to eliminate seasonality to easily see trends in sales.
- %4 Use the Solver to develop an additive or multiplicative model to estimate trends and seasonality.

Using Moving Averages to Smooth Data and Eliminate Seasonality

Moving averages smooth out noise in the data. For instance, suppose you work for Amazon.com and you are wondering whether sales are trending upward. For each January sales are less than the previous month (December sales are always high because of Christmas), so the unsuspecting marketing analyst might think there is

%4 downward trend in sales during January because sales have dropped. This conclusion is incorrect, though, because it ignores the fact that seasonal influences tend to drop January sales below December sales. You can use moving averages to smooth out seasonal data and better understand the trend and seasonality characteristics of your data.

NOTE All work in this chapter uses the file `airlinemiles.xlsx`, which contains monthly airlines miles (in thousands) traveled in the United States during the period from January 2003 through April 2012. A sample of this data is shown in Figure 12-1.

%4 Part III: Forecasting

	D	E	F
8	MonthNumber	Month	AirlineMiles (000'S)
9	1	Jan2003	32,854,790.00
10	2	Feb2003	30,814,269.00
11	3	Mar2003	37,586,654.00
12	4	Apr2003	35,226,398.00
13	5	May2003	36,569,670.00
14	6	Jun2003	39,750,216.00
15	7	Jul2003	43,367,508.00
16	8	Aug2003	42,092,669.00
17	9	Sep2003	32,549,732.00
18	10	Oct2003	36,442,428.00
19	11	Nov2003	34,350,366.00
20	12	Dec2003	37,389,382.00
21	13	Jan2004	33,537,392.00
22	14	Feb2004	33,909,139.00
23	15	Mar2004	40,805,211.00
24	16	Apr2004	40,172,829.00
25	17	May2004	39,671,007.00
26	18	Jun2004	43,652,277.00
27	19	Jul2004	46,262,249.00
28	20	Aug2004	44,701,691.00
29	21	Sep2004	35,470,844.00
30	22	Oct2004	39,627,851.00
31	23	Nov2004	37,567,116.00
32	24	Dec2004	39,117,678.00

Figure 12-1: US airline miles

To further illustrate the concept of moving averages, take a look at the graph of United States airline miles shown in Figure 12-2. To obtain this graph select the data from the Moving average worksheet of the *airlinemiles.xlsx* file in the range E8:F120 and select Insert > Charts > Scatter and choose the second option (Scatter with Smooth Lines and Markers). You obtain the graph shown in Figure 12-2.

Due to seasonality (primarily because people travel more in the summer), miles traveled usually increase during the summer and then decrease during the winter. This makes it difficult to ascertain the trend in airline travel. Graphing the *moving average* of airline miles can help to better understand the trend in this data. A 12-month moving average, for example, graphs the average of the current month's miles and the last 11 months. Because moving averages smooth out noise in the data, you can use a 12-month moving average to eliminate the influence of seasonality. This is because a 12-month moving average includes one data point for each month. When analyzing a trend in quarterly data, you should plot a four-quarter moving average.

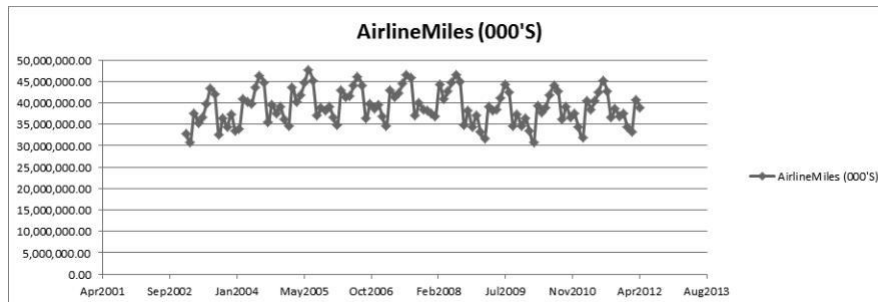


Figure 12-2: Graph of US airline miles

To overlay a 12-month moving average on the scatterplot, you return to an old friend, the Excel Trendline. Right-click the data series and select Add Trendline... Choose Moving Average and select 12 periods. Then you can obtain the trendline, as shown in Figure 12-3.

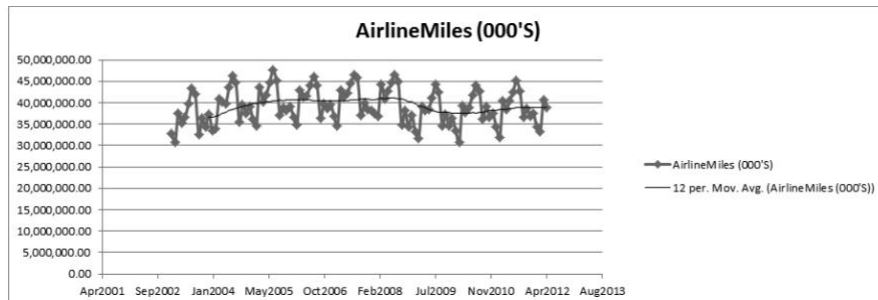


Figure 12-3: Moving average trendline

The moving average trendline makes it easy to see how airline travel trended between 2003 and 2012. You can now see the following:

- %4 In 2003 and 2004 there was a sharp upward trend in airline travel (perhaps a rebound from 9/11).
- %4 In 2005–2008 airline travel appeared to stagnate.
- %4 In late 2008 there was a sharp drop in airline travel, likely due to the financial crisis.
- %4 In 2010 a slight upward trend in air travel occurred.

%4 Part III: Forecasting

The next section uses the Excel Solver to quantify the exact nature of the trend in airline miles and also to learn how to determine how seasonality influences demand for air travel.

An Additive Model with Trends and Seasonality

Based on the previous section's discussion it should be clear that to accurately forecast sales when the data has seasonality and trends, you need to identify and separate these from the data series. In this section you learn how this process can be modeled using Excel's Solver. These analyses enable you to identify and separate between the baseline, seasonality, and trend components of a data series.

When predicting product sales, the following additive model is often used to estimate the trend and seasonal influence of sales:

$$\%4 \quad \text{Predicted Period } t \text{ Sales} = \text{Base} + \text{Trend} * \text{Period Number} + \text{Seasonal Index for Month } t$$

In Equation 1 you need to estimate the base, trend, and seasonal index for each month of the year. The work for this appears in the Additive trend worksheet (see Figure 12-4). To simplify matters the data is rescaled in billions of miles. The base, trend, and seasonal index may be described as follows:

- %4 **Base:** The base is the best estimate of the level (without seasonality) of monthly airline miles at the beginning of the observed time period.
- %4 **Trend:** The trend is the best estimate of the monthly rate of increase in airline miles traveled. A trend of 5, for example, would mean that the level of airline travel is increasing at a rate of 5 billion miles per month.
- %4 **Seasonal Index:** Each month of the year has a seasonal index to reflect if travel during the month tends to be higher or lower than average. A seasonal index of +5 for June would mean, for example, that June airline travel tends to be 5 billion miles higher than an average month.

NOTE The seasonal indices must average to 0.

	A	B	C	D	E	F	G	H	I	J	K
1											
2	baseadd	37.37856									
3	trendadd	0.059026									
4										stddeverr	0.386323
5	1	-4.45733								RSQ	0.988934
6	2	-6.62334								SSE	4.9251
7	3	1.601041									
8	4	-0.319		MonthNumber	Month	Month	Airline Miles (billion)	Forecast	Error	Sq Error	
9	5	1.274636		1	7/1/2009	7	44.22	43.7288	0.49	0.236896	
10	6	3.795057		2	8/1/2009	8	42.40	41.95583	0.44	0.194662	
11	7	6.291206		3	9/1/2009	9	34.68	35.12698	-0.45	0.203932	
12	8	4.459215		4	10/1/2009	10	37.32	37.69339	-0.38	0.140881	
13	9	-2.42866		5	11/1/2009	11	34.58	35.31697	-0.74	0.54817	
14	10	0.078726		6	12/1/2009	12	36.46	36.41789	0.04	0.001696	
15	11	-2.35673		7	1/1/2010	1	33.49	33.33441	0.15	0.023327	
16	12	-1.31482		8	2/1/2010	2	30.72	31.22743	-0.51	0.259421	
17				9	3/1/2010	3	39.37	39.51084	-0.14	0.019948	
18	mean	0		10	4/1/2010	4	37.76	37.64982	0.11	0.012653	
19				11	5/1/2010	5	38.88	39.30248	-0.42	0.175395	
20				12	6/1/2010	6	41.90	41.88193	0.02	0.000401	
21				13	7/1/2010	7	44.02	44.43711	-0.42	0.172428	
22				14	8/1/2010	8	42.81	42.66414	0.15	0.02222	
23				15	9/1/2010	9	36.13	35.8353	0.30	0.087799	
24				16	10/1/2010	10	39.18	38.4017	0.78	0.611145	
25				17	11/1/2010	11	36.67	36.02528	0.65	0.41766	

Figure 12-4: Additive trend model

To estimate base, trend, and seasonal indices, you need to create formulas based on trial values of the parameters in Column H. Then in Column I, you will determine the error for each month's forecast, and in Column J, you compute the squared error for each forecast. Finally, you use the Solver to determine the parameter values that minimize squared errors. To execute this estimation process, perform the following steps:

- %4 Enter trial values of the base and trend in cells B2 and B3. Name cell B2 **baseadd** and cell B3 **trend**
- %4 Enter trial seasonal indices in the range B5:B16.
- %4 In cell B18, average the seasonal indices with the formula `=AVERAGE(B5:B16)`. The Solver model can set this average to 0 to ensure the seasonal indices average to 0.
- %4 Copy the formula `=baseadd+trend*D9+VLOOKUP(F9,$A$5:$B$16,2)` from H9 to H10:H42 to compute the forecast for each month.
- %4 Copy the formula `=G9-H9` from I9 to I10:I42 to compute each month's forecast error.
- %4 Copy the formula `=(I9^2)` from J9 to J10:J42 to compute each month's squared error.
- %4 In cell K6, compute the Sum of Squared Errors (SSE) using the formula `=SUM(J9:J42)`

%4 Part III: Forecasting

Now set up the Solver model, as shown in Figure 12-5. Change the parameters to minimize SSE and constrain the average of the seasonal indices to 0. Do not check the non-negative box because some seasonal indices must be negative. The forecasting model of Equation 1 is a *linear forecasting model* because each unknown parameter is multiplied by a constant. When the forecasts are created by adding together terms that multiply changing cells by constants, the GRG Solver Engine always finds a unique solution to the least square minimizing parameter estimates for a forecasting model.



Figure 12-5: Additive trend Solver model

Refer to the data shown in Figure 12-4 and you can make the following estimates:

- %4 At the beginning of July 2009, the base level of airline miles is 37.38 billion.
- %4 An upward trend in airline miles is 59 billion miles per month.
- %4 The busiest month is July (6.29 billion miles above average) and the slowest month is February with 6.62 billion miles below average.

Cell K5 uses the formula `=RSQ(G9:G42,H9:H42)` to show that the model explains 98.9 percent of the variation in miles traveled. Cell K4 also computes the standard deviation of the errors (989 billion) with the formula `=STDEV(I9:I42)`. You should expect 95 percent of the predictions to be accurate within $2 * 0.386 = 0.772$ billion miles. Looking at Column I, no outliers are found.

A Multiplicative Model with Trend and Seasonality

When **predicting** product sales, the following multiplicative model is often used to estimate the trend and seasonal influence of sales:

$$(2) \text{ Predicted Period } t \text{ Sales} = \text{Base} * (\text{Trend}^t) * (\text{Seasonal Index for Month } t)$$

As in the additive model, you need to estimate the base, trend, and seasonal indices. In Equation 2 the trend and seasonal index have different meanings than in the additive model.

%4 **Trend:** The trend now represents the percentage monthly increase in the level of airline miles. For example, a trend value of 1.03 means monthly air travel is increasing 3 percent per month, and a trend value of .95 means monthly air travel is decreasing at a rate of 5 percent per month. If per period growth is independent of the current sales value, the additive trend model will probably outperform the multiplicative trend model. On the other hand, if per period growth is an increasing function of current sales, the multiplicative trend model will probably outperform the additive trend model.

%4 **Seasonal Index:** The seasonal index for a month now represents the percentage by which airline travel for the month is above or below an average month. For example, a seasonal index for July of 1.16 means July has 16 percent more air travel than an average month, whereas a seasonal index for February of .83 means February has 17 percent less air travel than an average month. Of course, multiplicative seasonal indices must average to 1. This is because months with above average sales are indicated by a seasonal index exceeding 1, while months with below average sales are indicated by a seasonal index less than 1.

The work for this equation appears in the Multiplicative trend worksheet. All the formulas are the same as the additive model with the exception of the monthly forecasts in Column H. You can implement Equation 2 by copying the formula `=base*(trend^D9)*VLOOKUP(F9,$A$5:$B$16,2)` from H9 to H10:H42.

%4 Part III: Forecasting

The forecasting model in Equation 2 is a nonlinear forecasting model because you can raise the trend to a power and multiply, rather than add terms involving the seasonal indices. For nonlinear forecasting models, the GRG Solver Engine often fails to find an optimal solution unless the starting values for the changing cells are close to the optimal solution. The remedy to this issue is as follows:

%4 In Solver select Options, and from the GRG tab, select Multistart. This ensures the Solver will try many (between 50 and 200) starting solutions and find the optimal solution from each starting solution. Then the Solver reports the “best of the best” solutions.

%4 To use the Multistart option, input lower and upper bounds on the changing cells. To speed up solutions, these bounds should approximate sensible values for the estimated parameters. For example, a seasonal index will probably be between 0 and 3, so an upper bound of 100 would be unreasonable. As shown in Figure 12-6, you can choose an upper bound of 3 for each seasonal index and an upper bound of 2 for the trend. For this example, choose an upper bound of 100 for the base.

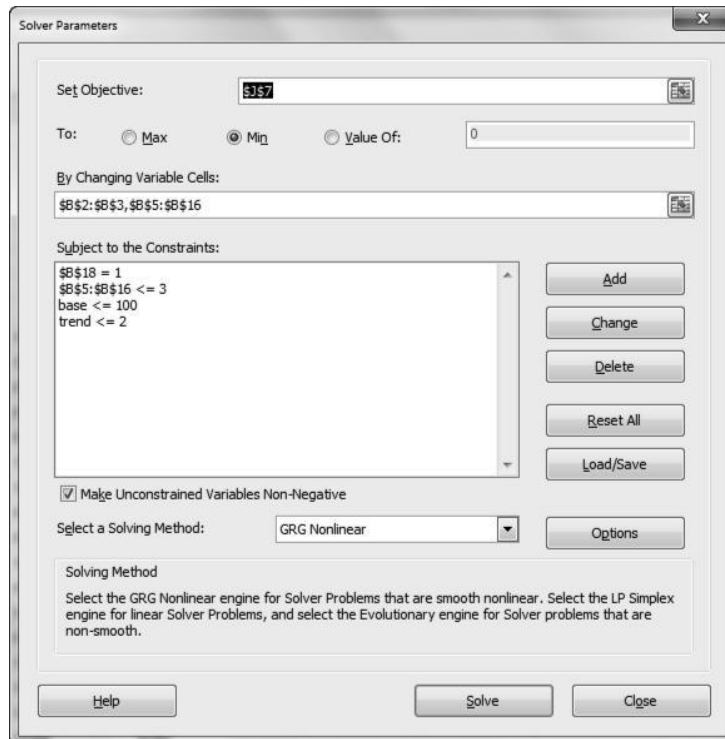


Figure 12-6: Solver window for multiplicative trend model

%4 Cell B18 averages the seasonal indices, so in the Solver window add the con-straint \$B\$18=1 to ensure that the seasonal indices average to 1.

%4 Select Solve, and the Solver will then find the optimal solution (refer to Figure 12-7).

	A	B	C	D	E	F	G	H	I	J	K
1											
2	base	3.74E+01									
3	trend	1.001493569					1.00E+06				
4										stddeverrors	0.411756002
5	1	0.884049011								RSQ	0.987429707
6	2	0.82837254								SSE	
7	3	1.041400111									5.59
8	4	0.991684904		MonthNumber	Month	Month	AirlineMiles (b Forecast	Error		Sq Error	
9	5	1.03315296		1	7/1/2009	7	44.22	43.63945105	0.58	0.33	
10	6	1.098599337		2	8/1/2009	8	42.40	41.89570886	0.50	0.25	
11	7	1.164327334		3	9/1/2009	9	34.68	35.19980697	-0.52	0.28	
12	8	1.116136195		4	10/1/2009	10	37.32	37.71609694	-0.40	0.16	
13	9	0.936353344		5	11/1/2009	11	34.58	35.38768125	-0.81	0.66	
14	10	1.00179316		6	12/1/2009	12	36.46	36.46161088	0.00	0.00	
15	11	0.938545346		7	1/1/2010	1	33.49	33.43255124	0.05	0.00	
16	12	0.965585734		8	2/1/2010	2	30.72	31.37379361	-0.66	0.43	
17				9	3/1/2010	3	39.37	39.50091231	-0.13	0.02	
18	mean	0.999999998		10	4/1/2010	4	37.76	37.67136647	0.09	0.01	
19				11	5/1/2010	5	38.88	39.30524076	-0.42	0.18	
20				12	6/1/2010	6	41.90	41.85750467	0.04	0.00	
21				13	7/1/2010	7	44.02	44.4280508	-0.41	0.16	
22				14	8/1/2010	8	42.81	42.65279779	0.16	0.03	
23				15	9/1/2010	9	36.13	35.8358956	0.30	0.09	
24				16	10/1/2010	10	39.18	38.39765695	0.79	0.62	
25				17	11/1/2010	11	36.67	36.02716493	0.64	0.42	

Figure 12-7: Multiplicative trend model

NOTE If the Solver assigns a changing cell, a value near its lower or upper bound should be relaxed. For example, if you set the upper bound for the base to 30, the Solver will find a value near 30, thereby indicating the bound should be relaxed.

From the optimal Solver solution you find the following:

- %4 The estimated base level of airline miles is 37.4 billion.
- %4 You can estimate airline miles increase at a rate of 0.15 percent per month or $1.00149^{12} - 1 = 1.8$ percent per year.
- %4 The busiest month for the airlines is July, when miles traveled are 16 percent above average, and the least busy month is February, during which miles traveled are 17 percent below average.

A natural question is whether the additive or multiplicative model should be used to predict airline miles for future months. Because the additive model has a lower standard deviation of residuals, you should use the additive model to forecast future airline miles traveled.

Summary

In this chapter you learned the following:

- %4 Using a 12-month or 4-quarter moving average chart enables you to easily see the trend in a product's sales.
- %4 You can often use seasonality and trend to predict sales by using the following equation:

$$\text{Predicted Period } t \text{ Sales} = \text{Base} + \text{Trend} * \text{Period Number} + \text{Seasonal Index for Month } t$$

- %4 You can often use the following equation to predict sales of a product: Predicted period t Sales = Base * (Trend ^{t}) * (Seasonal Index for Month t)

Exercises

The following exercises use the file `airlinedata.xlsx`, which contains monthly U.S. domestic air miles traveled during the years 1970–2004.

- %4 Determine the trend and seasonality for the years 1970–1980.
- %4 Determine the trend and seasonality for the years 1981–1990.
- %4 Determine the trend and seasonality for the years 1995–2004.

13

Ratio to Moving Average Forecasting Method

In Chapter 12, “Modeling Trend and Seasonality,” you learned how to estimate trend and seasonal indices. Naturally you would like to use your knowledge of trend and seasonality to make accurate forecasts of future sales. The *Ratio to Moving Average Method* provides an accurate, easy-to-use forecasting method for future monthly or quarterly sales. This chapter shows how to use this method to easily estimate seasonal indices and forecast future sales.

Using the Ratio to Moving Average Method

The **simple Ratio to Moving Average Forecasting Method** is described in this section via examples using data from the *Ratioma.xlsx* file, which includes sales of a product during 20 quarters (as shown in Figure 13-1 in rows 5 through 24). This technique enables you to perform two tasks:

- %4 Easily estimate a time series’ trend and seasonal indices.
- %4 Generate forecasts of future values of the time series.

Using the first 20 quarters for the data exemplified in this chapter, you will be able to forecast sales for the following four quarters (Quarters 21 through 24). Similar to the one in Chapter 12, this time series data has both trend and seasonality.

The Ratio to Moving Average Method has four main steps:

- %4 Estimate the deseasonalized level of the series during each period (using centered moving averages).
- %4 Fit a trend line to your deseasonalized estimates (in Column G).
- %4 Determine the seasonal index for each quarter and estimate the future level of the series by extrapolating the trend line.
- %4 Predict future sales by reseasonalizing the trend line estimate.

%4 Part III: Forecasting

	B	C	D	E	F	G	H	I	J	K	L
1				slope	6.9387868						
2				intercept	30.166176				quarter	seasonal index	normalized
3									1	0.818547	0.81373678
4	Quarter#	Year	Quarter	Sales	4 period MA	Centered MA	Actual/CMA	Forecast	2	0.93934	0.9338196
5	1	1	1	24					3	1.067364	1.06109143
6	2	1	2	44	52				4	1.198394	1.19135219
7	3	1	3	61	58	55.00	1.11				
8	4	1	4	79	63.5	60.75	1.30				
9	5	2	1	48	71	67.25	0.71				
10	6	2	2	66	77.5	74.25	0.89				
11	7	2	3	91	82.5	80.00	1.14				
12	8	2	4	105	87.25	84.88	1.24				
13	9	3	1	68	89.5	88.38	0.77				
14	10	3	2	85	94.5	92.00	0.92				
15	11	3	3	100	104.25	99.38	1.01				
16	12	3	4	125	114.25	109.25	1.14				
17	13	4	1	107	123.75	119.00	0.90				
18	14	4	2	125	132.25	128.00	0.98				
19	15	4	3	138	139.25	135.75	1.02				
20	16	4	4	159	146.75	143.00	1.11				
21	17	5	1	135	156	151.38	0.89				
22	18	5	2	155	164.25	160.13	0.97				
23	19	5	3	175							
24	20	5	4	192							
25	21	6	1			175.880699		143.121			
26	22	6	2			182.819485		170.72			
27	23	6	3			189.758272		201.351			
28	24	6	4			196.697059		234.335			

Figure 13-1: Example of Ratio to Moving Average Method

The following sections walk you through each main part of this process.

Calculating Moving Averages and Centered Moving Averages

To begin, you compute a four-quarter (four quarters eliminates seasonality) moving average for each quarter by averaging the prior quarter, current quarter, and next two quarters. To do this you copy the formula `=AVERAGE(E5:E8)` down from cell F6 to F7:F22. For example, for Quarter 2, the moving average is $(24 + 44 + 61 + 79) / 4 = 52$.

Because the moving average for Quarter 2 averages Quarters 1 through 4 and the numbers 1–4 average to 2.5, the moving average for Quarter 2 is centered at Quarter 2.5. Similarly, the moving average for Quarter 3 is centered at Quarter 3.5. Therefore, averaging these two moving averages gives a centered moving average that estimates the level of the process at the end of Quarter 3. To estimate the level of the series during each series (without seasonality), copy the formula `=AVERAGE(F6:F7)` down from cell G7.

Fitting a Trend Line to the Centered Moving Averages

You can use the centered moving averages to fit a trend line that can be used to estimate the future level of the series. To do so, follow these steps:

- %4 In cell F1 use the formula `=SLOPE(G7:G22,B7:B22)` to find the slope of the trend line.
- %4 In cell F2 use the formula `=INTERCEPT(G7:G22,B7:B22)` to find the intercept of the trend line.
- %4 Estimate the level of the series during Quarter t to be $6.94t + 30.17$.
- %4 Copy the formula `=intercept + slope*B25` down from cell G25 to G26:G28 to compute the estimated level (excluding seasonality) of the series from Quarter 21 onward.

Compute the Seasonal Indexes

Recall that a seasonal index of 2 for a quarter means sales in that quarter are twice the sales during an average quarter, whereas a seasonal index of .5 for a quarter would mean that sales during that quarter were one-half of an average quarter. Therefore, to determine the seasonal indices, begin by determining for each quarter for which you have sales (*Actual Sales*) / *Centered Moving Average*. To do this, copy the formula `=E7/G7` down from cell H7 to H8:H22. You find, for example, that during Quarter 1 sales were 77 percent, 71 percent, 90 percent and 89 percent of average, so you could estimate the seasonal index for Quarter 1 as the average of these four numbers (82 percent). To calculate the initial seasonal index estimates, you can copy the formula `=AVERAGEIF($D$7:$D$22,J3,$H$7:$H$22)` from cell K3 to K4:K6. This formula averages the four estimates you have for Q1 seasonality.

Unfortunately, the seasonal indices do not average exactly to 1. To ensure that your final seasonal indices average to 1, copy the formula `=K3/AVERAGE($K$3:$K$6)` from cell L3 to L4:L6.

Forecasting Sales during Quarters 21–24

To create your sales forecast for each future quarter, simply multiply the trend line estimate for the quarter's level (from Column G) by the appropriate seasonal index. Copy the formula `=VLOOKUP(D25,season,3)*G25` from cell G25 to G26:G28 to compute the final forecast for Quarters 21–24. This forecast includes estimates of trend and seasonality.

%4 Part III: Forecasting

If you think the trend of the series has changed recently, you can estimate the series' trend based on more recent data. For example, you could use the centered moving averages for Quarters 13–18 to get a more recent trend estimate by using the formula `=SLOPE(G17:G22,B17:B22)`. This yields an estimated trend of 8.09 units per quarter. If you want to forecast Quarter 22 sales, for example, you take the last centered moving average you have (from Quarter 18) of 160.13 and add 4 (8.09) to estimate the level of the series in Quarter 22. Then multiply the estimate of the Quarter 22 level by the Quarter 2 seasonal index of .933 to yield a final forecast for Quarter 22 sales of $(160.13 + 4(8.09)) * (.933) = 179.6$ units.

Applying the Ratio to Moving Average Method to Monthly Data

Often the **Ratio to Moving Average Method** is used to forecast monthly sales as well as quarterly sales. To illustrate the application of this method to monthly data, let's look at U.S. housing starts.

The `Housingstarts.xlsx` file gives monthly U.S. housing starts (in thousands) for the period January 2000 through May 2011. Based on the data through November 2010, you can apply the Ratio to Moving Average Method to forecast monthly U.S. housing starts for the period December 2010 through May 2011. You can forecast a total of 3.5 million housing starts, and in reality there were 3.374 million housing starts. The key difference between applying the method to monthly and quarterly data is that for monthly data you need to use 12-month moving averages to eliminate seasonality.

Summary

In this chapter you learned the following:

%4 Applying the Ratio to Moving Average Method involves the following tasks:

- Compute four-quarter moving averages and then determine the centered moving averages.
- Fit a trend line to the centered moving averages.
- Compute seasonal indices.
- Compute forecasts for future periods.

%4 You can apply the Ratio to Moving Average Method to monthly data as well by following the same process but use 12-month moving averages to eliminate seasonality.

Exercises

- %4** The file `Walmartdata.xls` contains quarterly revenues of Wal-Mart during the years 1994–2009. Use the Ratio to Moving Average Method to forecast revenues for Quarters 3 and 4 in 2009 and Quarters 1 and 2 in 2010. Use Quarters 53–60 to create a trend estimate that you use in your forecasts.
- %4** Based on the data in the file `airlinemiles.xlsx` from Chapter 12, use the Ratio to Moving Average Method to forecast airline miles for the remaining months in 2012.



14

Winter's Method

Predicting future values of a time series is usually difficult because the characteristics of any time series are constantly changing. For instance, as you saw in Chapter 12, “Modeling Trend and Seasonality,” the trend in U.S. airline passenger miles changed several times during the 2000–2012 period. *Smoothing* or *adaptive* methods are usually best suited for forecasting future values of a time series. Essentially, smoothing methods create forecasts by combining information from a current observation with your prior view of a parameter, such as trend or a seasonal index. Unlike many other smoothing methods, Winter’s Method incorporates both trend and seasonal factors. This makes it useful in situations where trend and seasonality are important. Because in an actual situation (think U.S. monthly housing starts) trend and seasonality are constantly changing, a method such as Winter’s Method that changes trend and seasonal index estimates during each period has a better chance of keeping up with changes than methods like the trend and seasonality approaches based on curve fitting discussed in Chapter 12, which use constant estimates of trend and seasonal indices.

To help you understand how Winter’s Method works, this chapter uses it to forecast airline passenger miles for April through December 2012 based on the data studied in Chapter 12. This chapter describes the three key characteristics of a time series (level, trend, and seasonality) and explains the initialization process, notation, and key formulas needed to implement Winter’s Method. Finally, you explore forecasting with Winter’s Method and the concept of Mean Absolute Percentage Error (MAPE).

Parameter Definitions for Winter’s Method

In this chapter you will develop **Winter’s exponential smoothing method** using the three time series characteristics, level (also called base), trend, and seasonal index,

Part III: Forecasting

discussed in Chapter 12 in the “Multiplicative Model with Trend and Seasonality” section. After observing data through the end of month t you can estimate the following quantities of interest:

L = Level of series

T = Trend of series

S = Seasonal index for current month

The key to Winter’s Method is the use of the following three equations, which are used to update L , T , and S . In the following equations, α , β , and γ are called *smoothing parameters*. The values of these parameters will be chosen to optimize your forecasts. In the following equations, c equals the number of periods in a seasonal cycle ($c = 12$ months for example) and x_t equals the observed value of the time series at time t

$$L_t = \alpha(x_t / (s_{t-c}) + (1 - \alpha)(L_{t-1} * T_{t-1}))$$

$$T_t = \beta(L_t / L_{t-1}) + (1 - \beta) T_{t-1}$$

$$S_t = \gamma(x_t / L_t) + (1 - \gamma)s_{(t-c)}$$

Equation 1 indicates that the new base estimate is a weighted average of the current observation (deseasonalized) and last period’s base is updated by the last trend estimate. Equation 2 indicates that the new trend estimate is a weighted average of the ratio of the current base to last period’s base (this is a current estimate of trend) and last period’s trend. Equation 3 indicates that you update the seasonal index estimate as a weighted average of the estimate of the seasonal index based on the current period and the previous estimate. In equations 1–3 the first term uses an estimate of the desired quantity based on the current observation and the second term uses a past estimate of the desired quantity.

NOTE Note that larger values of the smoothing parameters correspond to putting more weight on the current observation.

You can define F_{t+k} as your forecast (F) after period t for the period $t + k$. This results in the following equation:

$$F_{t,k} = L_t * (T_t)^k * s_{t+k-c}$$

Equation 4 first uses the current trend estimate to update the base k periods forward. Then the resulting base estimate for period $t + k$ is adjusted by the appropriate seasonal index.

Initializing Winter's Method

To start **Winter's Method**, you must have initial estimates for the series base, trend, and seasonal indices. You can use the data from the `airline winters.xls` file, which contains monthly U.S. airline passenger miles for the years 2003 and 2004 to obtain initial estimates of level, trend, and seasonality. See Figure 14-1.

	A	B	C	D	E	F	G	H	I
1									
2	base	3.51E+01							SSE
3	trend	1.006491							5.63
4			MonthNu	Month	Month	AirlineMiles (billions)	Forecast	Error	Sq Error
5	1	0.90305	1	Jan2003	1	32.85	31.87289	0.98	0.96
6	2	0.875947	2	Feb2003	2	30.81	31.11699	-0.30	0.09
7	3	1.053815	3	Mar2003	3	37.59	37.6785	-0.09	0.01
8	4	1.008006	4	Apr2003	4	35.23	36.27456	-1.05	1.10
9	5	1.011706	5	May2003	5	36.57	36.644	-0.07	0.01
10	6	1.099865	6	Jun2003	6	39.75	40.09572	-0.35	0.12
11	7	1.173706	7	Jul2003	7	43.37	43.06529	0.30	0.09
12	8	1.129149	8	Aug2003	8	42.09	41.69932	0.39	0.15
13	9	0.879644	9	Sep2003	9	32.55	32.696	-0.15	0.02
14	10	0.977359	10	Oct2003	10	36.44	36.5638	-0.12	0.01
15	11	0.918146	11	Nov2003	11	34.35	34.57155	-0.22	0.05
16	12	0.969609	12	Dec2003	12	37.39	36.74627	0.64	0.41
17			13	Jan2004	1	33.54	34.44595	-0.91	0.83
18	mean	1	14	Feb2004	2	33.91	33.62903	0.28	0.08
19			15	Mar2004	3	40.81	40.72023	0.08	0.01
20			16	Apr2004	4	40.17	39.20296	0.97	0.94
21			17	May2004	5	39.67	39.60222	0.07	0.00
22			18	Jun2004	6	43.65	43.33259	0.32	0.10
23			19	Jul2004	7	46.26	46.54189	-0.28	0.08
24			20	Aug2004	8	44.70	45.06565	-0.36	0.13
25			21	Sep2004	9	35.47	35.3355	0.14	0.02
26			22	Oct2004	10	39.63	39.51555	0.11	0.01
27			23	Nov2004	11	37.57	37.36247	0.20	0.04
28			24	Dec2004	12	39.12	39.71275	-0.60	0.35

Figure 14-1: Data for Winter's Method

In the Initial worksheet you can fit the Multiplicative Trend Model from Chapter 12 to the 2003–2004 data. As shown in Figure 14-2, you use the trend and seasonal index from this fit as the original seasonal index and the December 2004 trend. Cell C25 determines an estimate of the base for December 2004 by deseasonalizing the observed December 2004 miles. This is accomplished with the formula $= (B25/H25)$

%4 Part III: Forecasting

#	A	B	C	D	E	F	G	H	I	J	K
1	DATE	Airline Miles(billions)									
2	Jan2003	32.85									
3	Feb2003	30.81									
4	Mar2003	37.59									
5	Apr2003	35.23									
6	May2003	36.57									
7	Jun2003	39.75									
8	Jul2003	43.37									
9	Aug2003	42.09									
10	Sep2003	32.55						alp	bet	gam	
11	Oct2003	36.44						0.548512014	0.049142	0.58877279	
12	Nov2003	34.35									
13	Dec2003	37.39									
14	Jan2004	33.54						0.903049602			
15	Feb2004	33.91						0.875947455			
16	Mar2004	40.81						1.053814727			
17	Apr2004	40.17						1.00800602			
18	May2004	39.67						1.011705575			
19	Jun2004	43.65						1.099865309			
20	Jul2004	46.26						1.173705527			
21	Aug2004	44.70				SSE	77.8196	1.129148564			
22	Sep2004	35.47				stdeverror	0.9369659	0.879643964		34 sign changes of 87	
23	Oct2004	39.63						0.977358616		MAPE	
24	Nov2004	37.57	Base	Trend	Forecast	Error	Sq Error	0.918146041	34	0.02062096	
25	Dec2004	39.12	40.34378	1.006491				0.969608616	Sign char	APE	
26	Jan2005	36.12	40.27083	1.006083	36.6689	-0.55	0.3038	0.899411078		0.01526167	
27	Feb2005	34.56	39.93414	1.005373	35.4897	-0.93	0.8628	0.869764432	0	0.02687615	
28	Mar2005	43.64	40.8425	1.006227	42.3093	1.33	1.7767	1.062490091	1	0.03054212	
29	Apr2005	40.24	40.45404	1.005453	41.42583	-1.18	1.3953	1.000244124	1	0.02935135	
30	May2005	41.80	41.02748	1.005882	41.15077	0.65	0.4235	1.015922148	1	0.01556855	
31	Jun2005	44.68	40.91303	1.005456	45.39013	-0.71	0.5089	1.095230183	1	0.0159679	
32	Jul2005	47.56	40.80036	1.005052	48.28184	-0.72	0.5166	1.169022868	0	0.01511097	

Figure 14-2: Initialization of Winter's Method

The next part of Winter's Method includes choosing the smoothing parameters to optimize the one-month-ahead forecasts for the years 2005 through 2012.

Estimating the Smoothing Constants

After observing each month's airline miles (in billions), you are now ready to update the smoothing constants. In Column C, you will update the series base; in Column D, the series trend; and in Column H, the seasonal indices. In Column E, you compute the forecast for next month, and in Column G, you compute the squared error for each month. Finally, you'll use Solver to choose smoothing constant values that minimize the sum of the squared errors. To enact this process, perform the following steps:

- %4 In H11:J11, enter trial values (between 0 and 1) for the smoothing constants.
- %4 In C26:C113, compute the updated series level with Equation 1 by copying the formula $=\text{alp}*(\text{B26}/\text{H14})+(1-\text{alp})*(\text{C25}*\text{D25})$ from cell C26 to C27:C113.
- %4 In D26:D113, use Equation 2 to update the series trend. Copy the formula $=\text{bet}*(\text{C26}/\text{C25})+(1-\text{bet})*\text{D25}$ cell from D26 to D27:D113.

- %4 In H26:H113, use Equation 3 to update the seasonal indices. Copy the formula $=\text{gam}*(B26/C26)+(1-\text{gam})*H14$ from cell H26 to H27:H113.
- %4 In E26:E113, use Equation 4 to compute the forecast for the current month by copying the formula $=(C25*D25)*H14$ from cell E26 to E27:E113.
- %4 In F26:F113 compute each month's error by copying the formula $=(B26-E26)$ from cell E26 to E27:E113.
- %4 In G26:G113, compute the squared error for each month by copying the formula $=F26^2$ from cell F26 to F27:F113. In cell G21 compute the Sum of Squared Errors (SSE) using the formula $=\text{SUM}(G26:G113)$
- %4 Now use the Solver to determine smoothing parameter values that minimize SSE. The Solver Parameters dialog box is shown in Figure 14-3.



Figure 14-3: Solver Window for optimizing smoothing constants

- %4 Choose the smoothing parameters (H11:J11) to minimize SSE (cell G21). The Excel Solver ensures you can find the best combination of smoothing constants. Smoothing constants must be α . The Solver finds that $\text{alp} = 0.55$, $\text{bet} = 0.05$, and $\text{gamma} = 0.59$.

Forecasting Future Months

Now that you have estimated the Winter's Method smoothing constants (α , β , γ , etc.), you are ready to use these estimates to forecast future airline miles. This can be accomplished using the formula in cell D116. Copying this formula down to cells D117:D123 enables you to forecast sales for the months of May through December of 2012. Figure 14-4 offers a visual summary of the forecasted sales.

	A	B	C	D	E	F	G	H
101	Apr2011	38.51	38.09331	0.999196	38.63321	-0.13	0.0162	1.011448481
102	May2011	40.43	38.7357	1.000064	39.16698	1.26	1.5942	1.03767775
103	Jun2011	42.57	38.84163	1.000195	42.36397	0.21	0.0425	1.095009094
104	Jul2011	45.07	39.00168	1.000388	44.75388	0.32	0.1025	1.154171797
105	Aug2011	42.78	38.66139	0.99994	43.50483	-0.72	0.5220	1.110060016
106	Sep2011	36.70	39.65425	1.001205	35.05387	1.65	2.7064	0.917771587
107	Oct2011	38.70	39.72586	1.001235	38.66143	0.04	0.0018	0.974072493
108	Nov2011	36.83	40.16433	1.001716	36.182	0.65	0.4171	0.913943328
109	Dec2011	37.49	39.82658	1.001219	38.1972	-0.70	0.4955	0.944695094
110	Jan2012	34.31	39.32558	1.000541	35.19792	-0.88	0.7821	0.876725748
111	Feb2012	33.26	39.53299	1.000773	32.97971	0.28	0.0809	0.840092107
112	Mar2012	40.78	38.94334	1.000002	41.9811	-1.20	1.4396	1.05291499
113	Apr2012	38.81	38.6274	0.999604	39.38927	-0.58	0.3396	1.00743814
114			Base	Trend	Forecast	Error	Sq Error	Seasonal Indices
115			forecasts					
116		1	May-12	40.06691				
117		2	Jun-12	42.26383				
118		3	Jul-12	44.52966				
119		4	Aug-12	42.81079				
120		5	Sep-12	35.38093				
121		6	Oct-12	37.53649				
122		7	Nov-12	35.20542				
123		8	Dec-12	36.37556				
124			total	314.1696				

Figure 14-4: Forecasting with Winter's Method

Figure 14-4 shows the forecasted sales for May through December 2012 by copying the formula `=(C$113*$D$113^B116)*H102` from cell D116 to D117:D123. Cell D124 adds up these forecasts and predicts the rest of 2012 to see 314.17 billion airline miles traveled.

Cell G22 computes the standard deviation (0.94 billion) of the one-month-ahead forecast errors. This implies that approximately 95 percent of the forecast errors should be at most 1.88 billion. From Column F you see none of the one-month-ahead forecasts are outliers.

Mean Absolute Percentage Error (MAPE)

Statisticians like to estimate parameters for a forecast model by minimizing squared errors. In reality, however, most people are more interested in measuring forecast accuracy by looking at the Mean of Absolute Percentage Error (MAPE). This is probably because MAPE, unlike SSE, is measured in the same units as the data. Figure 14-5 shows that the one-month-ahead forecasts are off by an average of 2.1 percent. To compute the Absolute Percentage Error (APE) for each month, copy the formula =ABS(B26-E26)/B26 from cell G26 to J26:J113. In cell J24 the formula =AVERAGE(J26:J113) computes the MAPE.

	C	D	E	F	G	H	I	J	K	L
22				stdeverrors	0.9369659	0.879643964			34 sign changes of 87	
23						0.977358616			MAPE	
24	Base	Trend	Forecast	Error	Sq Error	0.918146041		34	0.0206	
25	40.34378	1.006491				0.969608616	Sign change		APE	
26	40.27083	1.006083	36.6689	-0.55	0.3038	0.899411078			0.0153	
27	39.93414	1.005373	35.4897	-0.93	0.8628	0.869764432		0	0.0269	
28	40.8425	1.006227	42.3093	1.33	1.7767	1.062490091		1	0.0305	
29	40.45404	1.005453	41.42583	-1.18	1.3953	1.000244124		1	0.0294	
30	41.02748	1.005882	41.15077	0.65	0.4235	1.015922148		1	0.0156	
31	40.91303	1.005456	45.39013	-0.71	0.5089	1.095230183		1	0.016	
32	40.80036	1.005052	48.28184	-0.72	0.5166	1.169022868		0	0.0151	
33	40.43957	1.00437	46.30243	-1.17	1.3621	1.121476998		0	0.0259	
34	41.43753	1.005368	35.72786	1.32	1.7346	0.888092908		1	0.0356	
35	40.61218	1.004125	40.71671	-1.87	3.4855	0.965138667		1	0.0481	
36	41.20776	1.004643	37.44173	0.72	0.5134	0.922768159		1	0.0188	
37	40.85333	1.003992	40.14091	-0.96	0.9307	0.963331236		1	0.0246	
38	40.88625	1.003835	36.89062	-0.21	0.0456	0.898023391		0	0.0058	
39	40.44253	1.003114	35.6978	-0.95	0.9068	0.863505332		0	0.0274	
40	40.45961	1.002981	43.10358	-0.21	0.0445	1.061104844		0	0.0049	

Figure 14-5: Computation of MAPE

Winter's Method is an attractive forecasting method for several reasons:

- %4 Given past data, the method can easily be programmed to provide quick forecasts for thousands of products.
- %4 Winter's Method catches changes in trend or seasonality.
- %4 Smoothing methods "adapt" to the data. That is, if you underforecast you raise parameter estimates and if you overforecast you lower parameter estimates.

Summary

In this chapter you learned the following:

%4 Exponential smoothing methods update time series parameters by computing a weighted average of the estimate of the parameter from the current observation with the prior estimate of the parameter.

%4 Winter's Method is an exponential smoothing method that updates the base, trend, and seasonal indices after each equation:

$$L_t = \alpha(x_t / (s_{t-c}) + (1-\alpha)(L_{t-1} * T_{t-1})$$

$$(2) T_t = \beta(L_t / L_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma(x_t / L_t) + (1-\gamma)s_{t-c}$$

%4 Forecasts for k periods ahead at the end of period t are made with Winter's Method using Equation 4:

$$F_{t,k} = L_t * (T_t)^k s_{t+k-c}$$

Exercises

All the data for the following exercises can be found in the file Quarterly.xlsx

- %4 Use Winter's Method to forecast one-quarter-ahead revenues for Wal-Mart.
- %4 Use Winter's Method to forecast one-quarter-ahead revenues for Coca-Cola.
- %4 Use Winter's Method to forecast one-quarter-ahead revenues for Home Depot.
- %4 Use Winter's Method to forecast one-quarter-ahead revenues for Apple.
- %4 Use Winter's Method to forecast one-quarter-ahead revenues for Amazon.com.
- %4 Suppose at the end of 2007 you were predicting housing starts in Los Angeles for the years 2008 and 2009. Why do you think Winter's Method would provide better forecasts than multiple regression?

15

Using Neural Networks to Forecast Sales

Recall from Chapter 10 “Using Multiple Regression to Predict Sales,” that to use multiple regression you had to posit the form of the relationship between the dependent and independent variables. Usually you assumed that the independent variables influenced the dependent variable via a linear relationship. However, if the relationship between the independent variables and the dependent variable is highly complex, there is little chance that multiple regression can find the relationship. Neural nets are an amazing form of artificial intelligence that can capture these complex relationships. Essentially a neural network is a “black box” that searches many models (including nonlinear models involving interactions) to find a relationship involving the independent variables that best predict the dependent variable. In a neural network the independent variables are called *input cells* and the dependent variable is called an *output cell* (more than one output is OK).

This chapter shows how to use Palisade Corporation’s great Excel add-in, NeuralTools, to easily fit a neural network to data. You can download a 15-day trial version of NeuralTools at Palisade.com

Regression and Neural Nets

As in **regression**, **neural nets have a certain number** of observations (say, N). Each observation contains a value for each independent variable and dependent variable. Also similar to regression, the goal of the neural network is to make accurate predictions for the output cell or dependent variable. As you will see, the usage of neural networks is increasing rapidly because neural networks are great at finding patterns. In regression you only find a pattern if you know what to look for. For example, if $y = Ln\ x$ and you simply use x as an independent variable, you cannot predict y very well. A neural network does not need to be “told” the nature of the relationship between the independent variables and the dependent variable. If a

Part III: Forecasting

relationship or pattern exists and you provide the neural network enough data, it can find the pattern on its own by “learning” it from the data. A major advantage of neural networks over regression is that this method requires no statistical assumptions about your data. For example, unlike regression you do not assume that your errors are independent and normally distributed.

Using Neural Networks

Neural networks have been successfully applied in many situations. This section briefly describes some actual applications of neural networks.

Predicting the Stock Market

The efficient market hypothesis of financial markets states that the “past history” of a stock’s returns yields no information about the future return of the stock. The late Halbert White, formerly an economics professor at UC San Diego, examines returns on IBM to see if the market is efficient in his 1988 white paper “Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns” (see <http://goo.gl/8vG9W>). He begins by estimating a multiple regression where the dependent variable is the next day’s return on IBM stock and the five independent variables are the return on IBM during each of the last five days. This regression yielded $R^2 = .0079$, which is consistent with the efficient market hypothesis. White then estimated a neural network (containing one hidden layer) with the output cell corresponding to the next day’s return on IBM and five input cells corresponding to the last five days’ return on IBM. This neural network yielded $R^2 = .179$. This implies that the past five days of IBM returns do contain information that can be used to make predictions about tomorrow’s return on IBM. This lends support to those investors who believe that *momentum* or trends in recent returns can be used to improve predictions of future market changes.

Fidelity, a financial services corporation specializing in investment banking, also uses neural networks for predicting trends in the stock market. According to the October 9, 1993 *Economist*, Fidelity managed \$2.6 billion in assets using neural networks.

For another example of an investment firm that uses neural networks to guide their investing strategy, see the following Ward Systems Group, Inc. article “Interviews with Real Traders” describing a hedge fund that had success using neural networks: <http://www.neuroshell.com/traders.asp?task=interviews&id=15>

Driving Your Car

In 1995, researchers at Carnegie-Mellon University developed ALVINN (short for Automated Land Vehicle in a Neural Network, a neural network that can drive a car! ALVINN can tell if a car is nearby and then slow down the car based on information received via video cameras installed in the car. Using improved versions of ALVINN, in 10 years a neural network may be driving your car! Exercise 1 at the end of this chapter prompts you to look deeper into this scenario. The following article from *Discover* magazine discusses the future of computer driven cars: [http://discovermagazine.com/2011](http://discovermagazine.com/2011/apr/10-future-tech-finally-ready-self-driving-cars)

<http://discovermagazine.com/2011/apr/10-future-tech-finally-ready-self-driving-cars>

Direct Market Targeting

Although most of us would not guess it based on the number of pieces of direct/junk mail delivered to our mailboxes, marketers often put a lot of effort into determining who receives direct mail offers. Direct mail campaigns target people who are most likely to respond to the mailing. An article from SSRN found at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=370877 shows how neural networks were used to determine who a Dutch charity should target to maximize the response rate. The following independent variables were chosen to reflect the RFM method:

- %4 **Recency:** Time since last donation
- %4 **Frequency:** How many donations were made in the last five years
- %4 **Monetary Value:** How much money has been donated in the last five years

The neural network outperformed all other methods used to identify the members of the mailing list who received a donor request. When mailing to 10 percent of the mailing list as chosen by the neural network, a 70 percent response rate was achieved, compared to a 30 percent overall response rate.

Bankruptcy Prediction

In finance and accounting it is helpful and important to accurately predict whether a company will go bankrupt during the next year. Edward Altman, an NYU finance professor, developed a method called Altman's Z-statistic in 1968 to predict whether a firm will go bankrupt during the next year based on the firm's financial ratios. This method uses a version of regression called *discriminant analysis* that Chapter 39, "Classification Algorithms: Naive Bayes Classifier and Discriminant Analysis," discusses in greater

%4 Part III: Forecasting

detail. As discussed by Rick Wilson and Ramesh Sharda (“Bankruptcy Prediction Using Neural Networks,” *Decision Support Systems Journal*, 1994, pages 545–57), neural networks that use financial ratios as input cells have been more accurate than Altman’s Z in their determination of whether a firm will go bankrupt during the next year.

Analyzing Scanner Data

Ann Furr Peterson and Thomas Grucca of the University of Iowa and Bruce Klemz of the University of Nebraska used neural networks to predict catsup market share based on price, which products were on display, and recent market share in 1999. The neural network outperformed (had a higher R^2 and a lower MAPE) other statistical techniques. You can read more about the study here: <http://dl.acm.org/citation.cfm?id=846174>

Neural Networks and Elevators

On September 22, 1993, the *New York Times* reported that Otis Elevator used neural networks to direct elevators. For example, if elevator 1 is on floor 10 and going up, elevator 2 is on floor 6 and going down, and elevator 3 is on floor 2 and going up, the neural network will recommend which elevator should answer a call to go down from floor 8. This system is used, for example, in the Marriott Hotel in Times Square.

Credit Cards and Loans

Many banks (Mellon and Chase are two examples) and credit card companies use neural networks to predict (on the basis of past usage patterns) whether a credit card transaction should be disallowed. If you have ever had a stop put on your credit card when you visit another city, there is a neural network to blame. Essentially the neural network attempts to spot patterns that indicate a credit card has been stolen before the theft is reported. This often happens to the author when he visits his daughter in Los Angeles and takes her shopping at the 3rd Street Promenade Mall in Santa Monica.

AVCO Financial also uses a neural net to determine whether to lend people money. The inputs to the neural network include information used to create credit scores (such as time at same address, annual income, number of times mortgage or credit card payments are late, etc.) By utilizing the neural network to determine if a person is a good or bad risk and giving loans only to people who are predicted to be good risks, AVCO increased its loan volume by 25 percent and decreased its default rate by 20 percent!

Using NeuralTools to Predict Sales

To **demonstrate how neural networks can find patterns in data**, you can use the data in the Data worksheet of the *Neuralpriceads.xlsx* file. A subset of this data is shown in Figure 15-1. You are given weekly sales of a product, the price, and advertising amount (in hundreds of dollars). You can construct sales under the assumption that when the price is high, advertising has no effect. More specifically, you can create weekly sales with the following rules:

- %4 If Price is less than or equal to \$8, then Sales = $500 - 15 * \text{Price} + 0.1 * \text{Advertising}$.
- %4 If Price is greater than \$8, then Sales = $500 - 15 * \text{Price}$. In this case advertising has no effect on sales.

	C	D	E
3	Sales	Price	Advertising
4	400.3	7	53
5	365	9	68
6	387.6	8	76
7	432	5	70
8	401.4	7	64
9	387.8	8	78
10	431	5	60
11	400.1	7	51
12	404.7	7	97
13	388.8	8	88
14	388.9	8	89
15	415.2	6	52
16	416.5	6	65
17	365	9	51

Figure 15-1: Price and advertising neural networks example

To get started, perform the following steps:

- %4 Run a multiple linear regression to predict Sales from Price and Advertising. The result is in the regression worksheet, as shown in Figure 15-2. The regression has a high R^2 and a standard error of 2.03 units. You soon see that a neural network can find much better forecasts with a standard deviation of forecast errors under .03!
- %4 From the Start Menu select Programs and then choose Palisade Decision Tools. Click NeuralTools; you'll see the toolbar, as shown in Figure 15-3.

%4 Part III: Forecasting

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.997413					
5	R Square	0.994833					
6	Adjusted R Square	0.994802					
7	Standard Error	2.034422					
8	Observations	332					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	262191.2	131095.6	31674.24	0	
13	Residual	329	1361.689	4.138871			
14	Total	331	263552.9				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	513.1581	0.812466	631.6054	0	511.5598	514.7564
18	Price	-16.7327	0.067173	-249.099	0	-16.8648	-16.6005
19	Advertising	0.064978	0.007716	8.421615	1.16E-15	0.0498	0.080156

Figure 15-2: Regression model to predict sales

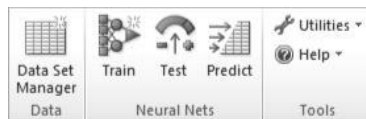


Figure 15-3: Neural Nets toolbar

%4 Select the data that will be analyzed. In this case select the data in the range C3:E335, and NeuralTools can automatically recognize the data. Columns of data may be classified as Independent or Dependent Variables and as Categorical (such as subscriber or nonsubscriber) or Numeric.

%4 After selecting the data click on the NeuralTools tab on the Ribbon. Then select the Data Set Manager. The Data Set Manager enables you to specify the nature of the input variables and dependent variable (categorical or numeric.) All the variables are numeric, so fill in the Data Set Manager dialog box, as shown in Figure 15-4.

%4 Select NeuralTools again from the Ribbon, then select Train and fill in the dialog box as shown in Figure 15-5 and click Next.

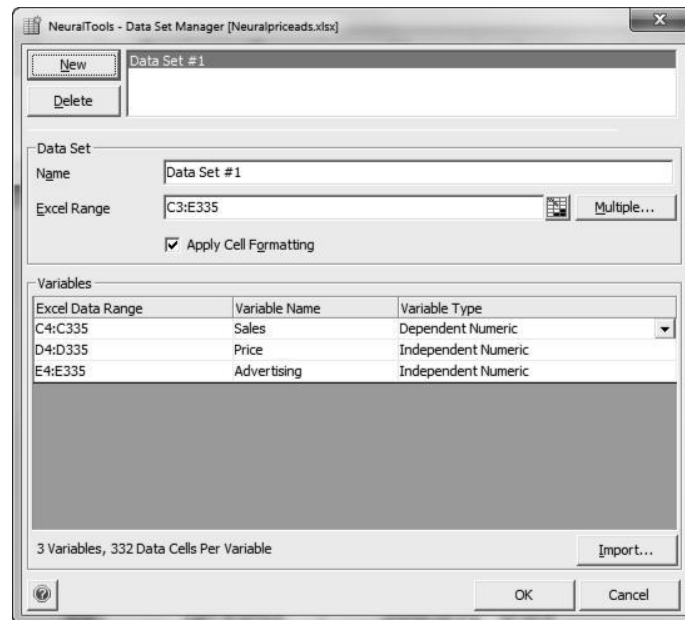


Figure 15-4: NeuralTools Data Set Manager dialog box

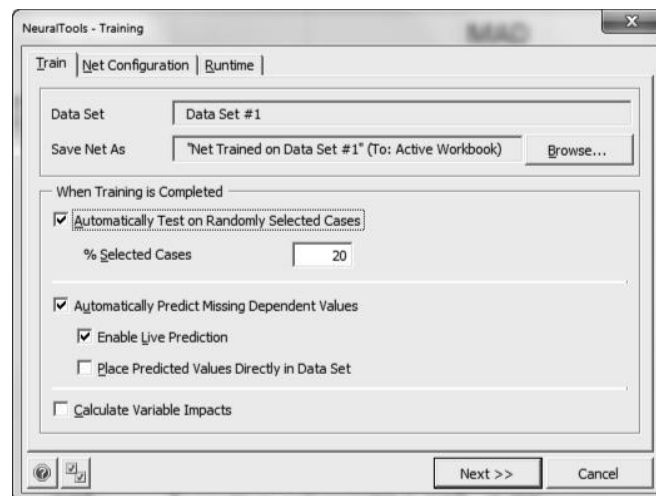


Figure 15-5: Training dialog box

%4 As also shown in Figure 15-5, check Automatically Test on Randomly Selected Cases to have NeuralTools “hold out” a randomly chosen 20 percent of the data for testing the network. This helps avoid a network

%4 Part III: Forecasting

that overfits the data used to fit the network and performs poorly on data that has not been seen by the network. Check Enable Live Prediction to place the Neural Network's predictions in the spreadsheet.

NOTE It is beyond the scope of this book to discuss how neural networks create predictions. It is sufficient to know that training a neural network is equivalent to trying many, many forms of the functional relationship between the dependent and independent variables. A neural network may seem like a “black box,” but if the analyst verifies that predictions for the testing set are nearly as accurate as predictions for the data used to fit the network, then the analyst has shown that the neural network works almost as well on data that has not yet been seen as on data used for fitting the neural network. This gives the analyst some confidence that the neural network is not simply overfitting data and will be useful for making predictions using unseen data. When estimating the accuracy of a forecast from a neural network, the analyst should use estimates of forecast accuracy based on the testing set, rather than error estimates based on the data used to fit the neural network.

%4 Select Next and NeuralTools will fit a neural network to your data. At this point Train NeuralTools uses 80 percent of the data set to train the network. A summary report sheet appears. The most important portion of this report is shown in Figure 15-6.

	B	C
15	<i>Training</i>	
16	Number of Cases	266
17	Training Time	0:00:00
18	Number of Trials	67
19	Reason Stopped	Auto-Stopped
20	% Bad Predictions (30% Tolerance)	0.0000%
21	Root Mean Square Error	0.03238
22	Mean Absolute Error	0.02137
23	Std. Deviation of Abs. Error	0.02432
24	<i>Testing</i>	
25	Number of Cases	66
26	% Bad Predictions (30% Tolerance)	0.0000%
27	Root Mean Square Error	0.05476
28	Mean Absolute Error	0.03369
29	Std. Deviation of Abs. Error	0.04317

Figure 15-6: NeuralTools report

You can find that 266 data points were used to train the network, and 67 data points were used to test the network. The observations used for the testing and training sets are randomly chosen. That is, each observation has an 80 chance of being in the training set and a 20 percent chance of being in the testing set. In the training set the mean of the absolute errors is 0.02, and in the testing set the mean absolute error is worse (0.03) but still impressive. Also note that the Root Mean Square error (which approximates the Standard Error of the Regression) for the training set is 0.03 and for the testing set is 0.05. Recall that for the multiple regression the standard error was 2.03, far inferior to the neural net.

If you now select Predict from the Neural Net toolbar and fill in the dialog box, as shown in Figure 15-7, NeuralTools places predictions in the spreadsheet, as shown in Figure 15-8. You can see Columns C and H are nearly identical, indicating that the neural net figured out the pattern in the data.

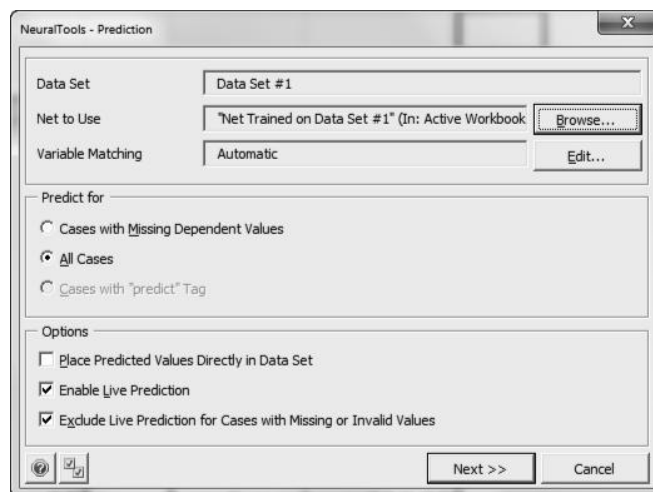


Figure 15-7: Prediction dialog box

To create forecasts for new data, you simply enter the values of Price and Advertising below the original data and copy down the formula in Column H. Figure 15-9 shows the forecasts for two new data points. Note that the forecasts are extremely accurate.

%4 Part III: Forecasting

	C	D	E	F	G	H
1						
2					Prediction Report: "Ne	
3	Sales	Price	Advertising		Tag Used	Prediction
4	400.3	7	53		predict	400.29
5	365	9	68		predict	365.00
6	387.6	8	76		predict	387.57
7	432	5	70		predict	431.94
8	401.4	7	64		predict	401.39
9	387.8	8	78		predict	387.84
10	431	5	60		predict	430.99
11	400.1	7	51		predict	400.08
12	404.7	7	97		predict	404.67
13	388.8	8	88		predict	388.83
14	388.9	8	89		predict	388.86
15	415.2	6	52		predict	415.23
16	416.5	6	65		predict	416.54

Figure 15-8: NeuralTools predictions

	C	D	E	F	G	H
3	Sales	Price	Advertising		Tag Used	Prediction
334	432.7	5	77		predict	432.66
335	385.1	8	51		predict	385.10
336	350	10	49			350.00
337	400.4	7	54			400.37

Figure 15-9: New forecasts of sales

Using NeuralTools to Forecast Air line Miles

In Chapter 14, “Winter’s Method,” you used Winter’s Method to forecast airline miles. You can also use neural networks to forecast U.S airline miles (in thousands) from the last 12 months of airline miles. Use the data from the workbook `neuralnetsairlinemiles.xlsx` to make this prediction. A subset of the data is shown in Figure 15-10.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O
7	Month	Miles	Miles-1	Miles-2	Miles-3	Miles-4	Miles-5	Miles-6	Miles-7	Miles-8	Miles-9	Miles-10	Miles-11	Miles-12
8	Jan1971	7.97E+06	7.99E+06	6.66E+06	8.02E+06	7.37E+06	7.78E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06
9	Feb1971	6.78E+06	6.66E+06	8.02E+06	7.37E+06	7.73E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06
10	Mar1971	7.71E+06	8.02E+06	7.37E+06	7.73E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06
11	Apr1971	8.48E+06	7.37E+06	7.73E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06
12	May1971	7.89E+06	7.73E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06
13	Jun1971	8.91E+06	9.18E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06
14	Jul1971	9.71E+06	9.57E+06	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06
15	Aug1971	1.01E+07	1.03E+07	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06	9.71E+06
16	Sep1971	7.85E+06	7.75E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06	9.71E+06	1.01E+07
17	Oct1971	8.10E+06	7.44E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06	9.71E+06	1.01E+07	7.85E+06
18	Nov1971	7.52E+06	7.03E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06	9.71E+06	1.01E+07	7.85E+06	8.10E+06
19	Dec1971	9.07E+06	8.47E+06	7.97E+06	6.78E+06	7.71E+06	8.48E+06	7.89E+06	8.91E+06	9.71E+06	1.01E+07	7.85E+06	8.10E+06	7.52E+06

Figure 15-10: Airline miles data

You first run a multiple linear regression to predict Column C from Columns D through O. You can see that in Column D of the regression worksheet this returns a MAD (Mean Absolute Deviation, which is the average of the absolute forecast errors) of 961,855. After running NeuralTools you find that the neural net yields a MAD of 497,000. To compute the MAD you simply average over all observations the absolute value of (Actual miles Neural Network prediction). Note that the neural net has one-half the average error of the multiple regression and yields more accurate forecasts.

Whenever you are using a neural net and you find that it yields forecasts that are not much better than a multiple regression, this tells you that the multiple linear regression is just about as good as any other functional relationship, and you are justified in modeling the relationship between the dependent and independent variables as a linear relationship.

Summary

In this chapter you learned the following:

- %4 You can use neural networks to search for a nonlinear relationship that best explains the relationship between a dependent variable and a set of independent variables.
- %4 Neural nets have been used to forecast stock prices, sales of consumer packaged goods, and even to drive a car!
- %4 The Palisades add-in, NeuralTools, makes it easy to fit a neural set to data.

- %4 If the MAD for the testing set is far larger than the MAD for the training set, then the neural net is overfitting the data and should not be used for forecasting.
- %4 If the MAD for a neural net is not much better than the MAD for a multiple linear regression, then you can be confident that the relationship between the dependent and independent variables can be accurately modeled as a linear relationship.

Exercises

- %4 **What independent variables and dependent variables would be used to train a neural network that can drive a car?** Hint: Assume the car has video cameras that can see in front of, in back of, and to the side of the car.

%4 Part III: Forecasting

The file `windchill.xlsx` gives the wind chill index (a measure of the danger from wind and temperature) for a variety of winds and temperatures. Use a neural network to predict the wind chill index based on temperature and wind speed. How much better is the neural network at predicting wind chill than a multiple linear regression?

The file `movierevenue.xls` gives revenue for several movies during weeks 1 and 2 of release as well as their total revenue. Use a neural network to predict total revenue from week 1 and week 2 revenues.

For the neural network you found in Problem 3, predict total revenue for the following two movies:

Movie 1: Week 1 revenue \$50 million; Week 2 revenue \$50 million

Movie 2: Week 1 revenue: \$80 million; Week 2 revenue \$20 million

Note that both movies made a total of \$100 million during their first two weeks of release. Why is your prediction for total revenue from Movie 2 higher than your predicted revenue for Movie 1?

%4 The file `Quarterly.xlsx` discussed in Chapter 14 gives quarterly revenues and quarterly profits for several U.S. companies. Use a neural network to predict next quarter's revenues based on the last eight quarters.

What do IV Customers Want?

Chapter 16: Conjoint Analysis

Chapter 17: Logistic Regression

Chapter 18: Discrete Choice Analysis



16

Conjoint Analysis

Often the marketing analyst is asked to determine the attributes of a product that are most (and least) important in driving a consumer's product choice. For example, when a soda drinker chooses between Coke and Pepsi, what is the relevant importance of the following:

- %4 Price
- %4 Brand (Coke or Pepsi)
- %4 Type of soda (diet or regular)

After showing a consumer several products (called *product profiles*) and asking the consumer to rank these product profiles, the analyst can use *full profile conjoint analysis* to determine the relative importance of various attributes. This chapter shows how the basic ideas behind conjoint analysis are simply an application of multiple regression as explained in Chapter 10, "Using Multiple Regression to Forecast Sales."

After understanding how to estimate a conjoint model, you will learn to use conjoint analysis to develop a market simulator, which can determine how a product's market share can change if the product's attributes are changed or if a new product is introduced into the market.

The chapter closes with a brief discussion of two other forms of conjoint analysis: *adaptive/hybrid conjoint* and *choice-based conjoint*

Products, Attributes, and Levels

Essentially, **conjoint analysis enables the marketing analyst** to determine the product characteristics that drive a consumer's preference for products. For example, in purchasing a new car what matters most: brand, price, fuel efficiency, styling, or engine power? Conjoint analysis analyzes the consumer decision process by identifying the number of product choices available; listing the main characteristics used by consumers when choosing among products; and ranking each attribute offered

Part IV: What do Customers Want?

in each product. After learning a few definitions, you will work through a detailed illustration of conjoint analysis.

The *product set* is a set of objects from which the consumer must make a choice (choosing no product is often an option). For example, a product set might be luxury sedans, laptop computers, shampoos, sodas, and so on. Conjoint analysis is also used in fields such as human resources, so product sets don't necessarily have to be consumer goods. For example, the HR analyst might want to determine what type of compensation mix (salary, bonus, stock options, vacation days, and telecommuting) is most attractive to prospective hires.

Each product is defined by the *level* of several *product attributes*. Attributes are the variables that describe the product. The levels for each attribute are the possible values of the attributes. Table 16-1 shows four examples of levels and attributes.

Table 16-1: Examples of Product Attributes and Levels

	Attribute 1	Attribute 2	Attribute 3
Blood Pressure Drug	Price: Low, Medium, High	Effi cacy: Reduces Systolic 5, 10, or 15 points	Side Effects: Percentage with serious side effects: 5%, 8%, and 11%
Plane Flight	Airline: USAir, Delta, United	Price: \$200, \$300, \$400	Number of Stops: 0, 1, or 2
Compensation for Microsoft Programmer	Starting Salary: 100K, 130K, 160K	Days of Vacation: 10, 20, 30	6-Month Sabbatical after 5 Years: Yes or No
Soda	Brand: Coke or Pepsi	Calories: 0 or 150	Price: High, Medium, or Low

The purpose of conjoint analysis is to help the marketing analyst understand the relative importance of the attributes and within each attribute the ranking of the levels. For example, a customer might rank the attributes in order of importance for price, brand, and food service. The customer might rank the brands in the order of Marriott, Hilton, and Holiday Inn. The best-known application of conjoint analysis is its use to design the Courtyard Marriott Hotel chain. This study is described in Wind, Green, Shifflet, and Scarborough's journal article, "Courtyard by Marriott: Designing a Hotel Facility with Consumer-Based Marketing Models" (*Interfaces*, 1989, pp. 25–47). This study used the following product attributes:

- %4 External décor
- %4 Room décor

- %4 Food service
- %4 Lounge facilities
- %4 General services
- %4 Leisure facilities
- %4 Security features

Other industries in which conjoint analysis has been used to ascertain how consumers value various product attributes include comparisons of the following:

- %4 Credit card offers
- %4 Health insurance plans
- %4 Automobiles
- %4 Overnight mail services (UPS versus FedEx)
- %4 Cable TV offerings
- %4 Gasoline (Shell versus Texaco)
- %4 Ulcer drugs
- %4 Blood pressure drugs
- %4 E-ZPass automated toll paying versus toll booths

The best way to learn how to use conjoint analysis is to see it in action. The following section works through a complete example that shows how conjoint analysis can give you knowledge of customer preferences.

Full Profile Conjoint Analysis

To illustrate the use of full profile **conjoint analysis**, this section uses a classic example that is described in Paul Green and Yorman Wind's article "New Way to Measure Consumers' Judgments" (*Harvard Business Review*, August 1975, pp. 107–117, <http://hbr.org/1975/07/new-way-to-measure-consumers-judgments/ar/1>). The goal of this conjoint study was to determine the role that five attributes play in influencing a consumer's preference for carpet cleaner. The five attributes (and their levels) deemed relevant to the consumer preference are as follows:

- %4 Package design (either A, B, or C)
- %4 Brand (1, 2, or 3)
- %4 Price (either \$1.19, \$1.39, or \$1.59)
- %4 Did *Good Housekeeping* magazine approve product?
- %4 Is product guaranteed?

These attributes were chosen not just because they are measurable, but because the researchers believed these attributes were likely to drive consumer product choices.

Determining the Product Profiles

In full profile conjoint analysis, the consumer is shown a set of products (called *product profiles*) and asked to rank them in order from best (#1) to worst. In the carpet clean-ing situation there are a total of $3 \times 3 \times 3 \times 2 \times 2 = 108$ possible product profiles. It seems unlikely that any consumer could rank the order of 108 product combinations; therefore, the marketing analyst must show the consumer a much smaller number of combinations. Green and Wind chose to show the consumer the 18 possible combinations shown in Figure 16-1. After being shown these combinations, the consumer ranked the product profiles (refer to Figure 16-1). Here a rank of 1 indicates the most preferred product profile and a rank of 18 the least preferred product profile. For example, the consumer felt the most preferred product was Package C, Brand 3, \$1.19 price, with a guarantee and Good Housekeeping Seal. This data is available in the Data worksheet of the *conjoint.xls* file that is available for download from the companion website.

The 18 product profiles cannot be randomly chosen. For example, if every profile with a guarantee also had *Good Housekeeping* approval, then the analyst could not determine whether the consumer preferred a profile due to a guarantee or a *Good Housekeeping* approval. Essentially, you want the attributes to be uncorrelated, so with few product profiles, a multiple regression will be less likely to be “confused” by the correlation between attributes. Such a combination of product profiles is called an *orthogonal design*. Figure 16-2 shows that the product profiles from Figure 16-1 yield an orthogonal design.

	A	B	C	D	E	F
1	Design	Brand	Price	Approved	Guarantee?	Rank
2	A	1	1.19	No	No	13
3	A	2	1.39	No	Yes	11
4	A	3	1.59	Yes	No	17
5	B	1	1.39	Yes	Yes	2
6	B	2	1.59	No	No	14
7	B	3	1.19	No	No	3
8	C	1	1.59	No	Yes	12
9	C	2	1.19	Yes	No	7
10	C	3	1.39	No	No	9
11	A	1	1.59	Yes	No	18
12	A	2	1.19	No	Yes	8
13	A	3	1.39	No	No	15
14	B	1	1.19	No	No	4
15	B	2	1.39	Yes	No	6
16	B	3	1.59	No	Yes	5
17	C	1	1.39	No	No	10
18	C	2	1.59	No	No	16
19	C	3	1.19	Yes	Yes	1

Figure 16-1: Data for conjoint example

	H	I	J	K	L	M	N	O	P	Q	R
1											
2											
3	Design	Brand	Price	Approval	Guarantee						
4	1	1	1	0	0						
5	1	2	2	1	1						
6	1	3	3	0	0						
7	2	1	2	1	1						
8	2	2	3	0	0						
9	2	3	1	0	0						
10	3	1	3	1	1						
11	3	2	1	0	0						
12	3	3	2	0	0						
13	1	1	3	0	0	Design	1				
14	1	2	1	1	1	Brand	0	1			
15	1	3	2	0	0	Price	0	0	1		
16	2	1	1	0	0	Approval	0	0	0	1	
17	2	2	2	0	0	Guarantee	0	0	0	1	1
18	2	3	3	1	1						
19	3	1	2	0	0						
20	3	2	3	0	0						
21	3	3	1	1	1						

Figure 16-2: Proof design is orthogonal.

Sidney Adelman's article "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments" (*Technometric*, 1962, Vol. 4 No. 1, pp. 36–39) is an excellent source on orthogonal designs. Table 16-2 illustrates an orthogonal design with nine product profiles and four attributes, each having three levels. For example, the first product profile sets each attribute to level 1.

Table 16-2: Example of an Orthogonal Design

Product Profile	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	1	1	1
2	1	2	2	3
3	1	3	3	1
4	2	1	2	2
5	2	2	3	1
6	2	3	1	3
7	3	1	3	3
8	3	2	1	2
9	3	3	2	1

%4 Part IV: What do Customers Want?

Running the Regression

You can determine the relative importance of product attributes by using regression with dummy variables. You begin by rescaling the consumer's rankings so that the highest ranked product combination receives a score of 18 and the lowest ranked product combination receives a score of 1. This ensures that the larger regression coefficients for product attributes correspond to more preferred attributes. Without the rescaling, the larger regression coefficients would correspond to less preferred attributes. The analysis applies the following steps to the information in the Data worksheet:

- %4 Rescale the product profile rankings by subtracting 19 from the product combination's actual ranking. This yields the rescaled rankings called the *inverse rankings*
- %4 Run a multiple linear regression using dummy variables to determine the effect of each product attribute on the inverse rankings.
- %4 This requires leaving out an arbitrarily chosen level of each attribute; for this example you can leave out Design C, Brand 3, a \$1.59 price, no Good Housekeeping Seal, and no guarantee.

After rescaling the rankings a positive coefficient for a dummy variable indicates that the given level of the attribute makes the product more preferred than the omitted level of the attribute, and a negative coefficient for a dummy variable indicates that the given level of the attribute makes the product less preferred than the omitted level of the attribute. Figure 16-3 displays the coding of the data.

	A	B	C	D	E	F	G	H	I	J	K
	A?	B?	Brand 1?	Brand 2?	1.19?	1.39?	Approved	Guarantee	Rank(1=Best)	Rank(1=worst)	Forecast
20	1	0	1	0	1	0	0	0	13	6	6.5
21	1	0	0	1	0	1	0	0	11	8	7.666667
22	1	0	0	0	0	0	1	0	17	2	1.833333
23	0	1	1	0	0	1	1	1	2	17	17.66667
24	0	1	0	1	0	0	0	0	14	5	6.333333
25	0	1	0	0	1	0	0	0	3	16	16
26	0	0	1	0	0	0	0	1	12	7	7.833333
27	0	0	0	1	1	0	1	0	7	12	12
28	0	0	0	0	0	1	0	0	9	10	9.666667
29	1	0	1	0	0	0	1	0	18	1	0.333333
30	1	0	0	1	1	0	0	1	8	11	10.5
31	1	0	0	0	0	1	0	0	15	4	5.166667
32	0	1	1	0	1	0	0	0	4	15	14.5
33	0	1	0	1	0	1	1	0	6	13	12.66667
34	0	1	0	0	0	0	0	1	5	14	12.83333
35	0	0	1	0	0	1	0	0	10	9	8.166667
36	0	0	0	1	0	0	0	0	16	3	2.833333
37	0	0	0	0	1	0	1	1	1	18	18.5

Figure 16-3: Coding of conjoint data

Notice how Row 21 indicates that if you charge \$1.19 for Brand 1 and Package design A with no guarantee or *Good Housekeeping* approval, the combination is rated 6th from

worst (or 13th overall). If you run a regression with Y-range J21:J38 and X-range A21:H38, you can obtain the equation shown in Figure 16-4 (see the Regression worksheet).

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.99153625					
5	R Square	0.983144135					
6	Adjusted R Square	0.968161144					
7	Standard Error	0.952579344					
8	Observations	18					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	8	476.3333333	59.54167	65.61735	4.49E-07	
13	Residual	9	8.166666667	0.907407			
14	Total	17	484.5				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	4.833333333	0.635052896	7.610915	3.29E-05	3.396743	6.269924
18	A?	-4.5	0.549971941	-8.18224	1.85E-05	-5.74412	-3.25588
19	B?	3.5	0.549971941	6.363961	0.000131	2.255876	4.744124
20	Brand 1?	-1.5	0.549971941	-2.72741	0.023323	-2.74412	-0.25588
21	Brand 2?	-2	0.549971941	-3.63655	0.00543	-3.24412	-0.75588
22	1.19?	7.666666667	0.549971941	13.94011	2.13E-07	6.422543	8.910791
23	1.39?	4.833333333	0.549971941	8.788327	1.04E-05	3.589209	6.077457
24	Approved?	1.5	0.476289672	3.149344	0.01175	0.422557	2.577443
25	Guarantee?	4.5	0.476289672	9.448032	5.73E-06	3.422557	5.577443

Figure 16-4: Conjoint regression data

All independent variables are significant at the .05 level. (All p-values are smaller than .05.) The R^2 value of 0.98 indicates that the attributes explain 98 percent of the variation in this consumer's ranking. As discussed in Chapter 10, "Using Multiple Regression to Forecast Sales," the standard error of 0.95 indicates that 95 percent of the predicted ranks would be accurate within 1.9. Thus it appears that the multiple linear regression model adequately captures how this consumer processes a product and creates a product ranking.

If the multiple linear regression model for a conjoint analysis does not produce a high R^2 and a low standard error, it is likely that one of the following must have occurred:

- %4 You omitted some attributes that the consumer feels are important.
- %4 The consumer's preferences are related to the current set of attributes via an interaction and nonlinear relationship. You can test for interactions or nonlinear relationships using the techniques described in Chapter 10.

The best prediction for the rescaled rank of a product is as follows:

$$\begin{aligned} \%4 \text{ Predicted Rescaled Rank} = & 4.833 - 4.5A + 3.5B - 1.5(\text{Brand 1}) - \\ & 2(\text{Brand 2}) + 7.667(\$1.19 \text{ Price}) + 4.83(\$1.39 \text{ Price}) + 1.5(\text{Approved?}) + \\ & 4.5(\text{Guarantee}) \end{aligned}$$

To interpret this equation, recall from the discussion of dummy variables in Chapter 11 that all coefficients are interpreted relative to the level of the attribute that was coded with all 0s. This observation implies the following:

- %4 Design C leads to a rank 4.5 higher than Design A and 3.5 lower than Design B.
- %4 Brand 3 leads to a rank 1.5 higher than Brand 1 and 2 higher than Brand 2.
- %4 A \$1.19 price leads to a rank 7.67 higher than \$1.59 and 2.83 higher than \$1.39.
- %4 A *Good Housekeeping* approval yields a rank 1.5 better than no approval.
- %4 A guarantee yields a rank 4.5 higher than no guarantee.

By array entering the formula `=TREND(J21:J38,A21:H38,A21:H38)` in the range K21:K38, you create the predicted inverse rank for each product profile. The close agreement between Columns J and K shows how well a simple multiple linear regression explains this consumer's product profile rankings.

Ranking the Attributes and Levels

Which attributes have the most influence on the customer's likelihood to purchase the product? To rank the importance of the attributes, order the attributes based on the attributes' spread from the best level of the attribute to the worst level of the attribute. Table 16-3 displays this ranking.

Table 16-3: Product Attribute Rankings

Attribute	Spread	Ranking
Design	$4.5 - (-3.5) = 8$	1 st
Brand	$0 - (-2) = 2$	4 th
Price	$7.67 - 0 = 7.67$	2 nd
Approval	$1.5 - 0 = 1.5$	5 th
Guarantee	$4.5 - 0 = 4.5$	3 rd

For example, you can see the package design is the most important attribute and the *Good Housekeeping* approval is the least important attribute. You can also rank the levels within each attribute. This customer ranks levels as follows:

- %4 Design: B, C, A
- %4 Price: \$1.19, \$1.39, \$1.59

- %4 Brand: Brand 3, Brand 1, Brand 2
- %4 Guarantee: Guarantee, No Guarantee
- %4 Approval: Approval, No Approval

Within each attribute, you can rank the levels from most preferred to least preferred. In this customer's example, Brand 3 is most preferred, Brand 1 is second most preferred, and Brand 2 is least preferred.

Using Conjoint Analysis to Segment the Market

You can also use conjoint analysis to segment the market. To do so, simply determine the regression equation described previously for 100 representative customers. Let each row of your spreadsheet be the weights each customer gives to each attribute in the regression. Thus for the customer analyzed in the preceding section, his row would be (-4.5, 3.5, -1.5, -2, 7.67, 4.83, 1.5, 4.5). Then perform a *cluster analysis* (see Chapter 23, "Cluster Analysis") to segment customers. This customer, for example, would be representative of a brand-insensitive, price-elastic decision maker who thought package design was critical.

Value-Based Pricing with Conjoint Analysis

Many companies price their products using *cost plus pricing*. For example, a cereal manufacturer may mark up the cost of producing a box of cereal by 20 percent when selling to the retailer. Cost plus pricing is prevalent for several reasons:

- %4 It is simple to implement, even if your company sells many different products.
- %4 Product margins are maintained at past levels, thereby reducing shareholder criticism that "margins are eroding."
- %4 Price increases can be justified if costs increase.

Unfortunately, cost plus pricing ignores the voice of the customer. Conjoint analysis, on the other hand, can be used to determine the price that can be charged for a product feature based on the value consumers attach to that feature. For example, how much is a Pentium 4 worth versus a Pentium 2? How much can a hotel charge for high-speed wireless Internet access? The key idea is that Equation 1 enables you to impute a monetary value for each level of a product attribute. To illustrate how conjoint analysis can aid in value-based pricing, assume the preferences of all customers in the carpet cleaning example are captured by Equation 1. Suppose the carpet cleaning fluid without guarantee currently sells for \$1.39. What can you

%4 Part IV: What do Customers Want?

charge if you add a guarantee? You can implement value-based pricing by keeping the product with a guarantee at the same value as a product without a guarantee at the current price of \$1.39.

The regression implies the following:

\$1.19 has a score of 7.67.

\$1.39 has a score of 4.83.

\$1.59 has a score of 0.

Thus, increasing the product price (in the range \$1.39 to \$1.59) by 1 cent costs you $4.83/20 = 0.24$ points.

Because a guarantee increases ranking by 4.5 points, value-based pricing says that with a guarantee you should increase price by x cents where $0.24x = 4.5$ or $x = 4.5/0.24 = 19$ cents

You should check that if the product price is increased by 19 cents (to \$1.58) and the product is guaranteed, the customer will be as happy as she was before. This shows that you have priced the guarantee according to the customer's imputed value. Although this approach to value-based pricing is widely used, it does not ensure profit maximization. When you study discrete choice in Chapter 18, "Discrete Choice Analysis," you will see how choice-based conjoint analysis can be used to incorporate consumer preferences into profit maximization. Of course, this analysis assumes that for a price between \$1.39 and \$1.59 the effect of price on product ranking is a linear function of price. If this is not the case then Price^2 should have been added to the regression as an independent variable. This would allow the regression to capture the nonlinear effect of price on the consumer's product rankings.

Using Evolutionary Solver to Generate Product Profiles

The **leading statistical packages** (SAS and SPSS) enable a user to input a desired number of product profiles, the number of attributes, and the number of levels, and output (if one exists) an orthogonal design. In many situations the user must exclude product profiles that are unreasonable. For example, in analyzing an automobile it is unreasonable to set the car's size to a Mini Cooper and assume the car can carry six passengers. This section shows how to use the Excel 2010 or 2013 Evolutionary Solver to easily design something that is close to being orthogonal but that excludes

unreasonable product profiles. The work here is based on Steckel, DeSarbo, and Mahajan's study "On the Creation of Acceptable Conjoint Analysis Experimental Designs" (*Decision Sciences*, 1991, pp. 436–42, <http://goo.gl/2PO0J>) and can be found in the `evconjoint.xlsx` file. If desired, the reader may omit this advanced section without loss of continuity.

Assume that you want to use a conjoint analysis to evaluate how consumers value the following attributes for a new car:

- %4 Miles per gallon (MPG): 20 or 40
- %4 Maximum speed in miles per hour (MPH): 100 or 150
- %4 Length: 12 feet or 14 feet
- %4 Price: \$25,000 or \$40,000
- %4 Passenger capacity: 4, 5, or 6

Now suppose the analyst wants to create 12 product profiles and exclude the following types of product profiles as infeasible:

- %4 40 MPG, 6 passengers, 14-foot car
- %4 14-foot car, 150 MPH, and \$20,000 price
- %4 40 MPG and 150 MPH

You can now see how to design 12 product profiles that include each level of an attribute an equal number of times, which are close to orthogonal and exclude infeasible product profiles.

- %4 To begin, list in B3:G28 each attribute and level the number of times you want the combination to occur in the profiles. For example, Figure 16-5 lists 20 MPG six times and four passengers four times.
- %4 Give a range name to each listing of attribute values. For the example file the range D16:E28 is named `lookpassengers`
- %4 The key to the model is to determine in cells K5:O16 how to "scramble" (as shown in Figure 16-6) for each attribute the integers 1 through 12 to select for each product profile the level for each attribute.

For example, the 7 in K5 in `evconjoint.xlsx` (to be selected by Solver) means "look up the attribute level for MPG in the 7th row of B4:C15." This yields 40 MPG in Q5. The formulas in Q5:U16 use a little known Excel function: the `INDIRECT` function. With this function a cell reference in an Excel formula after the word `INDIRECT` tells Excel to replace the cell reference with the contents of the cell. (If the contents of the

%4 Part IV: What do Customers Want?

cell are a range name, Excel knows to use the named range in the formula.) Therefore, copying the formula =VLOOKUP(K5,INDIRECT(Q\$2),2, FALSE) from Q5:U16 translates the values 1–12 in K5:O16 into actual values of the attributes.

	B	C	D	E	F	G
3		MPG		MPH		Length
4	1	20	1	100	1	12
5	2	20	2	100	2	12
6	3	20	3	100	3	12
7	4	20	4	100	4	12
8	5	20	5	100	5	12
9	6	20	6	100	6	12
10	7	40	7	150	7	14
11	8	40	8	150	8	14
12	9	40	9	150	9	14
13	10	40	10	150	10	14
14	11	40	11	150	11	14
15	12	40	12	150	12	14
16		Price		Passengers		
17	1	25	1	4		
18	2	25	2	4		
19	3	25	3	4		
20	4	25	4	4		
21	5	25	5	5		
22	6	25	6	5		
23	7	30	7	5		
24	8	30	8	5		
25	9	30	9	6		
26	10	30	10	6		
27	11	30	11	6		
28	12	30	12	6		

Figure 16-5: Listing attribute-level combinations

	K	L	M	N	O	P	Q	R	S	T	U
3	Coded						Actual				
4	MPG	MPH	Length	Price	Passengers		MPG	MPH	Length	Price	Passenger
5	7	1	8	4	6		40	100	14	25	5
6	2	10	9	12	9		20	150	14	30	6
7	1	12	2	10	7		20	150	12	30	5
8	3	7	4	5	2		20	150	12	25	4
9	9	3	11	1	1		40	100	14	25	4
10	5	11	10	2	11		20	150	14	25	6
11	6	9	7	7	3		20	150	14	30	4
12	10	5	6	11	4		40	100	12	30	4
13	4	8	3	6	5		20	150	12	25	5
14	12	6	12	8	8		40	100	14	30	5
15	8	2	1	3	12		40	100	12	25	6
16	11	4	5	9	10		40	100	12	30	6

Figure 16-6: Selecting attribute levels

%4 Next, in the cell range V5:W16 (see Figure 16-7), you can determine if a product profile represents an infeasible combination. Copy the formula =COUNTIFS(Q5:40,U5:6,S5,14) from V5 to V6:V16. This determines if the product profile is infeasible due to getting 40 MPG in a six-passenger 14-foot car. In a similar fashion Columns W and X yield a 1 if a product profile is infeasible for being either 14-foot, 150 MPH, and \$20,000 price or 40 MPG with 150 MPH.

	Q	R	S	T	U	V	W	X
3	Actual					Excluded	Excluded	Excluded
4	MPG	MPH	Length	Price	Passenger	40 MPG 6 PASS 14 FT	14 ft 150 MPH \$20	40 MPG 150 MPH
5	40	100	14	25	5	0	0	0
6	20	150	14	30	6	0	0	0
7	20	150	12	30	5	0	0	0
8	20	150	12	25	4	0	0	0
9	40	100	14	25	4	0	0	0
10	20	150	14	25	6	0	0	0
11	20	150	14	30	4	0	0	0
12	40	100	12	30	4	0	0	0
13	20	150	12	25	5	0	0	0
14	40	100	14	30	5	0	0	0
15	40	100	12	25	6	0	0	0
16	40	100	12	30	6	0	0	0

Figure 16-7: Excluding infeasible product profiles

%4 Determine the correlations between each pair of attributes. A slick use of the INDIRECT function makes this easy: apply Create from Selection to the range Q4:U16 names Q5:Q16 MPG, R5:R16 MPH, and so on. Then, as shown in Figure 16-8, compute the absolute value of the correlation between each pair of attributes by copying the formula =ABS(CORREL(INDIRECT(O\$19),INDIRECT(\$N20))) from O20 to O20:S24.

	L	M	N	O	P	Q	R	S	T
19				MPG	MPH	Length	Price	Passengers	
20			MPG	1	1	0	0	0	
21			MPH	1	1	0	0	0	
22			Length	0	0	1	0	0	
23			Price	0	0	0	1	0	
24			Passengers	0	0	0	0	1	
25									
26		avgcorr	penalties	target					
27		0.1	0	0.1					

Figure 16-8: Correlations and target cell

%4 Now set up a target cell for Solver to minimize the sum of the average of the nondiagonal correlations (all diagonal correlations are 1) added to the number of infeasible product profiles. In cell M27 the formula =(SUM(O20:S24)-5)/20

calculates the average absolute value of the nondiagonal correlations. The target cell tries to minimize this average. This minimization moves you closer to an orthogonal design. Then determine the number of infeasible product profiles used in cell N27 with the formula `=SUM(V5:X16)`. The final target cell is computed in cell O27 with the formula `=N27+M27`.

%4 Finally, you are ready to invoke Solver! Figure 16-9 shows the Solver window. Simply change K5:O16 and invoke the AllDifferent option for each column. This ensures that columns K through O will contain the integers 1–12 once. In general, if you constrain a range of n cells to be AllDifferent, Excel ensures that these cells always assume the integer values 1, 2, ..., n and each value will be used once. From the way you set up the formulas in Columns Q through U, this ensures that each level of all attributes appear the wanted number of times. After going to Evolutionary Solver options and resetting the Mutation rate to .5, you obtain the solution shown in Figures 16-7 and 16-8. Note that no infeasible product profiles occur, and all correlations except for MPH and MPG are 0. MPH and MPG have a correlation of -1 because the infeasible profiles force a low MPG to be associated with a large MPH, and vice versa.

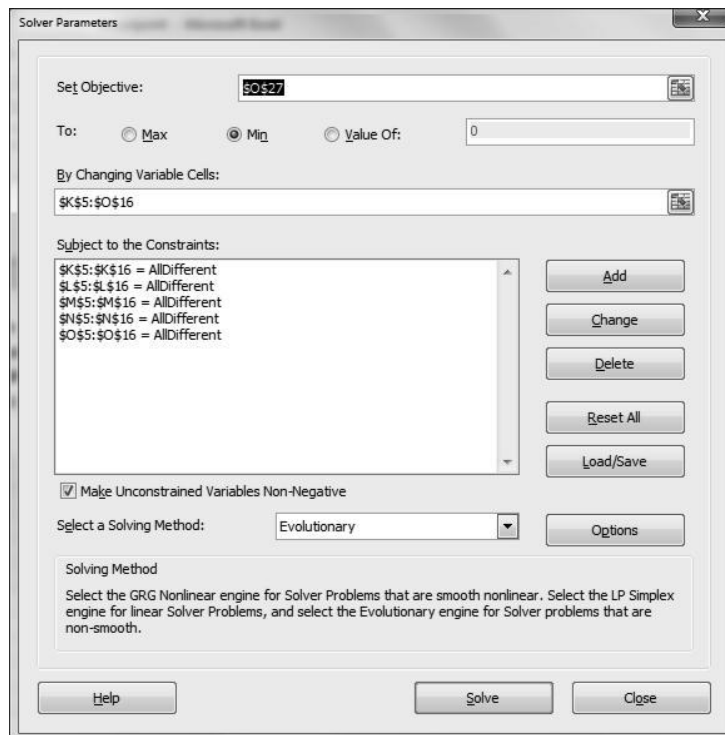


Figure 16-9: Solver window for selecting product profiles

Developing a Conjoint Simulator

Conjoint analysis is often used to predict how the introduction of a new product (or changes in existing products) result in changes in product market share. These insights require the marketing analyst to develop a *conjoint simulator* that predicts how changes in product attributes change market share. This section illustrates how to calibrate a conjoint simulator to current share data and how to make predictions about how changes in products or the introduction of a new product can change market share. The work for this section is in the Segments worksheet of the *conjoint.xls* file and is shown in Figure 16-10.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Current share			A	B	Brand	Brand	1.19	1.39	Approved?	Guarantee	Alpha	2.07561	Squared Error				SSE
2	0.3	Our brand		1	0	1	0	1	0	0	0			Actual	0.00193	0.00035	0.00063	0.00291
3	0.5	Comp 1		0	1	0	1	0	1	1	0			Predicted	0.3	0.5	0.2	
4	0.2	Comp 2		0	0	0	0	0	0	0	1			Share	0.3439	0.48124	0.17486	
5	Segment	Size(000)	Intercept	A	B	Brand	Brand	1.19	1.39	Approved?	Guarantee	Score	Comp 1 score	Comp 2 score	Our share	Comp 1 share	Comp 2 share	
6	1	10	4.83	-4.5	3.5	-1.5	-2	7.7	4.83	1.5	4.5	6.5	12.66	9.33	0.14071	0.56136	0.29793	
7	2	15	2	-6	5	1	2	9	6	2	3	6	17	5	0.09643	0.83753	0.06605	
8	3	20	5	-2	2	-4	-6	8	5	3	5	7	9	10	0.20914	0.35236	0.43849	
9	4	12	2	4	1	4	2	7	4	1	4	17	10	6	0.69882	0.22964	0.07954	
10	5	22	4.7	1	4	3	2	6	4	2	2	14.7	16.7	6.7	0.40016	0.52148	0.07834	
11	6	14	9	-6	-4	2	1	5	3	2	3	10	11	12	0.27183	0.3313	0.39687	
12	7	10	8	1	3	0	-2	4	3	1	2	13	13	10	0.38758	0.38758	0.22483	
13	8	15	2	4	-3	5	-3	9	7	2	1	20	5	3	0.92956	0.05232	0.01812	
14	9	8	2	-2	4	2	5	7	4	3	4	9	18	6	0.17711	0.74655	0.07634	
15	10	12	7	-4	-2	1	3	6	4	2	2	10	14	9	0.26219	0.52713	0.21069	
16	11	9	2	1	2	2	4	8	5	4	3	13	17	5	0.34689	0.68537	0.04774	
17	12	11	4.4	-6	5	-4	-2	6	4	5	2	0.4	16.4	6.4	0.00039	0.87544	0.12416	
18	13	8	7	-4	3	-1	3	4	3	2	3	6	18	10	0.07317	0.71557	0.21126	
19	14	14	2	2	2	2	-1	8	5	3	4	14	11	6	0.56228	0.34085	0.09687	

Figure 16-10: Conjoint simulator

Suppose you're using cluster analysis (as shown in Chapter 23) and through a conjoint analysis the analyst has identified 14 market segments for a carpet cleaner. The regression equation for a typical member of each segment is given in the cell range D6:K19 and the size of each segment (in thousands) is the cell range B6:B19. For example, Segment 1 consists of 10,000 people and has Equation 1 as representative of consumer preferences for Segment 1. The range D2:K4 describes the product profile associated with the three products currently in the market. The brand currently has a 30 percent market share, Comp 1 has a 50 percent market share, and Comp 2 has a 20 percent market share.

To create a market simulator, you need to take the score that each segment associates with each product and translate these scores for each segment to a market share for each product. You can then calibrate this rule, so each product gets the observed market share. Finally, you can change attributes of the current products or introduce new products and "simulate" the new market shares for each product.

Consider Segment 1. Given that there are %4 products, you can assume fraction of people in Segment 1 that will purchase Product i is given by following:

$$\frac{(Product\ i\ Score\ for\ Segment\ 1)^\alpha}{\sum_{k=1}^K (Product\ k\ Score\ for\ Segment\ 1)^\alpha}$$

When you choose the value of α , ensure that the predicted market share over all segments for each product matches as closely as possible the observed market share. To complete this process proceed as follows:

- %4 Copy the formula =SUMPRODUCT(\$D\$2:\$K\$2,D6:K6)+C6 from L6 to L7:L19 to compute each segment's score for the product. Similar formulas in Columns M and N compute each segment's score for Comp 1 and Comp 2.
- %4 Copy the formula =L6^Alpha/(\$L6^Alpha+\$M6^Alpha+\$N6^Alpha) from O6 to the range O6:Q19 to compute for each segment the predicted market share for each product.
- %4 Copy the formula =SUMPRODUCT(O6:O19,\$B\$6:\$B\$19)/SUM(\$B\$6:\$B\$19) from O4 to P4:Q4 to compute the predicted market share for each product by accounting for the size of each segment.
- %4 Copy the formula =(O4-O3)^2 from O2 to P2:Q2 to compute (based on the trial value of α) the squared error in trying to match the actual market share for each product.
- %4 Use the Solver window, as shown in Figure 16-11, to yield the value of α (2.08) that “calibrates” the simulator to actual market share data.

Suppose you consider adding (with no price increase) a guarantee and believe you can obtain *Good Housekeeping* approval. In the worksheet *GuaranteeApp* you can change J2 and K2 to 1, and see that the conjoint simulator predicts that the market share will increase from 34 percent to 52 percent.

As an illustration of the use of a conjoint simulator, consider the problem of estimating the fraction of motorists who would use E-ZPass (an electronic method to collect tolls). Before New York and New Jersey implemented E-ZPass a conjoint simulator was used to choose the levels of the following attributes:

- %4 Lanes available
- %4 Method used to acquire E-ZPass
- %4 Toll prices
- %4 Other uses of E-ZPass

The conjoint simulator predicted that 49 percent of motorists would use E-ZPass. In reality, after seven years 44 percent of all motorists used E-ZPass.



Figure 16-11: Solver window to calibrate conjoint simulator

Examining Other Forms of Conjoint Analysis

The **approach to conjoint analysis** thus far in this chapter involves using a set of a few (in the chapter example, five) product attributes to generate product profiles (in the chapter example, 18) which are ranked by a group of customers. Then multiple regression is used to rank the importance of the product attributes and rank the levels of each attribute. This approach to conjoint analysis is called *full profile conjoint*. The two shortcomings of full profile conjoint are as follows:

- %4 Full profile conjoint has difficulty dealing with many attributes because the number of profiles that must be ranked grows rapidly with the number of attributes.
- %4 Consumers have difficulty ranking product profiles. Rather than having a consumer rank product profiles, it is much easier for a consumer to choose the best available option.

Part IV: What do Customers Want?

This section provides a brief description of two alternate forms of conjoint analysis: adaptive/hybrid conjoint analysis and choice-based conjoint analysis, which attempt to resolve the problems with full profile conjoint analysis.

Adaptive/Hybrid Conjoint Analysis

If a product has many attributes, the number of product profiles needed to analyze the relative importance of the attributes and the desirability of attribute levels may be so large that consumers cannot accurately rank the product profiles. In such situations *adaptive* or *hybrid conjoint analysis* (developed by Sawtooth Software in 1985) may be used to simplify the consumer's task.

In step 1 of an adaptive conjoint analysis, the consumer is asked to rank order attribute levels from best to worst. In step 2 the consumer is asked to evaluate the relative desirability of different attribute levels. Based on the consumer's responses in steps 1 and 2 (this is why the method is called adaptive conjoint analysis), the consumer is asked to rate on a 1–9 scale the strength of his preference for one product profile over another. If, for example, a consumer preferred Design A to Design B and Brand 1 to Brand 2, the adaptive conjoint software would never create a paired comparison between a product with Design A and Brand 1 to Design B and Brand 2 (all other attributes being equal).

Choice-Based Conjoint Analysis

In *choice based conjoint analysis*, the consumer is shown several product profiles and is not asked to rank them but simply to state which profile (or none of the available choices) she would choose. This makes the consumer's task easier, but an understanding of choice-based conjoint analysis requires much more mathematical sophistication than ordinary conjoint analysis. Also, choice-based conjoint analysis cannot handle situations in which attributes such as price or miles per gallon for a car are allowed to assume a range of values. In Chapter 18 you will learn about the theory of *discrete choice*, which is mathematically more difficult than multiple linear regression. Discrete choice generalizes choice-based conjoint analysis by allowing product attributes to assume a range of values.

Summary

In this chapter you learned the following:

- %4 Conjoint analysis is used to determine the importance of various product attributes and which levels of the attributes are preferred by the customer.
- %4 In full profile conjoint analysis, the consumer is asked to rank a variety of product profiles.
- %4 To make the estimates of each attribute's importance and level preferences more accurate, it is usually preferable to make the correlation between any pair of attributes in the product profiles close to 0.
- %4 Multiple linear regression (often using dummy variables) can easily be used to rank the importance of attributes and the ranking of levels within each attribute.
- %4 A conjoint simulator can predict how changes in existing products (or introduction of new products) will change product market shares.
- %4 Adaptive/hybrid conjoint analysis is often used when there are many product attributes. After having the consumer answer questions involving ranking attributes and levels, the software adapts to the consumer's preferences before asking the consumer to make paired comparisons between product profiles.
- %4 Choice-based conjoint analysis ascertains the importance of attributes and the ranking of levels within each attribute by asking the consumer to simply choose the best of several product profiles choices (usually including an option to choose none of the available product profiles.) The extension of choice-based conjoint analysis to a situation where an attribute such as price is continuous is known as discrete choice and will be covered in Chapter 18.

Exercises

- %4 **Determine** how various attributes impact the purchase of a car. There are four attributes, each with three levels:

Brand: Ford = 0, Chrysler = 1, GM = 2

MPG: 15 MPG = 0, 20 MPG = 1, 25 MPG = 2

Horsepower (HP): 100 HP = 0, 150 HP = 1, 200 HP = 2

Price: \$18,000 = 0, \$21,000 = 1, \$24,000 = 2

%4 Part IV: What do Customers Want?

The nine product profiles ranked in Figure 16-12 were evaluated by a consumer.

1	A	B	C	D	E	F
2	Trial	(F,C,GM) Brand	(15,20,25) MPG	(100,150,200) Power	(18000,21000,24000) Price	Rank
3	1	0	0	0	0	4
4	2	0	1	1	2	7
5	3	0	2	2	1	3
6	4	1	0	1	1	6
7	5	1	1	2	0	2
8	6	1	2	0	2	9
9	7	2	0	2	2	8
10	8	2	1	0	1	5
11	9	2	2	1	0	1

Figure 16-12: Auto data for Exercise 1

%4 For this market segment, rank the product attributes from most important to least important.

%4 Consider a car currently getting 20 MPG selling for \$21,000. If you could increase MPG by 1 mile, how much could you increase the price of the car and keep the car just as attractive to this market segment?

%4 Is this design orthogonal?

NOTE When you run the regression in Excel, you can obtain a #NUM error for the p-values. You still may use the Coefficients Column to answer the exercise.

%4 The soda.xlsx file (see Figure 16-13) gives a consumer's ranking on an orthogonal design with 12 product profiles involving a comparison of Coke and Pepsi. The attributes and levels are as follows:

Brand: Coke or Pepsi

Packaging: 12 oz. can or 16 oz. bottle

Price per ounce: 8 cents, 10 cents, 12 cents

Calories per ounce: 0 or 15

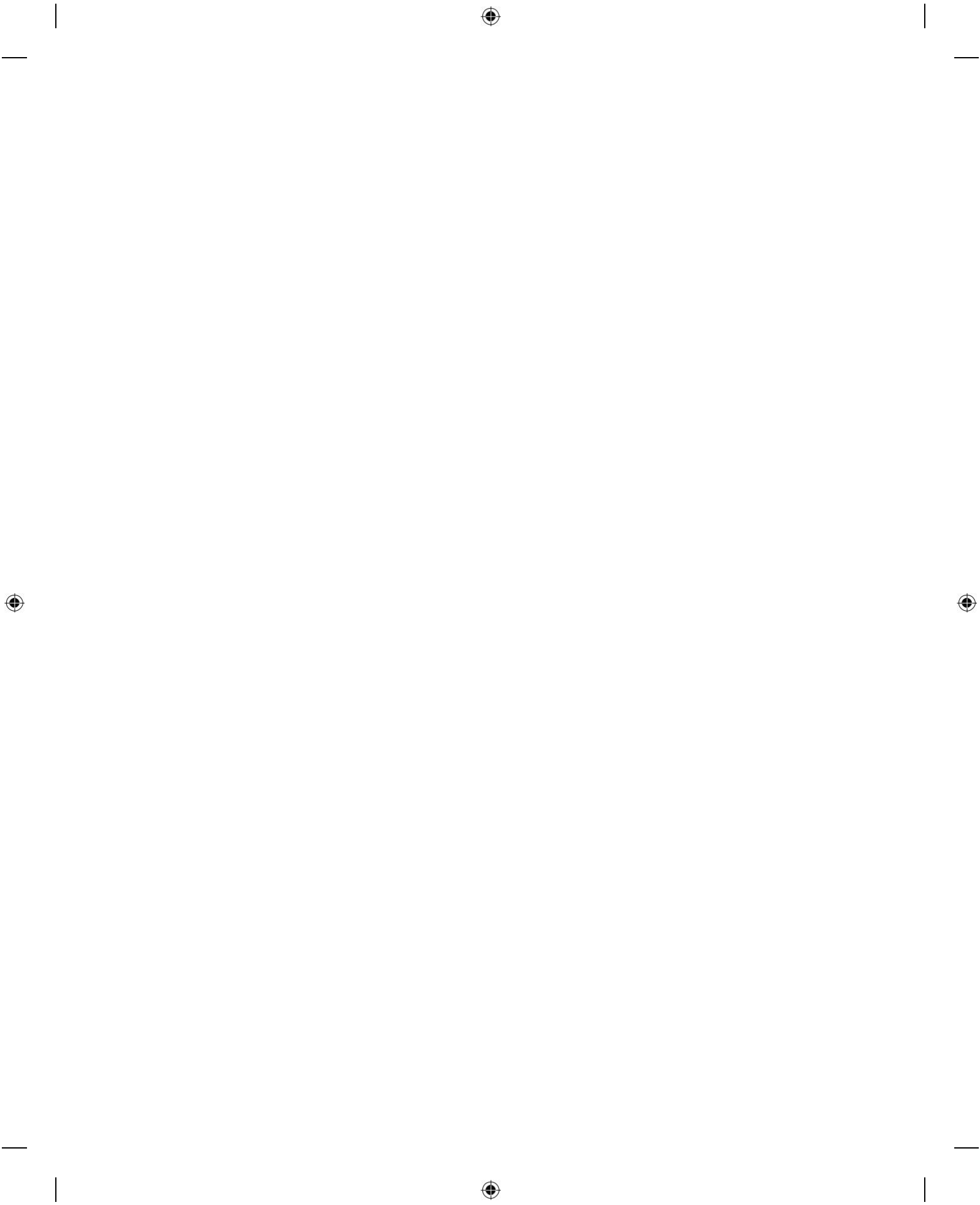
Determine the ranking of the attributes' importance and the ranking of all attribute levels.

	F	G	H	I	J	K
1						
2				per oz	per oz	
3	Profile	Brand	Packaging	Calories	Price	Rank
4		1 coke	can	0	8	4
5		2 coke	bottle	0	8	2
6		3 pepsi	can	15	8	6
7		4 pepsi	bottle	15	8	5
8		5 coke	can	15	10	11
9		6 coke	bottle	15	10	9
10		7 pepsi	can	0	10	3
11		8 pepsi	bottle	0	10	1
12		9 coke	can	15	12	12
13		10 coke	bottle	0	12	8
14		11 pepsi	can	0	12	7
15		12 pepsi	bottle	15	12	10

Figure 16-13: Soda data for Exercise 2

%4 Show that the list of product profiles in the following table yields an orthogonal design.

Product Profile	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	1	1	1
2	1	2	2	3
3	1	3	3	2
4	2	1	2	2
5	2	2	3	1
6	2	3	1	3
7	3	1	3	3
8	3	2	1	2
9	3	3	2	1



17

Logistic Regression

Many marketing problems and decisions deal with understanding or estimating the probability associated with certain events or behaviors, and frequently these events or behaviors tend to be dichotomous—that is, of one type or of another type. When this is the case, the marketing analyst must predict a binary dependent variable (one that assumes the value 0 or 1, representing the inherent dichotomy) from a set of independent variables. Some examples follow:

%4 Predicting from demographic behavior whether a person will (dependent variable = 1) or will not (dependent variable = 0) subscribe to a magazine or use a product.

%4 Predicting whether a person will (dependent variable = 1) or will not (dependent variable = 0) respond to a direct mail campaign. Often the independent variables used are recency (time since last purchase), frequency (how many orders placed in last year), and monetary value (total amount purchased in last year).

%4 Predicting whether a cell phone customer will “churn” (dependent variable = 1) by end of year and switch to another carrier. The dependent variable = 0 if you retain the customer. You will see in Chapters 19, “Calculating Lifetime Customer Value,” and 20, “Using Customer Value to Value a Business,” that a reduction in churn can greatly increase the value of a customer to the company.

In this chapter you learn how to use the widely used tool of *logistic regression* to predict a binary dependent variable. You learn the following:

%4 Why multiple linear regression is not equal to the task of predicting a binary dependent variable

%4 How the Excel Solver may be used to implement the technique of maximum likelihood to estimate a logistic regression model

%4 Part IV: What do Customers Want?

How to interpret the coefficients in a logistic regression model

How Palisade's StatTools program can easily be used to estimate a logistic regression model and test hypotheses about the individual coefficients

Why Logistic Regression Is Necessary

To explain **why you need logistic regression**, suppose you want to predict the chance (based on the person's age) that a person will subscribe to a magazine. The linear regression worksheet in the `subscribers.xlsx` file provides the age and subscription status (1 = subscriber, 0 = nonsubscriber) for 41 people. Figure 17-1 displays a subset of this data.

	D	E
6	Age	Subscribe
7	20	0
8	23	0
9	24	0
10	25	0
11	25	1
12	26	0
13	26	0
14	28	0
15	28	0
16	29	0
17	30	0
18	30	0
19	30	0
20	30	0
21	30	0
22	30	1
23	32	0
24	32	0
25	33	0
26	33	0
27	34	0
28	34	0
29	34	1
30	34	0
31	34	0

Figure 17-1: Age and subscriber status

Using this data, you can run a linear regression to predict a subscriber's status from the subscriber's age. The Y-range is E6:E47 and the X-range is E6:E47. Check the residual box because you will soon see that the residuals indicate two of the key assumptions (from Chapter 10, "Using Multiple Regression to Forecast Sales")

of regression are violated. Figure 17-2 and Figure 17-3 show the results of running this regression.

	H	I	J	K	L	M	N
6							
7							
8							
9	SUMMARY OUTPUT						
10							
11	Regression Statistics						
12	Multiple R	0.413481783					
13	R Square	0.170967185					
14	Adjusted R Square	0.149709934					
15	Standard Error	0.400905926					
16	Observations	41					
17							
18	ANOVA						
		df	SS	MS	F	gnificance F	
19	Regression	1	1.292678718	1.292678718	8.04277	0.007207	
20	Residual	39	6.268296892	0.160725561			
21	Total	40	7.56097561				
22							
23							
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
24	Intercept	-0.543019266	0.284454263	-1.908986215	0.063642	-1.11838	0.032344
25	Age	0.023194673	0.008178722	2.83597778	0.007207	0.006652	0.039738

Figure 17-2: Flawed linear regression

	H	I	J	K
	Observation	Predicted Subscribe?	Residuals	abs(Residuals)
32				
33	1	-0.07912581	0.07912581	0.07912581
34	2	-0.009541791	0.009541791	0.009541791
35	3	0.013652882	-0.013652882	0.013652882
36	4	0.036847555	-0.036847555	0.036847555
37	5	0.036847555	0.963152445	0.963152445
38	6	0.060042228	-0.060042228	0.060042228
39	7	0.060042228	-0.060042228	0.060042228
40	8	0.106431573	-0.106431573	0.106431573
41	9	0.106431573	-0.106431573	0.106431573
42	10	0.129626246	-0.129626246	0.129626246
43	11	0.152820919	-0.152820919	0.152820919
44	12	0.152820919	-0.152820919	0.152820919
45	13	0.152820919	-0.152820919	0.152820919
46	14	0.152820919	-0.152820919	0.152820919
47	15	0.152820919	-0.152820919	0.152820919
48	16	0.152820919	0.847179081	0.847179081
49	17	0.199210265	-0.199210265	0.199210265
50	18	0.199210265	-0.199210265	0.199210265
51	19	0.222404937	-0.222404937	0.222404937
52	20	0.222404937	-0.222404937	0.222404937
53	21	0.24559961	-0.24559961	0.24559961

Figure 17-3: Residuals for flawed regression

%4 Part IV: What do Customers Want?

From Figure 17-2 you can find the following equation:

$$(\text{Predicted subscriber status}) = -0.543 + .0231 * \text{Age}$$

Using the Residuals column shown in Figure 17-3, you can see three problems with this equation.

- %4 Some observations predicted a negative subscriber status. Because the sub-scriber status must be 0 or 1, this is worrisome. Particularly, if you want to predict the probability that a person is a subscriber, a negative probability does not make sense.
- %4 Recall from Chapter 10 that the residuals should indicate that the error term in a regression should be normally distributed. Figure 17-4, however, shows a histogram of the residuals in which this is not the case.

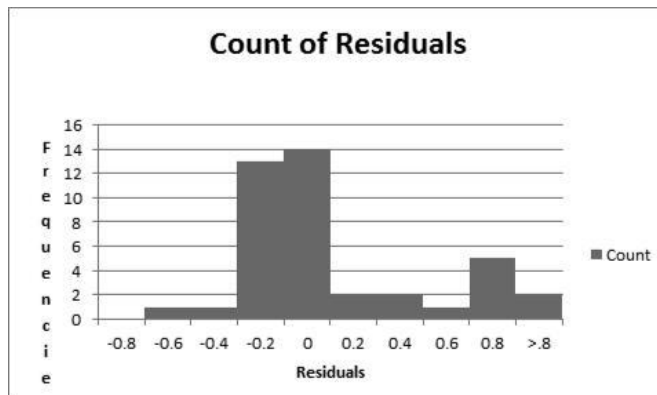


Figure 17-4: Histogram of residuals

- %4 Figure 17-5 gives the standard deviation of the residuals as a function of the predicted value of the subscriber variable. The spread of the residuals is not independent of the predicted value of the dependent variable. This implies that the assumption of homoscedasticity is violated. It can be shown that the variance of the error term in predicting a binary dependent variable is largest when the probability of the dependent variable being 1 (or 0) is near .5 and decreases as this probability moves away from .5.

	P	Q	R
41	Prediction	Sigma Residuals	
42	<.13		0.32544
43	.13-.25		0.325747
44	.26-.31		0.419726
45	.31-.74		0.507828

Figure 17-5: Residual spread as function of predicted dependent variable

The Logistic Regression Model, described in the following section, resolves these three problems.

Logistic Regression Model

Let p = **probability and a binary dependent variable** = 1. In virtually all situations it appears that the relationship between p and the independent variable(s) is nonlinear. Analysts have often found that the relationship in Equation 1 does a good job of explaining the dependence of a binary-dependent variable on an independent variable.

$$(1) \ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

The transformation of the dependent variable described by Equation 1 is often called the *logit transformation*.

In this magazine example, Equation 1 takes on the following form:

$$(2) \ln \frac{P}{1-P} = \text{Intercept} + \text{Slope} * \text{Age}$$

If you take e to both sides of Equation 2 and use the fact that $e^{\ln x} = x$, you can rewrite Equation 2 as one of the following:

$$p = \frac{1}{1 + e^{-(\text{Intercept} + \text{Slope} * \text{Age})}}$$

or

$$p = \frac{e^{(\text{Intercept} + \text{Slope} * \text{Age})}}{1 + e^{(\text{Intercept} + \text{Slope} * \text{Age})}}$$

Equation 3 is often referred to as the *logistic regression model* (or sometimes the *logit regression model*) because the function $y = \frac{1}{1 + e^{-x}}$ is known as the logis-

tic function. Note that $0 < p < 1$. Because p is a probability, this is desirable. In the next section you find the best estimate of slope and intercept to be slope

0.1281 and intercept = 5.662. Substituting these values into Equation 3 shows that as a function of age, a person's chance of being a subscriber varies according to the S-curve, as shown in Figure 17-6. Note that this relationship is highly nonlinear.

%4 Part IV: What do Customers Want?

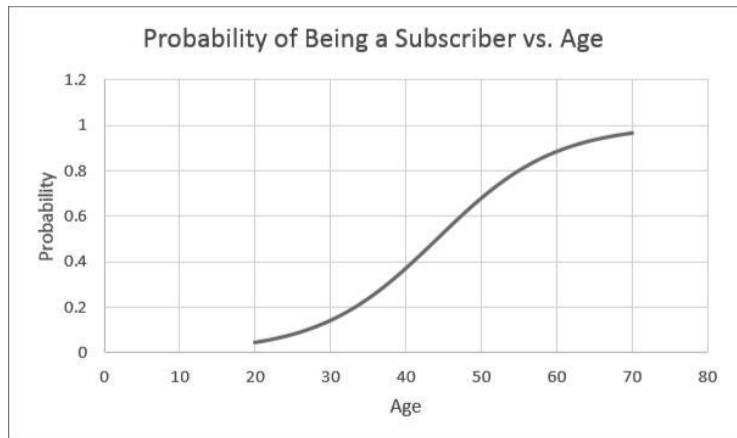


Figure 17-6: S-curve for Being a Subscriber versus Age relationship

Maximum Likelihood Estimate of Logistic Regression Model

This section **demonstrates how to** use the maximum likelihood method to estimate the coefficients in a logistic regression model. Essentially, in the magazine example, the maximum likelihood estimation chooses the slope and intercept to maximize, given the age of each person, the probability or likelihood of the observed pattern of subscribers and nonsubscribers. For each observation in which the person was

%4 subscriber, the probability that the person was a subscriber is given by Equation 4, and for each observation in which the person is not a subscriber, the probability that the person is not a subscriber is given by $1 -$ (right side of Equation 4). If you choose slope and intercept to maximize the product of these probabilities, then you are “maximizing the likelihood” of what you have observed. Unfortunately, the product of these probabilities proves to be a small number, so it is convenient to maximize the natural logarithm of this product. The following equation makes it easy to maximize the log likelihood.

$$\%4 \quad \ln(p_1 * p_2 * \dots p_n) = \ln p_1 + \ln p_2 + \dots \ln p_n$$

The work for this equation is shown in Figure 17-7 and is located in the data worksheet.

	C	D	E	F	G	H	I	J
1	intercept	-5.661581					product	
2	slope	0.128134					3.85064E-09	-19.3750
3	Person#	Age	Subscribe?	Score	Prob	1-Prob	Likelihood	Ln Likelihood
4	1	20	0	-3.0989	0.043	0.95685	0.9568	-0.0441
5	2	23	0	-2.7145	0.062	0.93788	0.9379	-0.0641
6	3	24	0	-2.58637	0.07	0.92998	0.9300	-0.0726
7	4	25	0	-2.45824	0.079	0.92116	0.9212	-0.0821
8	5	25	1	-2.45824	0.079	0.92116	0.0788	-2.5404
9	6	26	0	-2.3301	0.089	0.91134	0.9113	-0.0928
10	7	26	0	-2.3301	0.089	0.91134	0.9113	-0.0928
11	8	28	0	-2.07383	0.112	0.88833	0.8883	-0.1184
12	9	28	0	-2.07383	0.112	0.88833	0.8883	-0.1184
13	10	29	0	-1.9457	0.125	0.87498	0.8750	-0.1336
14	11	30	0	-1.81757	0.14	0.86027	0.8603	-0.1505
15	12	30	0	-1.81757	0.14	0.86027	0.8603	-0.1505
16	13	30	0	-1.81757	0.14	0.86027	0.8603	-0.1505
17	14	30	0	-1.81757	0.14	0.86027	0.8603	-0.1505
18	15	30	0	-1.81757	0.14	0.86027	0.8603	-0.1505
19	16	30	1	-1.81757	0.14	0.86027	0.1397	-1.9681
20	17	32	0	-1.5613	0.173	0.82654	0.8265	-0.1905

Figure 17-7: Maximum likelihood estimation

To perform the maximum likelihood estimation of the slope and intercept for the subscriber example, proceed as follows:

- %4 Enter trial values of the intercept and slope in D1:D2, and name D1:D2 using Create from Selection.
- %4 Copy the formula =intercept+slope*D4 from F4 to F5:F44, to create a “score” for each observation.
- %4 Copy the formula =EXP(F4)/(1+EXP(F4)) from G4 to G5:G44 to use Equation 4 to compute for each observation the estimated probability that the person is a subscriber.
- %4 Copy the formula =1-G4 from H4 to H5:H44 to compute the probability of the person not being a subscriber.
- %4 Copy the formula =IF(E4=1,G4,1-G4) from I4 to I5:I44 to compute the likelihood of each observation.
- %4 In I2 the formula =PRODUCT(I5:I44) computes the likelihood of the observed subscriber and nonsubscriber data. Note that this likelihood is a small number.
- %4 Copy the formula =LN(I4) from J4 to J5:J44, to compute the logarithm of each observation's probability.
- %4 Use Equation 5 in cell J2 to compute the Log Likelihood with the formula =SUM(J4:J44)

%4 Part IV: What do Customers Want?

Use the Solver window (in Figure 17-8), to determine the slope and intercept that maximize the Log Likelihood.

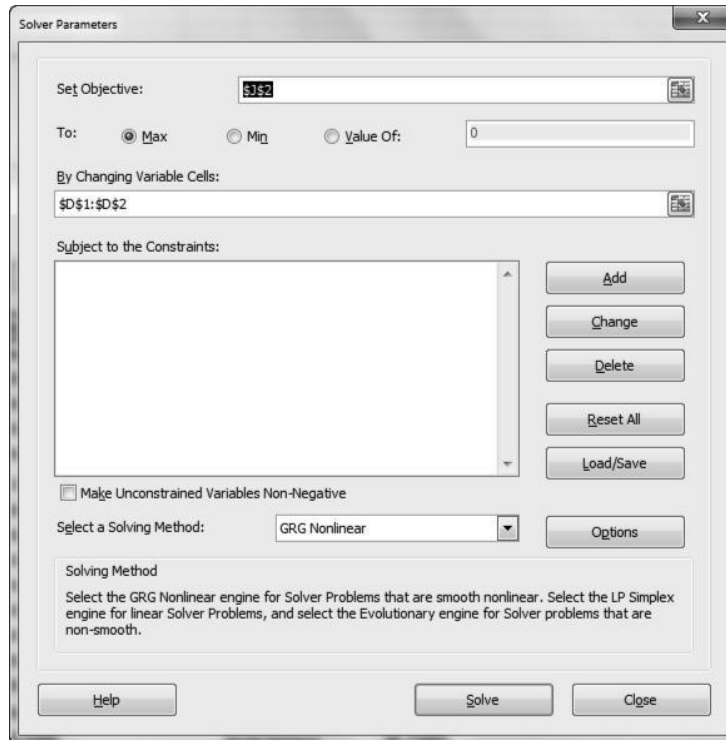


Figure 17-8: Maximum Likelihood Solver window

%4 Press Solve and you find the maximum likelihood estimates of slope = - 5.661 and Intercept = 0.1281.

Using a Logistic Regression to Estimate Probabilities

You can use logistic regression with Equation 4 to predict the chance that a person will be a subscriber based on her age. For example, you can predict the chance that a 44-year-old is a subscriber with the following equation:

$$\frac{e^{(-5.661 + 44 * .1281)}}{1 + e^{(-5.661 + 44 * .1281)}} = .494$$

Since e raised to any power is a positive number, you can see that unlike ordinary least squares regression, logistic regression can never give a negative number for

%4 probability. In Exercise 7 you will show that logistic regression also resolves the problems of heteroscedasticity and non-normal residuals.

Interpreting Logistic Regression Coefficients

In a multiple linear regression, you know how to interpret the coefficient of an independent variable: If β_i is the coefficient of an independent variable x_i , then a unit increase in the independent variable can increase the dependent variable by

%4 i. In a logistic regression, the interpretation of the coefficient of an independent variable is much more complex: Suppose in a logistic regression β_i is the coefficient

of an independent variable x_i . It can be shown (see Exercise 6) that a unit increase in x_i increases the odds ratio ($\frac{\text{Probability } Y=1}{\text{Probability } Y=0}$) by e^{β_i} percent. In the magazine example,

this means that for any age a one-year increase in age increases the odds ratio by $e^{.1281} = 13.7$ percent.

Using StatTools to Estimate and Test Logistic Regression Hypotheses

Chapter 10 explains how easy it is to use the Excel Analysis ToolPak to estimate a multiple linear regression and test relevant statistical hypotheses. The last section showed that it is relatively easy to use the Excel Solver to estimate the coefficients of a logistic regression model. Unfortunately, it is difficult to use “garden variety” Excel to test logistic regression statistical hypotheses. This section shows how to use Palisade’s add-in StatTools (a 15-day trial version is downloadable from Palisade.com) to estimate a logistic regression model and test hypotheses of interest. You can also use full-blown statistical packages such as SAS or SPSS to test logistic regression statistical hypotheses. The work for this section is in the data worksheet of the subscribers.xlsx workbook.

Running the Logistic Regression with StatTools

To start StatTools simply click it on the desktop or start it from the All Programs menu. After bringing up StatTools, proceed as follows:

- %4 Choose Data Set Manager from the StatTools Toolbar, and select the data range (D3:E44), as shown in Figure 17-9.
- %4 From the Regression and Classification menu, select Logistic Regression..., and fill in the dialog box, as shown in Figure 17-10.

%4 Part IV: What do Customers Want?

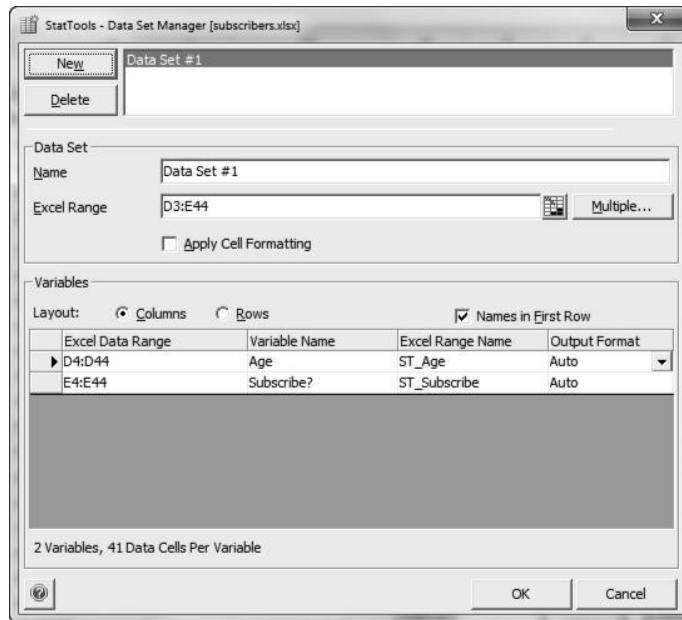


Figure 17-9: Selecting data for the subscriber example

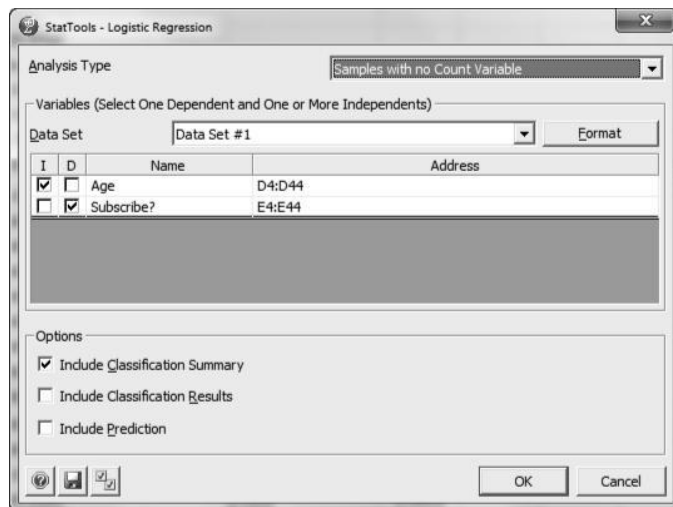


Figure 17-10: Dialog box for Logistic Regression

This dialog box tells StatTools to use Logistic Regression to predict Subscribe from Age. After selecting OK you can obtain the StatTools printout, as shown in Figure 17-11.

	M	N	O	P	Q	R	S	T
16	Summary Measures							
17	Null Deviance	45.554						
18	lo	38.75005						
19	Improvement	6.803948						
20	p-Value	0.0091						
21								
22		Coefficient	Standard	Wald	p-Value	Lower	Upper	Exp(Coeff)
23	Regression Coefficients							
24	Constant	-5.66156	2.017251	-2.80657	0.0050	-9.61537	-1.70775	0.003477
25	Age	0.128133	0.054733	2.34107	0.0192	0.020857	0.235409	1.136704

Figure 17-11: StatTools logistic regression output

Interpreting the StatTools Logistic Regression Output

It is not necessary for the marketing analyst to understand all numbers in the StatTools output, but the key parts of the output are explained here:

- %4 The maximum likelihood estimates of the intercept (-5.661 in cell N24) and slope (0.1281 in cell N25) agree with the maximum likelihood estimates you obtained with the Excel Solver.
- %4 The p-value in cell N20 is used to test the null hypothesis: Adding the age variable improves the prediction of the subscriber over predicting the subscriber with just an intercept. The p-value of 0.009 indicates (via a Likelihood Ratio Chi Squared test) there are only 9 chances in 1,000 that age does not help predict whether a person is a subscriber.
- %4 The p-value in Q25 (0.0192) uses a different test (based on Wald's Statistic) to test whether the age coefficient is significantly different from 0. This test indicates there is a 1.9 percent chance that the age coefficient is significantly different from 0.
- %4 T24:T25 exponentiates each coefficient. This data is useful for using Equation 4 to predict the probability that a dependent variable equals 1.

If you are interested in a more complete discussion of hypothesis testing in logistic regression, check out Chapter 14 of *Introduction to Linear Regression Analysis* (Montgomery, Peck, and Vining, 2006).

%4 Part IV: What do Customers Want?

%4 Logistic Regression with More Than One Independent Variable

The subscribers2.xlsx workbook shows how to use StatTools to run a logistic regression with more than one independent variable. As shown in Figure 17-12, you can try to predict the likelihood that people will subscribe to a magazine (indicated by a 1 in the Subscribe column) based on their age and annual income (in thousands of dollars). To do so, perform the following steps:

	D	E	F	G	H
1					
2	constant	age	income		
3	-3.231111	0.023059	0.015936		
4	Subscribe	Age	Income	Score	Prob
5	0	25	132	-0.55103	0.365624
6	0	64	84	-0.41668	0.39731
7	0	31	100	-0.92265	0.284419
8	0	66	72	-0.5618	0.36313
9	0	39	78	-1.08878	0.251849
10	1	53	54	-1.14842	0.240777
11	1	49	102	-0.47571	0.383265
12	1	63	145	0.532377	0.630037
13	0	66	130	0.362508	0.589647
14	0	72	87	-0.1844	0.454029
15	0	35	89	-1.00571	0.26782
16	0	42	97	-0.71681	0.328096
17	1	65	90	-0.29801	0.426045
18	1	48	90	-0.69001	0.334031
19	0	28	67	-1.51772	0.179797
20	0	46	98	-0.60864	0.352371
21	0	31	55	-1.63978	0.162495
22	0	41	103	-0.64425	0.344287
23	0	57	128	0.123105	0.530737
24	0	70	72	-0.46957	0.384719
25	0	53	81	-0.71814	0.327802
26	0	49	134	0.034252	0.508562
27	1	51	142	0.207861	0.551779

Figure 17-12: Logistic data with two independent variables

- %4 In StatTools use the Data Set Manager to select the range D4:F411.
- %4 From Regression and Classification select Logistic Regression....
- %4 Fill in the dialog box, as shown in Figure 17-13.

You should obtain the output shown in Figure 17-14.

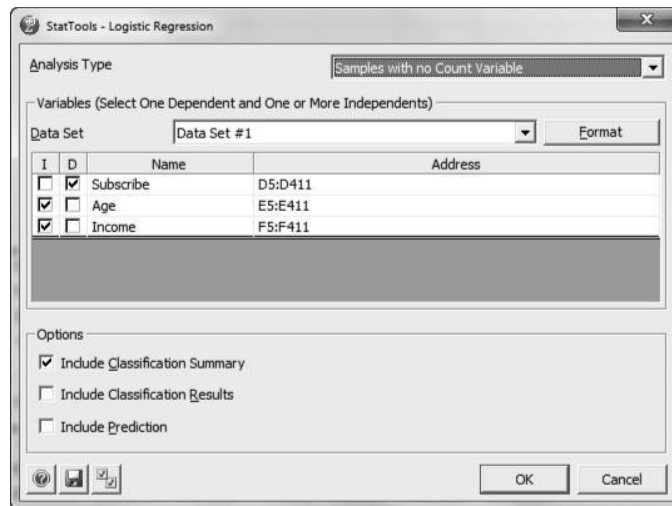


Figure 17-13: Variable selection for a Two Variable Logistic Model

	J	K	L	M	N	O	P	Q
10	Summary Measures							
11	Null Deviance	553.7964						
12	Model Deviance	514.5033						
13	Improvement	39.29306						
14	p-Value	< 0.0001						
15								
16								
17	Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coeff)
18	Constant	-3.23111	0.534179	-6.04874	< 0.0001	-4.2781	-2.18412	0.039514
19	Age	0.023059	0.007163	3.219241	0.0013	0.00902	0.037098	1.023327
20	Income	0.015936	0.002993	5.324267	< 0.0001	0.01007	0.021803	1.016064

Figure 17-14: Output for Two Variable Logistic Regression

The p-value in K14 (< 0.0001) indicates that age and income together have less than 1 chance in 1,000 of not being helpful to predict whether a person is a subscriber. The small p-values in N19 and N20 indicate that each individual independent variable has a significant impact on predicting whether a person is a subscriber. Your estimate of the chance that a person is a subscriber is given by the following equation:

$$(5) \text{ Probability person is a subscriber} = \frac{e^{-3.23 + .023 \text{ Age} + .016 \text{ Income}}}{1 + e^{-3.23 + .023 \text{ Age} + .016 \text{ Income}}}$$

%4 Part IV: What do Customers Want?

In the cell range H5:H411, you use Equation 5 to compute the probability that a person is a subscriber. For example, from H6 you see the probability that a 64-year-old with an \$84,000 annual income is a subscriber is 39.7 percent.

Performing a Logistic Regression with Count Data

In the **previous section** you performed logistic regression from data in which each data point was listed in an individual row. Sometimes raw data is given in a grouped, or *count format*. In count data each row represents multiple data points. For example, suppose you want to know how income (either low, medium, or high) and age (either young or old) influence the likelihood that a person purchases a product. Use the data from countdata2.xlsx, also shown in Figure 17-15. You can find, for example, that the data includes 500 young, low-income people, 35 of whom purchased your product.

	A	B	C	D	E	F	G	H	I	J	K
1						0.8625	1.48651	1.00484			
2	Joint category	Age	Income	Number	Purchasers	Young	LowInc	MidInc	Score	Predicted Prob	Actual Prob
3	1	Young	Low	500	35	1	1	0	-2.54	0.0731	0.07
4	2	Young	Middle	300	15	1	0	1	-3.022	0.0465	0.05
5	3	Young	High	200	4	1	0	0	-4.026	0.0175	0.02
6	4	Old	Low	200	8	0	1	0	-3.402	0.0322	0.04
7	5	Old	Middle	400	7	0	0	1	-3.884	0.0202	0.0175
8	6	Old	High	200	1	0	0	0	-4.889	0.0075	0.005
9											
10	Summary Measures										
11	Null Deviance	29.58									
12	Model Deviance	0.92									
13	Improvement	28.66									
14	p-Value	####									
15											
16		Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coeff)			
17	Regression Coefficients										
18	Constant	-4.889	0.496323	-9.85023	< 0.0001	-5.8617	-3.9161	0.00753			
19	Young	0.863	0.298032	2.89411	0.0038	0.2784	1.446679	2.36916			
20	LowInc	1.487	0.479329	3.101233	0.0019	0.547	2.425994	4.42164			
21	MidInc	1.005	0.500595	2.007282	0.0447	0.0237	1.986002	2.73146			

Figure 17-15: Count data for logistic regression

In F3:H8 IF statements are used to code age and income as dummy variables. The data omits old in the age variable and high in the income variable. For example, the range F8:H8 tells StatTools that row 8 represents old, high-income people.

To run a logistic regression with this count data, perform the following steps:

- %4 Use the Data Set Manager to select the range D2:H8.
- %4 Select Logistic Regression... from Regression and Classification, and fill in the dialog box, as shown in Figure 17-16. Note now the upper-left corner of the dialog box indicates that you told StatTools that count data was used.

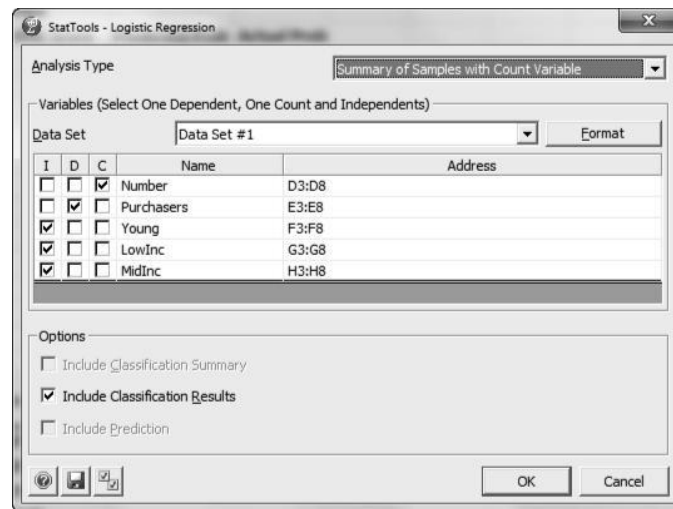


Figure 17-16: Variable selection with count data

- %4 Select Number as the raw Count column and tell StatTools to predict purchasers using the *Young*, *LowInc*, and *MidInc* variables.
- %4 The p-values in E19:E21 indicate that each independent variable is a significant predictor of purchasing behavior. If you generalize Equation 4 you can predict in Column J the probability that a person would subscribe with the formula. Use the following equation to do so:

$$\text{Probability of subscribing} = \frac{e^{-4.89 + .8663 * \text{Young} + 1.487 * \text{LowInc} * 1.005 * \text{MidInc}}}{1 + e^{-4.89 + .8663 * \text{Young} + 1.487 * \text{LowInc} * 1.005 * \text{MidInc}}}$$

Refer to Figure 17-15 to see that the actual and predicted probabilities for each grouping are close.

%4 Part IV: What do Customers Want?

Summary

In this chapter you learned the following:

- %4 When the dependent variable is binary, the assumptions of multiple linear regression are violated.
- %4 To predict the probability p that a binary-dependent variable equals 1, use the logit transformation:

$$\text{Ln} \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- %4 The coefficients in Equation 1 are estimated using the method of Maximum Likelihood.
- %4 You can use programs such as SAS, SPSS, and StatTools to easily estimate the logistic regression model and test the significance of the coefficients.
- %4 Manipulation of Equation 1 shows the probability that a dependent variable equals one and is estimated by the following equation: β

$$\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}}$$

Exercises

- %4 **For 1,022** NFL field goal attempts the file `FGdata.xlsx` contains the distance of the field goal and whether the field goal was good. How does the length of the field goal attempt affect the chance to make the field goal? Estimate the chance of making a 30, 40, or 50 yard field goal.
- %4 The file `Logitsubscribedata.xls` gives the number of people in each age group who subscribe and do not subscribe to a magazine. How does age influence the chance of subscribing to the magazine?
- %4 The file `Healthcaredata.xlsx` gives the age, party affiliation, and income of 300 people. You are also told whether they favor Obamacare. Develop a model to predict the chance that a person favors Obamacare. For each person generate a prediction of whether the person favors Obamacare. Interpret the coefficients of the independent variables in your logistic regression.
- %4 How would you incorporate nonlinearities and/or interactions in a logistic regression model?

%4 The file RFMdata.xlsx contains the following data:

Whether a person responded (dependent variable = 1) or did not respond (dependent variable = 0) to the most recent direct mail campaign.

R: recency (time since last purchase) measured on a 1–5 scale; 5 indicates a recent purchase.

F: frequency (how many orders placed in the last year) measured on a 1–5 scale with a 5 indicating many recent purchases.

M: monetary value (total amount purchased in last year) measured on a 1–5 scale with a 5 indicating the person spent a lot of money in the last year on purchases.

Use this data to build a model to predict based on R, F, and M whether a person will respond to a mailing. What percentage of people responded to the most recent mailing? If you mailed to the top 10 percent (according to the logistic regression model) what response would be obtained?

%4 Show that in a logistic regression model if an independent variable x_i is increased by 1, then the odds ratio will increase by a factor of e^{β_i}

%4 Using the data in the data worksheet of the file subscribers.xlsx, show that for the magazine example logistic regression resolves the problems of non-normal residuals and heteroscedasticity.

%4 The file shuttledata.xlsx contains the following data for several launches of the space shuttle:

Temperature (degrees Fahrenheit)

Number of O-rings on the shuttle and the number of O-rings that failed during the mission

Use logistic regression to determine how temperature affects the chance of an O-ring failure. The *Challenger* disaster was attributed to O-ring failure. The temperature at launch was 36 degrees. Does your analysis partially explain the *Challenger* disaster?

%4 In 2012 the Obama campaign made extensive use of logistic regression. If you had been working for the Obama campaign how would you have used logistic regression to aid his re-election campaign?



Discrete Choice 18 Analysis

In Chapter 16, “Conjoint Analysis,” you learned how the marketing analyst could use a customer’s ranking of product profiles to determine the relative importance the customer attaches to product attributes and how a customer ranks the level of each product attribute. In this chapter you will learn how to use *discrete choice analysis* to determine the relative importance the customer attaches to product attributes and how a customer ranks the level of each product attribute. To utilize discrete choice analysis, customers are asked to make a choice among a set of alternatives (often including none of the given alternatives). The customer need not rank the product attributes. Using maximum likelihood estimation and an extension of the logit model described in Chapter 17, “Logistic Regression,” the marketing analyst can estimate the ranking of product attributes for the population represented by the sampled customers and the ranking of the level of each product attribute.

In this chapter you learn how discrete choice can be used to perform the following:

- %4 Estimate consumer preferences for different types of chocolate.
- %4 Estimate price sensitivity and brand equity for different video games, and determine the profit maximizing price of a video game.
- %4 Model how companies should dynamically change prices over time.
- %4 Estimate price elasticity.
- %4 Determine the price premium a national brand product can command over a generic product.

Random Utility Theory

The concept of *random utility theory* provides the theoretical basis for discrete choice analysis. Suppose a decision maker must choose among n alternatives. You can observe certain attributes and levels for each alternative.

Part IV: What do Customers Want?

The decision maker associates a utility U_j with the j th alternative. Although the decision maker knows these utilities, the marketing analyst does not. The random utility model assumes the following:

$$U_j = V_j + \epsilon_j$$

Here V_j is a deterministic “score” based on the levels of the attributes that define the alternative, and the ϵ_j ’s are random unobservable error terms. The decision maker is assumed to choose the alternative j ($j = 1, 2, \dots, n$) having the largest value of U_j . Daniel McFadden showed (see “Conditional Logit Analysis of Qualitative Choice Behavior” in *Frontiers of Econometric Behavior*, Academic Press, 1974) that if the ϵ_j ’s are independent *Gumbel* (also known as *extreme value*) random variables, each having the following distribution function: $F(x) = \text{Probability } \epsilon_j \leq x = 1 - e^{-x}$,

then the probability that the decision maker chooses alternative j (that is, $U_j = \max_{k=1,2,\dots,n} U_k$) is given by the following:

$$(1) \quad \frac{e^{V_j}}{\sum_{i=1}^n e^{V_i}}$$

Of course, Equation 1 is analogous to the logit model of Chapter 17 and is often called the *multinomial logit model*. Equation 1 is of crucial importance because it provides a reasonable method for transforming a customer’s score for each product into a reasonable estimate of the probability that the person will choose each product. In the rest of this chapter you will be given the alternative chosen by each individual in a set of decision makers. Then you use Equation 1 and the method of maximum likelihood (introduced in Chapter 17) to estimate the importance of each attribute and the ranking of the levels within each attribute.

NOTE If this discussion of random utility theory seems too technical, do not worry; to follow the rest of the chapter all you really need to know is that Equation

1 can be used to predict the probability that a decision maker chooses a given alternative.

NOTE Another commonly used assumption is that the errors in Equation 1 are not independent and have marginal distributions that are non-normal. This assumption can be analyzed via *probit regression*, which is beyond the scope of this book. In his book *Discrete Choice Methods with Simulation*, 2nd ed. (Cambridge University Press, 2009), Kenneth Train determined how changes in

the price of different energy sources (gas, electric, and so on) affect the fraction of homes choosing different sources of energy. You can refer to his book for an extensive discussion of probit regression.

Discrete Choice Analysis of Chocolate Preferences

Suppose that **eight** alternative types of chocolate might be described by the levels of the following attributes:

- %4 Dark or milk
- %4 Soft or chewy
- %4 Nuts or no nuts

The eight resulting types of chocolate are listed here:

- %4 Milk, chewy, no nuts
- %4 Milk, chewy, nuts
- %4 Milk, soft, no nuts
- %4 Milk, soft, nuts
- %4 Dark, chewy, no nuts
- %4 Dark, chewy, nuts
- %4 Dark, soft, no nuts
- %4 Dark, soft, nuts

Ten people were asked which type of chocolate they preferred. (People were not allowed to choose none of the above.) The following results were obtained:

- %4 Two people chose milk, chewy, nuts.
- %4 Two people chose dark, chewy, no nuts.
- %4 Five people chose dark, chewy, nuts.
- %4 One person chose dark, soft, nuts.

You can use discrete choice analysis to determine the relative importance of the attributes and, within each attribute, rank the levels of the attributes. The work for this process is shown in Figure 18-1 and is in the `chocolate.xls` file.

%4 Part IV: What do Customers Want?

	A	B	C	D	E	F	G	H
2								
3			Dark	Milk	Soft	Chewy	Nuts	No Nuts
4		weights	0	-1.38629458	2.401382	4.598606	5.923627529	5.07633
5		number choosing						
6				exp(score)				
7	Score	milk chewy no nuts	8.288641	3978.42311		we set dark weight =0		
8	Score	milk chewy nuts	9.135939	9282.98697				
9	Score	milk soft nonuts	6.091417	442.047234				
10	Score	milk soft nuts	6.938715	1031.44351				
11	Score	dark chewy no nuts	9.674935	15913.6959				
12	Score	dark chewy nuts	10.52223	37131.9559			sum	
13	Score	dark soft nonuts	7.477711	1768.18932			-14.36349699	
14	Score	dark soft nuts	8.325009	4125.77494	Number			
15	Prob	milk chewy no nuts	0.054		choosing			
16	Prob	milk chewy nuts	0.126		0			
17	Prob	milk soft nonuts	0.006		2			
18	Prob	milk soft nuts	0.014		0			
19	Prob	dark chewy no nuts	0.216		0			
20	Prob	dark chewy nuts	0.504		2			
21	Prob	dark soft nonuts	0.024		5			
22	Prob	dark soft nuts	0.056		1			
23		milk chewy no nuts			0			

Figure 18-1: Chocolate discrete choice example

Assume the score for each of the eight types of chocolate will be determined based on six unknown numbers which represent the “value” of various chocolate characteristics.

- %4 Value for dark
- %4 Value for milk
- %4 Value for soft
- %4 Value for chewy
- %4 Value for nuts
- %4 Value for no nuts

The score, for example of milk, chewy, no nuts chocolate, is simply determined by the following equation:

$$(2) (\text{Value of Milk}) + (\text{Value of Chewy}) + (\text{Value of No Nuts})$$

To estimate the six unknowns proceed as follows:

- %4 In cells C4:H4 enter trial values for these six unknowns.
- %4 In cells C7:C14 compute the score for each type of chocolate. For example, in C7 Equation 2 is used to compute the score of milk, chewy, no nuts with the formula =D4+F4+H4
- %4 Use Equation 1 to compute the probability that a person will choose each of the eight types of chocolate. To do so copy the formula =EXP(C7) from D7 to D8:D14 to compute the numerator of Equation 1 for each type of chocolate.

Then copy the formula `=D7/SUM($D$7:$D$14)` from C15 to C16:C22 to implement Equation 1 to compute the probability that a person will choose each type of chocolate.

%4 Use the technique of maximum likelihood (actually maximizing the logarithm of maximum likelihood) to determine the six values that maximize the likelihood of the observed choices. Note that Equation 2 implies that if you add the same constant to the score of each alternative, then for each alternative the probabilities implied by Equation 2 remain unchanged. Therefore when you maximize the likelihood of the observed choices, there is no unique solution k . Therefore you may arbitrarily set any of the changing cells to 0. For this example suppose the value of dark = 0.

The likelihood of what you have observed is given by the following equation:

$$(3) (P \text{ of } M, C, N)^2 (P \text{ of } D, C, NN)^2 (P \text{ of } D, C, N)^5 (P \text{ of } D, S, N)^1$$

Essentially you find the likelihood of what has been observed by taking the probability of each selected alternative and raising it to a power that equals the number of times the selected alternative is observed. For example, in the term $(P \text{ of } D, C, N)^5$ $(P \text{ of } D, C, N)$ is raised to the fifth power because five people preferred dark and chewy with nuts.

Equation 3 uses the following abbreviations:

%4 P = Probability
 %4 M = Milk
 %4 D = Dark
 %4 N = Nuts
 %4 NN = No Nuts
 %4 C = Chewy
 %4 S = Soft

Recall the following two basic laws of logarithms:

$$\begin{aligned} \%4 \quad \ln(x * y) &= \ln x + \ln y \\ \%4 \quad \ln p^n &= n \ln p \end{aligned}$$

You can apply Equations 4 and 5 to Equation 3 to show that the natural logarithm of the likelihood of the observed choices is given by the following equation:

$$\%4 \quad 2 * \ln(P \text{ of } M, C, N) + 2 * (P \text{ of } D, C, NN) + 5 * \ln(P \text{ of } D, C, N) + \ln(P \text{ of } D, S, N)$$

Cell G13 computes the Log Likelihood with the formula `=E16*LN(C16)+E19*LN(C19)+E20*LN(C20)+E21*LN(C21)`

%4 Part IV: What do Customers Want?

You can use the Solver window in Figure 18-2 to determine the maximum likelihood estimates of the values for each level of each attribute. The values enable you to determine the importance of each attribute and the rankings within each attribute of the attribute levels.

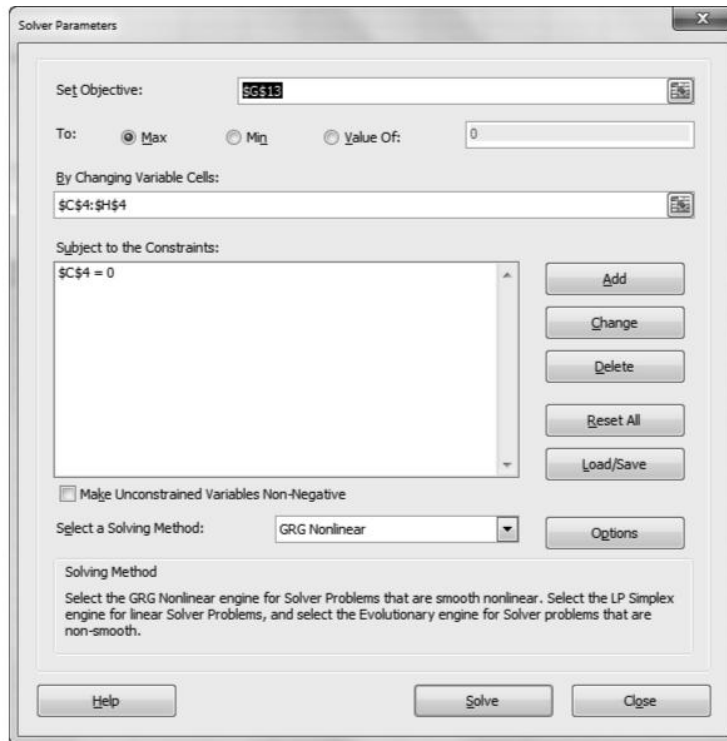


Figure 18-2: Solver window for chocolate example

Refer to Figure 18-1 and you can find the following maximum likelihood estimates:

- %4 Dark = 0, Milk = -1.38, so Dark is preferred to Milk overall and the difference between the Dark and Milk values is 1.38.
- %4 Chewy = 4.59 and Soft = 2.40, so Chewy is preferred to Soft overall and the difference between the Chewy and Soft scores is 2.19.
- %4 Nuts = 5.93 and No Nuts = 5.08, so Nuts is preferred to No Nuts and the difference between the Nuts and No Nuts scores is 0.85.

You can rank the importance of the attributes based on the difference in scores within the levels of the attribute. This tells you that the attributes in order of importance

are Chewy versus Soft, Dark versus Milk, and Nuts versus No Nuts. This is consistent with the fact that nine people prefer Chewy, eight prefer Dark, and seven prefer Nuts.

Incorporating Price and Brand Equity into Discrete Choice Analysis

In this section you learn how to incorporate price and brand equity into a discrete choice analysis. The first model is in the file *Xbox1.xls* and the data is shown in Figure 18-3.

	A	B	C	D	E	F	G	H	I	J
1		cost	\$180	\$160	\$155					
2										
3						Nothing				
4		Brand	1.879	4.967	0.489	0				
5		price	-0.012	-0.024	-0.01					
6		Choice#	Xbox	PS	Wii		Pick Xbox	Pick PS	Pick Wii	Pick nothing
7		1	\$221	\$267	\$275		28	11	14	47
8		2	\$193	\$295	\$275		38	1	14	47
9		3	\$278	\$294	\$176		10	2	43	45
10		4	\$288	\$191	\$250		2	44	10	44
11		5	\$162	\$224	\$221		30	16	17	37
12		6	\$172	\$249	\$157		23	15	24	38
13		7	\$167	\$251	\$169		24	12	25	39
14		8	\$213	\$255	\$158		21	25	9	45

Figure 18-3: Video game data

The data indicates that 100 people were shown eight price scenarios for Xbox, PlayStation (PS), and Wii. Each person was asked which product they would buy (or none) for the given prices. For example, if Xbox sold for \$221, PS for \$267, and Wii for \$275, 28 people chose Xbox, 11 chose PS, 14 chose Wii, and 47 chose to buy no video game console. You are also given the cost to produce each console.

You will now learn how to apply discrete choice analysis to estimate price sensitivity and product brand equity in the video game console market. For each possible product choice, assume the value may be computed as follows:

- %4 Value Xbox = Xbox brand weight + (Xbox price) * (Xbox price sensitivity)
- %4 Value PS = PS brand weight + (PS price) * (PS price sensitivity)
- %4 Value Wii = Wii brand weight + (Wii price) * (Wii price sensitivity)
- %4 Assume Weight for Nothing = 0, so Value Nothing = 0

Now you can use maximum likelihood estimation to estimate the brand weight for each product and the price sensitivity for each product. The strategy is to find the Log Likelihood of the data for each set of prices, and then, because of Equation 4,

%4 Part IV: What do Customers Want?

add the Log Likelihoods. By maximizing the Log Likelihood you can obtain the set of weights that maximize the chance to observe the actual observed choice data. The work is shown in Figure 18-4. Proceed as follows:

	G	H	I	J	K	L	M	N	O	P	Q	R	S
4													sum
5													-984.7550485
6	Pick Xbox	Pick PS	Pick Wii	Pick nothing	XBox score	PS Score	Wii Score	Nothing Score	XBox prob	PS prob	Wii Prob	Nothing Prob	Ln(Likelihood)
7	28	11	14	47	-0.882	-1.5539	-1.218	0	0.2155	0.11004	0.154	0.52045	-124.1354628
8	38	1	14	47	-0.532	-2.2376	-1.218	0	0.2952	0.053621	0.1487	0.5025	-108.3181076
9	10	2	43	45	-1.594	-2.2132	-0.603	0	0.1092	0.058804	0.2942	0.53777	-108.3357813
10	2	44	10	44	-1.719	0.30214	-1.062	0	0.0623	0.470087	0.1201	0.3475	-106.4650498
11	30	16	17	37	-0.145	-0.5038	-0.882	0	0.3001	0.209578	0.1435	0.34683	-133.2964867
12	23	15	24	38	-0.27	-1.1143	-0.485	0	0.2821	0.121213	0.2274	0.36938	-134.1584807
13	24	12	25	39	-0.207	-1.1631	-0.56	0	0.3014	0.115888	0.2119	0.37083	-132.1271517
14	21	25	9	45	-0.782	-1.2608	-0.492	0	0.1945	0.120471	0.26	0.42505	-137.918528

Figure 18-4: Maximum likelihood estimation for video console example

- %4 In C4:E5 enter trial values for brand coefficients and price sensitivities for each product. In F4 set the Nothing “brand” coefficient equal to 0.
- %4 Copy the formula =C\$4+C\$5*C7 from K7 to K7:M14 to generate scores for each product for each scenario.
- %4 Copy the formula =F\$4 from N7 to N7:N14 to generate for each scenario the score for no purchase.
- %4 Copy the formula =EXP(K7)/(EXP(\$K7)+EXP(\$L7)+EXP(\$M7)+EXP(\$N7)) from O7 to O7:R14 to utilize Equation 1 to compute for each price scenario the probability that each product (or no purchase) is chosen.
- %4 To see how to compute the likelihood of the observed results for each scenario, note that in the first scenario the likelihood is as follows:

$$(\text{Prob. Xbox chosen})^{28} (\text{Prob. PS chosen})^{11} (\text{Prob. GC chosen})^{14} (\text{Prob. nothing chosen})^{47}$$

The goal is to maximize the sum of the logarithms of the likelihoods for each scenario. For the first scenario the logarithm of the likelihood is as follows:

$$28 * \text{Ln}(\text{Prob. Xbox chosen}) + 11 * \text{Ln}(\text{Prob. PS chosen}) + 14 * \text{Ln}(\text{Prob. GC chosen}) + 47 * \text{Ln}(\text{Prob. nothing chosen})$$

This follows from Equations 4 and 5. Therefore copy the formula G7*LN(O7)+H7*LN(P7)+I7*LN(Q7)+J7*LN(R7) from S7 to S7:S14 to compute the logarithm of the likelihood for each scenario.

- %4 In cell S5 you can create the target cell needed to compute maximum likelihood estimates for the changing cells (the range C4:E5 and F4). The target cell is computed as the sum of the Log Likelihoods with the formula =SUM(S7:S14)

%4 Use the Solver window shown in Figure 18-5 to determine the brand weights and price sensitivities that maximize the sum of the Log Likelihoods. This is equivalent, of course, to maximizing the likelihood of the observed choices given by Equation 3. Note that you constrained the score for the Nothing choice to equal 0 with the constraint $FF\$4 = 0$.



Figure 18-5: Solver window for maximum likelihood estimation

The weights and price sensitivities are shown in the cell range C4:E5 of Figure 18-3. At first glance it appears that PS has the highest brand equity, but because PS also has the most negative price sensitivity, you are not sure. In the “Evaluating Brand Equity” section you constrain each product’s price weight to be identical so that you can obtain a fair comparison of brand equity for each product.

Pricing Optimization

After you have estimated the brand weights and price sensitivities, you can create a market simulator (similar to the conjoint simulator of Chapter 16) to help a company price to maximize profits. Suppose you work for Xbox and want to determine

%4 Part IV: What do Customers Want?

a profit maximizing price. Without loss of generality, assume the market consists of 100 customers. Then your goal is to price Xbox to maximize profit earned per 100 customers in the market. Make the following assumptions:

- Xbox price is \$180.
- PS price is \$215 and Wii price is \$190.
- Each Xbox purchaser buys seven games.
- Xbox sells each game for \$40 and buys each game for \$30.

To determine what price maximizes profit from Xbox look at the worksheet pricing in file Xbox1.xls (also shown in Figure 18-6).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		cost	\$180.00	\$160.00	\$155.00									
2														
3						Nothing								
4		Brand	1.87892	4.9665	0.488912494	0								
5		price	-0.0125	-0.0244	-0.00620515									
6		Price	X BOX	PS 2	Game Cube		XBOX score	PS	Wii	Nothing	XBox	PS	Wii	Nothing
7			1	\$207.46	\$215.00	\$190.00	-0.712716307	Score	Score	Score	Prob	Prob	Prob	Prob
8							XBOX Profit Computation	-0.28396	-0.69	0	0.179	0.274	0.183	0.36435
9							unit sales	17.86425						
10		Games sold	7				Revenues	\$8,708.12						
11		Game price	\$40.00				Costs	\$6,967.06						
12		Game cost	\$30.00				Total Profit	\$1,741.07						

Figure 18-6: Determining Xbox price

Proceed as follows:

%4 To perform the pricing optimization you only need one row containing prob-ability calculations from the model worksheet. Therefore in Row 7 of the pricing worksheet recreate Row 7 from the model worksheet. Also insert a trial price for Xbox in C7, and enter information concerning Xbox in B10 and B11.

%4 In cell H9, compute unit sales of Xbox with the formula =100*K7

%4 In cell H10, compute Xbox revenue from consoles and games with the formula
C7*H9+H9*B10*B11

%4 In cell H11, enter the cost of Xbox consoles and games with the formula
=H9*C1 +B10*H9*B12.

%4 In cell H12, compute the profit with the formula =H10-H11

%4 Use the Solver window in Figure 18-7 to yield a profit-maximizing price of \$207.47 for Xbox.



Figure 18-7: Solver window for Xbox price

Evaluating Brand Equity

Based on the pricing analyses performed in the previous section, you can also estimate the value of each brand, or brand equity, by forcing the price weight to be identical across all competitors. For instance, you might say PS has the largest brand equity because its brand coefficient is the largest. As previously pointed out, however, PS has the most negative price weight, so it is not clear that PS has the largest brand equity. For example, if PS had the largest brand equity then you would think that the scores and choice mechanism would indicate that the market would prefer a PS over an Xbox or Wii at a \$200 price. Due to the large negative price coefficient for PS, it is unclear whether PS would be preferred if all products sold for the same price. To correctly determine brand equity, you must rerun the Solver model assuming that the price weight is the same for each product. Setting the price weights equal for all products ensures that if all products sell for the same price the market would prefer the product with the largest brand weight. The results are in the worksheet Brand equity and price of work-book Xbox2.xls, and the changes from the previous model are shown in Figure 18-8.

%4 Part IV: What do Customers Want?

Set D5 and E5 equal to C5, and delete D5 and E5 as changing cells. After running Solver, you find that PS has the largest brand equity, followed by Xbox and Wii.

	A	B	C	D	E	F
1		Cost	\$180.00	\$160.00	\$155.00	
2						
3						pick nothing
4		Brand	1.904231	2.10939	1.74652854	0
5		Price	-0.01254	-0.01254	-0.01253911	
6		Price#	X Box	PS 2	Wii	
7		1	\$221.00	\$267.00	\$275.00	
8		2	\$193.00	\$295.00	\$275.00	
9		3	\$278.00	\$294.00	\$176.00	
10		4	\$288.00	\$191.00	\$250.00	
11		5	\$162.00	\$224.00	\$221.00	
12		6	\$172.00	\$249.00	\$157.00	
13		7	\$167.00	\$251.00	\$169.00	
14		8	\$213.00	\$255.00	\$158.00	

Figure 18-8: PS has the largest brand equity

Testing for Significance in Discrete Choice Analysis

Once you have estimated the brand values and price weights, you can plug the product prices into Equation 2 and forecast market shares. You might wonder whether including a factor such as price in market share forecasts improves the quality of the forecasts. To determine whether or not adding a factor such as price to a discrete choice analysis significantly improves the model's predictions for product market shares, proceed as follows:

%4 Let $LL(Full)$ = Log Likelihood when all changing cells are included.

%4 Let $LL(Full - q)$ = Log Likelihood when changing cell q is omitted from the model.

%4 Let $DELTA(q) = 2 * (LL(Full) - LL(Full - q))$. Then $DELTA(q)$ follows a Chi-Square random variable with 1 degree of freedom, and a p-value for the significance of q may be computed with the Excel formula

$CHIDIST(DELTA(q),1)$

%4 If this formula yields a p-value less than .05, you may conclude that the changing cell q is of significant use (after adjusting for other changing cells in the model) in predicting choice.

You can use this technique to test if incorporating price in your choice model is a significant improvement over just using the brand weights. In the Brand Equity and Price worksheet of workbook *Xbox2.xls*, you can find $LL(Full Model) = -996.85$. In worksheet Brand equity and price the model is run with the price weight set to 0, which obtains $LL(Model(no price)) = -1045.03$. Then $Delta(q) = 96.37$ and

$\text{CHIDIST}(96.37,1) = 10^{-22}$. It is therefore virtually certain that price is a significant factor in the choice of a video game console.

Dynamic Discrete Choice

Although the previous sections and analyses offer important insights regarding the drivers of consumer choice, as well as pricing and brand equity implications, they do not represent the “real world” in the sense that the previous analyses do not address competitive dynamics. In the real world, if one firm changes its price, it should expect competitors to react. This section addresses competitive dynamics in discrete choice analysis.

If Xbox sets a price of \$207 as recommended previously, then PS and Wii would probably react by changing their prices. In this section you will learn how to follow through the dynamics of this price competition. Usually one of two situations occurs:

- %4 A price war in which competitors continually undercut each other
- %4 A *Nash equilibrium* in which no competitor has an incentive to change its price. Nash equilibrium implies a stable set of prices because nobody has an incentive to change their price.

The file *Xbox3.xls* contains the analysis. The model of competitive dynamics is shown in Figure 18-9.

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Cost	\$180.00	\$160.00	\$155.00									
2													
3					Nothing								
4	Brand	1.87892	4.966548	0.48891	0								
5	Price#	-0.01249	-0.024421	-0.0062									
6	Price	Xbox	PS	Wii		Xbox score	PS Score	Wii Score	Nothing Score	XBox Prob	PS Prob	Wii Prob	Nothing Prob
7		1	\$207.46	\$215.00	\$190.00	-0.712715	-0.284	-0.6900664	0	0.17864	0.274	0.183	0.364345
8						Xbox	PS	Wii					
9						unit sales	17.86	27.43	18.27				
10						revenue	3706.13	5896.97	3471.96				
11						cost	3215.56	4388.44	2832.39				
12						profit	1741.07	3428.47	1918.71				
13													
14													
15	Xbox	\$207.46	\$215.00	\$190.00									
16	PS	\$207.46	\$173.55	\$190.00									
17	Wii	\$207.46	\$173.55	\$260.74									
18	Xbox	\$202.38	\$173.55	\$260.74									
19	PS	\$202.38	\$175.16	\$260.74									
20	Wii	\$202.38	\$175.16	\$260.92									
21	Xbox	\$202.64	\$175.16	\$260.92									
22	PS	\$202.64	\$175.18	\$260.92									
23	Wii	\$202.64	\$175.18	\$260.94									

Figure 18-9: Dynamic pricing of Video game consoles

4 Part IV: What do Customers Want?

To begin, copy the formulas in H9:H12 to I9:J12. This computes profit for PS and Wii, respectively.

Change Solver to maximize I12 by changing D7. The resulting Solver solution shows that PS should now charge \$173.55.

Next, change the target cell to maximize J12 by changing E7. The resulting Solver solution shows that given current prices of other products, Wii should charge \$260.74.

Now it is Xbox's turn to change the Xbox price: maximize H12 by changing C7. Xbox now sells for \$202.38. Repeating the process, you find that prices appear to stabilize at the following levels:

Xbox: \$202.64

PS: \$175.18

Wii: \$260.94

With this set of prices, no company has an incentive to change its price. If it changes its price, its profit will drop. Thus you have found a stable Nash equilibrium.

Independence of Irrelevant Alternatives (IIA) Assumption

Discrete choice analysis implies that the ratio of the probability of choosing alternative j to the chance of choosing alternative i is independent of the other available choices. This property of discrete choice modeling is called the *Independence of Irrelevant Alternatives (IIA)* assumption. Unfortunately, in many situations you will learn that IIA is unreasonable.

From Equation 2 you can show (see Exercise 6) that for any two alternatives i and j , Equation 7 holds.

$$(7) \quad \frac{\text{Prob Choose } i}{\text{Prob Choose } j} = \frac{e^{\text{Score}_i}}{e^{\text{Score}_j}}$$

Equation 7 tells you that the ratio of the probability of choosing alternative i to the probability of choosing alternative j is independent of the other available choices. This property of discrete choice modeling is called the *Independence of Irrelevant Alternatives (IIA)* assumption. The following paradox (called the Red Bus/Blue Bus Problem) shows that IIA can sometimes result in unrealistic choice probabilities.

Suppose people have two ways to get to work: a Blue Bus or a Car. Suppose people equally like buses and cars so the following is true:

$$\text{Probability of Choosing Blue Bus} = \text{Probability of Choosing Car} = 0.5$$

Now add a third alternative: a Red Bus. Clearly the Red and Blue Buses should divide the 50-percent market share for buses and the new probabilities should be:

```
%4 Probability Car = 0.5
%4 Probability Blue Bus = 0.25
%4 Probability Red Bus = 0.25
```

The IIA assumption implies that the current ratio of Probability Blue Bus/Probability Car will remain unchanged. Because Probability Red Bus = Probability Blue Bus, you know that discrete choice will produce the following unrealistic result:

```
%4 Probability Car = 1/3
%4 Probability Blue Bus = 1/3
%4 Probability Red Bus = 1/3
```

The IIA problem may be resolved by using more advanced techniques including mixed logit, nested logit, and probit. See Train (2009) for details.

Discrete Choice and Price Elasticity

In Chapter 4, “Estimating Demand Curves and Using Solver to Optimize Price,” you learned that a product’s price elasticity is the percentage change in the demand for a product resulting from a 1 percent increase in the product’s price. In this section you will learn how discrete choice analysis can be used to estimate price elasticity. Define the following quantities as indicated:

```
%4 Prob(j) = Probability product j is chosen.
%4 Price(j) = Price of product j
%4 β (j) = Price coefficient in score equation.
```

From Equation 2 it can be shown that:

$$(8) \text{ Price Elasticity for Product } j = (1 - \text{Prob}(j)) * \text{Price}(j) * \beta (j)$$

The cross elasticity (call it $E(k, j)$) of product k with respect to product j is the percentage change in demand for product k resulting from a 1 percent increase in the price of product j . From Equation 2 it can be shown that:

$$(9) E(k, j) = - \text{Prob}(j) * \text{Price}(j) * \beta (j)$$

%4 Part IV: What do Customers Want?

NOTE All products other than product j have the same cross elasticity. This is known as the *property of uniform cross elasticity*

The file Xbox4.xls illustrates the use of Equations 8 and 9. Return to the video game example and assume that the current prices are Xbox = \$200, PS = \$210, and Wii = \$220, as shown in Figure 18-10.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Cost	\$180.00	\$160.00	\$155.00									
2													
3					Pick Nothing								
4	Brand	1.9042314	2.1093899	1.74653	0								
5	Price	-0.0125391	-0.012539	-0.0125									
6		Xbox	PS	Wii		Xbox score	PS Score	Wii Score	Nothing Score	Xbox Prob	PS Prob	Wii Prob	Nothing Prob
7	Price	\$200.00	\$210.00	\$220.00		-0.60359	-0.524	-1.012	0	0.219	0.24	0.145	0.39959
8													
9													
10		Xbox elasticity	-1.96										
11		PS Cross elasticity with Xbox	0.55										
12		Wii Cross elasticity with Xbox	0.55										

Figure 18-10: Elasticities for Xbox example

The price elasticity for Xbox is -1.96 (a 1 percent increase in Xbox price reduces Xbox sales by 1.96 percent) and the cross elasticities for Xbox price are .55. (A 1 percent increase in Xbox price increases demand for other products by 0.55 percent)

Summary

In this chapter you learned the following:

- %4 A discrete choice analysis helps the marketing analyst determine what attributes matter most to decision makers and how levels of each attribute are ranked by decision makers.
- %4 To begin a discrete choice analysis, decision makers are shown a set of alternatives and asked to choose their most preferred alternative.

- %4 The analyst must determine a model that is used to “score” each alternative based on the level of its attributes.
- %4 The fraction of decision makers that choose alternative j is assumed to follow the multinomial logit model:

$$(1) \frac{e^{V_j}}{\sum_{i=1}^n e^{V_i}}, \text{ where } V_j = \text{score for alternative } j$$

- %4 Maximum Likelihood is used to estimate the parameters (such as brand equity and price sensitivity) in the scoring equation.
- %4 In a discrete analysis, a Chi Square Test based on the change in the Log Likelihood Ratio can be used to assess the significance of a changing cell.
- %4 The Independence of Irrelevant Alternatives (IIA) that follows from Equation 2 implies that the ratio of the probability to choose alternative i to the probability to choose alternative j is independent of the other available choices. In situations like the Red Bus/Blue Bus example, discrete choice analysis may lead to unrealistic results.
- %4 The multinomial logit version of discrete choice enables you to easily compute price elasticities using the following equations:

$$\begin{aligned} \text{Price Elasticity for Product } j &= (1 - \text{Prob}(j)) * \text{Price}(j) * \beta(j) \\ E(k,j) &= -\text{Prob}(j) * \text{Price}(j) * \beta(j) \end{aligned}$$

Exercises

- %4 **In her book** *Discrete Choice Modeling and Air Travel Demand* (Ashgate Publishing, 2010), Per Laurie Garrow details how airlines have predicted how changes in flight prices will affect their market share. Use this example to perform some similar analysis. Delta Airlines wants to determine the profit maximizing price to charge for a 9 a.m. New York to Chicago flight. A focus group has been shown 16 different flights and asked if it would choose a flight if the flight were available. The results of the survey are in file *AirlineData.xlsx*, and a sample is shown in Figure 18-11. For example, when shown a Delta 8 a.m., four-hour flight with no music (audio), video, or meals, priced at \$300, 78 people chose that flight over choosing not to take a flight at all (Delta = 0, United = 1).

%4 Part IV: What do Customers Want?

	F	G	H	I	J	K	M	N	O	P	Q	R
	Price	Leave 8	Leave 9	Leave Noon	Leave other time	Length of flight	Audio	Video	Meals	Airline	Like flight?	Not?
13	300	1	0	0	0	4	0	0	0	0	78	22
14	300	0	1	0	0	5	0	1	1	0	60	40
15	300	0	0	1	0	6	1	0	1	1	50	50
16	300	0	0	0	1	7	1	1	0	1	28	72
17	400	1	0	0	0	5	1	0	1	1	60	40
18	400	0	1	0	0	4	1	1	0	1	50	50
19	400	0	0	1	0	7	0	0	0	0	20	80
20	400	0	0	0	1	6	0	1	1	0	33	67
21	500	1	0	0	0	6	0	1	0	1	10	90
22	500	0	1	0	0	7	0	0	1	1	13	87
23	500	0	0	1	0	4	1	1	1	0	38	60
24	500	0	0	0	1	5	1	0	0	0	22	78
25	600	1	0	0	0	7	1	1	1	0	30	70
26	600	0	1	0	0	6	1	0	0	0	7	93
27	600	0	0	1	0	5	0	1	0	1	12	88
28	600	0	0	0	1	4	0	0	1	1	15	85

Figure 18-11: Data for Exercise 1

%4 Which airline (Delta or United) has more brand equity on this route?

%4 Delta wants to optimally design a 9 a.m. flight. The flight will have audio and take six hours. There are 500 potential flyers on this route each day. The plane can seat at most 300 people. Determine the profit maximizing price, and whether Delta should offer a movie and/or meals on the flight. The only other flight that day is a \$350 United 8 a.m., five-hour flight with audio, movies, and no meals. Delta's cost per person on the flight breaks down as follows:

Item	Cost Per Person
Fuel	\$60
Food	\$40
Movie	\$15

Help Delta maximize its profit on this flight.

%4 P&G is doing a discrete choice analysis to determine what price to charge for a box of Tide. It collected the data shown in Figure 18-12. For example, when people were asked which they prefer: Generic for \$5, Tide for \$8, or None, 35 people said generic, 22 picked Tide, and 43 chose None.

	F	G	H	I	J
5	Generic	P and G	Generic	P and G	None
6	\$ 5.00	\$ 8.00	35	22	43
7	\$ 5.50	\$ 7.00	29	29	42
8	\$ 7.00	\$ 9.00	24	18	59
9	\$ 7.50	\$ 8.00	17	28	33
10	\$ 0.85	\$ 10.00	56	4	28
11	\$ 4.00	\$ 7.00	37	26	28
12	\$ 3.50	\$ 6.50	38	28	27

Figure 18-12: Data for Exercise 2

Using a discrete choice model with the same price weight for each product, answer the following questions:

Using the value-based approach to pricing outlined in Chapter 16, “Conjoint Analysis,” what price premium can Tide command over the generic product?

If the generic product sold for \$5, what price would you recommend for Tide?

%4 P&G wants to determine if it should introduce a new, cheaper version of Head and Shoulders shampoo. It asked a focus group which of three products it would prefer at different prices (see Figure 18-13). For example, if all three products cost \$3.60, 70 people prefer Head and Shoulders, 13 prefer Head and Shoulders Lite, 4 prefer CVS, and 13 would buy no dan-druff shampoo.

Use this data to calibrate a discrete choice model. Use the same price weight for each product.

Suppose the CVS shampoo sells for \$3.00. If the unit cost to produce Head and Shoulders is \$2.20 and the unit cost to produce Head and Shoulders Lite is \$1.40, what pricing maximizes P&G’s profit?

By what percentage does the introduction of Head and Shoulders Lite increase P&G’s profit?

%4 Part IV: What do Customers Want?

Scenario	HS	HSLite	CVS	HS Share	HS lite share	CVS share	None share
1	3.6	3.6	3.6	70	13	4	13
2	3.8	2.8	2.6	59	22	7	13
3	4	3	2.7	58	21	7	14
4	4.2	3	2.7	54	23	8	15
5	4.4	3.1	2.6	51	23	9	17
6	4.6	2.9	2.5	46	27	10	17
7	4.8	3	2.8	44	28	9	19
8	5	3	2.9	41	30	9	20
9	5.2	3	2.6	37	31	11	21
10	5.4	3.1	3	35	32	9	23
11	5.6	3.3	2.8	33	30	12	25
12	5.8	3.5	3.1	32	29	11	28
13	6	3.7	3.2	31	28	11	31
14	6.2	4	3.2	29	25	12	34
15	6.4	4.5	3.4	29	20	12	39
16	6.6	4.7	3.6	28	19	11	42

Figure 18-13: Data for Exercise 3

%4 CVS wants to determine how to price Listerine with whitener (LW), Listerine (LIST), and CVS mouthwash. One hundred people were shown the following price scenarios, and their choices are listed in Figure 18-14. For example, with LW at \$6.60, LIST at \$5, and CVS at \$3.25, 37 picked no mouthwash, 18 chose LISTW, 20 preferred LIST, and 25 chose CVS.

	E	F	G	H	I	J	K
18	LISTW	LIST	CVS	Pick Nothing	Pick LISTW	Pick LIST	Pick CVS
19	\$6.60	\$5.00	\$3.25	37	18	20	25
20	\$7.00	\$5.80	\$3.50	40	17	17	25
21	\$6.90	\$6.00	\$3.75	41	18	17	24
22	\$6.80	\$5.90	\$3.10	38	18	16	27
23	\$8.40	\$6.75	\$4.25	46	14	16	24

Figure 18-14: Data for Exercise 4

%4 Fit a discrete choice model to this data. Use only a single price variable.

%4 Using Value-Based Pricing, estimate the value customers place on the whitening feature.

%4 Using Value-Based Pricing, estimate the brand equity of P&G.

NOTE LIST and CVS have the same features; the only difference is the name on the bottle.

%4 Suppose the price of CVS mouthwash must be \$6. Also assume CVS pays \$4 for a bottle of LISTW, \$3 for a bottle of LIST, and \$2.50 for a bottle of CVS. What price for LISTW and LIST maximizes profits for CVS?

%4 Armed Forces recruiting has asked for your help to allocate recruiting bonuses among the Air Force, Navy, and Army. During the next year, 1,000,000 people are expected to show interest in enlisting. The United States needs 100,000 enlistees in the Air Force and Navy and 250,000 enlistees in the Army. Recruiting bonuses of up to \$30,000 are allowed. A discrete choice study has been undertaken to determine the minimum cost recruiting budget that will obtain the correct number of enlistees. The information shown in Figure 18-15 is available:

	C	D	E	F	L	M	N	O
7		Thousands of \$s						
8								
9		Bonus	Bonus	Bonus				
10	Set	AF	NAVY	ARMY	No Enlist	AF	NAVY	ARMY
11	1	\$10.00	\$10.00	\$10.00	59	25	10	6
12	2	\$10.00	\$20.00	\$20.00	52	11	27	10
13	3	\$10.00	\$30.00	\$30.00	3	3	47	14
14	4	\$20.00	\$10.00	\$20.00	49	43	1	7
15	5	\$20.00	\$20.00	\$30.00	42	28	12	18
16	6	\$20.00	\$30.00	\$10.00	37	20	37	6
17	7	\$30.00	\$10.00	\$30.00	31	60	0	9
18	8	\$30.00	\$20.00	\$10.00	33	59	2	6
19	9	\$30.00	\$30.00	\$20.00	29	44	21	6

Figure 18-15: Data for Exercise 5

For example, 100 potential recruits were offered \$10,000 to enlist in each service. Fifty-nine chose not to enlist, 25 chose the Air Force, 10 chose the Navy, and 6 chose the Army.

%4 Develop a discrete choice model that can determine how the size of the bonuses influences the number of recruits each service obtains. Assume the following:

Score for Not Enlisting = Weight for Not Enlisting

Score for Each Service = Weight for Each Service + (Bonus Weight) * (Size of Bonus)

Use the same bonus weight for each service.

Assuming the bonus for each service can be at most \$30,000, determine the minimum cost bonus plan that will fill the U.S. recruiting quotas.

%4 Verify Equation 7.

