#### THOMAS W. MILLER

FACULTY DIRECTOR OF NORTHWESTERN UNIVERSITY'S PREDICTIVE ANALYTICS PROGRAM

# MODELING TECHNIQUES IN PREDICTIVE ANALYTICS

# BUSINESS PROBLEMS and SOLUTIONS with **R**

# Modeling Techniques in Predictive Analytics

**Business Problems and Solutions with R** 

THOMAS W. MILLER

Vice President, Publisher: Tim Moore Associate Publisher and Director of Marketing: Amy Neidlinger Executive Editor: Jeanne Glasser Operations Specialist: Jodi Kemper Cover Designer: Alan Clements Managing Editor: Kristy Hart Project Editor: Sara Schumacher Senior Compositor: Gloria Schurick Manufacturing Buyer: Dan Uhrig

© 2014 by Thomas W. Miller Published by Pearson Education, Inc. Upper Saddle River, New Jersey 07458

Pearson offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales. For more information, please contact U.S. Corporate and Government Sales, 1-800-382-3419, corpsales@pearsontechgroup.com. For sales outside the U.S., please contact International Sales at international@pearsoned.com.

Company and product names mentioned herein are the trademarks or registered trademarks of their respective owners.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

First Printing August 2013

ISBN-10: 0-13-341293-8 ISBN-13: 978-0-13-341293-2

Pearson Education LTD.

Pearson Education Australia PTY, Limited.

Pearson Education Singapore, Pte. Ltd.

Pearson Education Asia, Ltd.

Pearson Education Canada, Ltd.

Pearson Educacin de Mexico, S.A. de C.V.

Pearson Education—Japan

Pearson Education Malaysia, Pte. Ltd.

Library of Congress Control Number: 2013946325

# Contents

Pr	v	
Fi	gures	ix
Ta	bles	xiii
Exhibits		xv
1	Analytics and Data Science	1
2	Advertising and Promotion	15
3	Preference and Choice	29
4	Market Basket Analysis	37
5	Economic Data Analysis	53
6	Operations Management	67
7	Text Analytics	83
8	Sentiment Analysis	113
9	Sports Analytics	149

10	0 Brand and Price							
11	1 Spatial Data Analysis							
12	2 The Big Little Data Game							
А	There's a Pack' for That	237						
	A.1 Regression	. 238						
	A.2 Classification	. 240						
	A.3 Recommender Systems	. 242						
	A.4 Product Positioning	. 244						
	A.5 Segmentation and Target Marketing							
	A.6 Finance and Risk Analytics							
	A.7 Social Network Analysis	. 250						
В	Measurement 2.							
С	Code and Utilities 2							
Bib	Bibliography							
Inc	Index 32							

# Preface

"Toto, I've got a feeling we're not in Kansas anymore."

—JUDY GARLAND AS DOROTHY GALE IN *The Wizard of Oz* (1939)

Data and algorithms rule the day. Welcome to the new world of business, a fast-paced, data-intensive world, an open-source world in which competitive advantage, however fleeting, is obtained through analytic prowess and the sharing of ideas.

Many books about predictive analytics talk about strategy and management. Some focus on methods and models. Others look at information technology and code. This is that rare book that tries to do all three, appealing to modelers, programmers, and business managers alike.

We recognize the importance of analytics in gaining competitive advantage. We help researchers and analysts by providing a ready resource and reference guide for modeling techniques. We show programmers how to build upon a foundation of code that works to solve real business problems. We translate the results of models into words and pictures that management can understand. We explain the meaning of data and models.

Growth in the volume of data collected and stored, in the variety of data available for analysis, and in the rate at which data arrive and require analysis, makes analytics more important with every passing day. Achieving competitive advantage means implementing new systems for information management and analytics. It means changing the way business is done. Covering a variety of applications, this book is for people who want to know about data, modeling techniques, and the benefits of analytics. This book is for people who want to make things happen in their organizations.

Predictive analytics is data science. The literature in the field is massive, drawing from many academic disciplines and application areas. The relevant code (even if we restrict ourselves to R) is growing quickly. Indeed, it would be a challenge to provide a comprehensive guide to predictive analytics. What we have done is offer a collection of vignettes with each chapter focused on a particular application area and business problem.

Our objective is to provide an overview of predictive analytics and data science that is accessible to many readers. There is scant mathematics in the book—statisticians and modelers may look to the references for details and derivations of methods. We describe methods in plain English and use data visualization to show solutions to business problems.

Given the subject of the book, some might wonder if I belong to either the classical or Bayesian camp. At the School of Statistics at the University of Minnesota, I developed a respect for both sides of the classical/Bayesian divide. I regard highly the perspective of empirical Bayesians and those working in statistical learning, an area that combines machine learning and traditional statistics. I am a pragmatist when it comes to modeling and inference. I do what works and express my uncertainty in statements that others can understand.

What made this book possible is the work of thousands of experts across the world, people who contribute time and ideas to the R community. The growth of R and the ease of growing it further ensures that the R environment for modeling techniques in predictive analytics will be around for many years to come. Genie out of the lamp, wizard from behind the curtain—rocket science is not what it used to be. Secrets are being revealed. This book is part of the process.

Most of the data in the book were obtained from public domain data sources. Bobblehead promotional data were contributed by Erica Costello. Computer choice study data were made possible through work supported by Sharon Chamberlain. The call center data of "Anonymous Bank" were provided by Avi Mandelbaum and Ilan Guedj. Movie information was obtained courtesy of The Internet Movie Database, used with permission. IMDb movie reviews data were organized by Andrew L. Mass and his colleagues at Stanford University. Some examples were inspired by working with NCR Comten, Hewlett-Packard Company, Union Cab Cooperative of Madison, Site Analytics Co. of New York, and Sunseed Research LLC of Madison, Wisconsin.

As with vignettes under the Comprehensive R Archive Network, program examples in the book show what can be done with R. We work in a world of open source, sharing with one another. The truth about what we do is in programs for everyone to see and for some to debug. The code in this book contains step-by-step comments to promote student learning. Each program example ends with suggestions to build on the analysis that has been presented.

Many have influenced my intellectual development over the years. There were those good thinkers and good people, teachers and mentors for whom I will be forever grateful. Sadly, no longer with us are Gerald Hahn Hinkle in philosophy and Allan Lake Rice in languages at Ursinus College, and Herbert Feigl in philosophy at the University of Minnesota. I am also most thankful to David J. Weiss in psychometrics at the University of Minnesota and Kelly Eakin in economics, formerly at the University of Oregon. Good teachers—yes, great teachers—are valued for a lifetime.

Thanks to Michael L. Rothschild, Neal M. Ford, Peter R. Dickson, and Janet Christopher who provided invaluable support during our years together at the University of Wisconsin–Madison and the A.C. Nielsen Center for Marketing Research.

Those who know me well are not surprised by my move to the Los Angeles area. Two Major League Baseball teams, movies, and good weather is a hard combination to beat. I am most fortunate to be involved with graduate distance education at Northwestern University's School of Continuing Studies. Distance learning faculty and students at this school can live and work anywhere they like.

Thanks to Glen Fogerty who offered me the opportunity to teach and take a leadership role in the Predictive Analytics program at Northwestern University. Thanks to colleagues and staff who administer this exceptional graduate program. And thanks to the many students and fellow faculty from whom I have learned. Amy Hendrickson of T<sub>E</sub>Xnology Inc. applied her craft, making words, tables, and figures look beautiful in print—another victory for open source. Thanks to Donald Knuth and the T<sub>E</sub>X/ $\text{LAT}_{E}$ X community for their contributions to this wonderful system for typesetting and publication.

Thanks to readers and reviewers who provided much needed assistance, including Suzanne Callender, Philip M. Goldfeder, Melvin Ott, and Thomas P. Ryan. Jennifer Swartz provided proofreading assistance. Candice Bradley served dual roles as a reviewer and copyeditor. I am most grateful for their feedback and encouragement. Thanks to my editor, Jeanne Glasser Levine, and publisher, Pearson/FT Press, for making this book possible. Any writing issues, errors, or items of unfinished business, of course, are my responsibility alone.

My good friend Brittney and her daughter Janiya keep me company when time permits. And my son Daniel is there for me in good times and bad, a friend for life. My greatest debt is to them because they believe in me.

Thomas W. Miller Glendale, California July 2013

# Figures

1.1	Data and models for research	3
1.2	Training-and-Test Regimen for Model Evaluation	6
1.3	Training-and-Test Using Multi-fold Cross-validation	7
1.4	Training-and-Test with Bootstrap Resampling	8
1.5	Importance of Data Visualization: The Anscombe Quartet	10
2.1	Dodgers Attendance by Day of Week	18
2.2	Dodgers Attendance by Month	18
2.3	Dodgers Weather, Fireworks, and Attendance	19
2.4	Dodgers Attendance by Visiting Team	21
2.5	Regression Model Performance: Bobbleheads and Attendance	22
3.1	Spine Chart of Preferences for Mobile Communication Services	31
4.1	Market Basket for One Shopping Trip	38
4.2	Market Basket Prevalence of Initial Grocery Items	41
4.3	Market Basket Prevalence of Grocery Items by Category	43
4.4	Market Basket Association Rules: Scatter Plot	44
4.5	Market Basket Association Rules: Matrix Bubble Chart	45
4.6	Association Rules for a Local Farmer: A Network Diagram	47
5.1	Multiple Time Series of Economic Data	55
5.2	Horizon Plot of Indexed Economic Time Series	57
5.3	Forecast of National Civilian Employment Rate (percentage)	59
5.4	Forecast of Manufacturers' New Orders: Durable Goods (billions of	
	dollars)	59
5.5	Forecast of University of Michigan Index of Consumer Sentiment (1Q 1966 = 100)	60
5.6	Forecast of New Homes Sold (millions)	60
6.1	Call Center Operations for Monday	69
6.2	Call Center Operations for Tuesday	69
6.3	Call Center Operations for Wednesday	70

6.4	Call Center Operations for Thursday	70
6.5	Call Center Operations for Friday	71
6.6	Call Center Operations for Sunday	71
6.7	Call Center Arrival and Service Rates on Wednesdays	72
6.8	Call Center Needs and Optimal Workforce Schedule	75
7.1	Movie Taglines from The Internet Movie Database (IMDb)	84
7.2	Movies by Year of Release	86
7.3	A Bag of 200 Words from Forty Years of Movie Taglines	88
7.4	Picture of Text in Time: Forty Years of Movie Taglines	89
7.5	Text Measures and Documents on a Single Graph	90
7.6	Horizon Plot of Text Measures across Forty Years of Movie Taglines	92
7.7	From Text Processing to Text Analytics	93
7.8	Linguistic Foundations of Text Analytics	94
7.9	Creating a Terms-by-Documents Matrix	96
7.10	An R Programmer's Word Cloud	100
8.1	A Few Movie Reviews According to Tom	114
8.2	A Few More Movie Reviews According to Tom	115
8.3	Fifty Words of Sentiment	117
8.4	List-Based Text Measures for Four Movie Reviews	119
8.5	Scatter Plot of Text Measures of Positive and Negative Sentiment	120
8.6	Word Importance in Classifying Movie Reviews as Thumbs-Up or Thumbs-Down	124
8.7	A Simple Tree Classifier for Thumbs-Up or Thumbs-Down	125
9.1	Predictive Modeling Framework for Picking a Winning Team	150
9.2	Game-day Simulation (offense only)	156
9.3	Mets' Away and Yankees' Home Data (offense and defense)	157
9.4	Balanced Game-day Simulation (offense and defense)	158
9.5	Actual and Theoretical Runs-scored Distributions	160
9.6	Poisson Model for Mets vs. Yankees at Yankee Stadium	162
9.7	Negative Binomial Model for Mets vs. Yankees at Yankee Stadium	163
9.8	Probability of Home Team Winning (Negative Binomial Model)	165
10.1	Computer Choice Study: One Choice Set	176
10.2	Computer Choice Study: A Mosaic of Top Brands and Most Valued Attributes	179
10.3	Framework for Describing Consumer Preference and Choice	181
10.4	Ternary Plot of Consumer Preference and Choice	181
10.5	Comparing Consumers with Differing Brand Preferences	182
10.6	Potential for Brand Switching: Parallel Coordinates for Individual Consumers	184

10.7	Potential for Brand Switching: Parallel Coordinates for Consumer	
	Groups	185
10.8	Market Simulation: A Mosaic of Preference Shares	188
11.1	California Housing Data: Correlation Heat Map for the Training Data	213
11.2	California Housing Data: Scatter Plot Matrix of Selected Variables	214
11.3	Tree-Structured Regression for Predicting California Housing Values .	216
11.4	Random Forests Regression for Predicting California Housing Values .	217
12.1	From Data to Explanation	232
A.1	Evaluating Predictive Accuracy for a Binary Classifier	241
<b>B.1</b>	Hypothetical Multitrait-Multimethod Matrix	255
<b>B.2</b>	Conjoint Degree-of-Interest Rating	258
<b>B.3</b>	Conjoint Sliding Scale for Profile Pairs	258
<b>B.4</b>	Paired Comparisons	259
<b>B.5</b>	Multiple-Rank-Orders	259
<b>B.6</b>	Best-worst Item Provides Partial Paired Comparisons	260
<b>B.</b> 7	Paired Comparison Choice Task	262
<b>B.8</b>	Choice Set with Three Product Profiles	262
<b>B.9</b>	Menu-based Choice Task	264
<b>B.10</b>	Elimination Pick List	265

This page intentionally left blank

# Tables

1.1	Data for the Anscombe Quartet	9
2.1	Bobbleheads and Dodger Dogs	17
2.2	Regression of Attendance on Month, Day of Week, and Bobblehead Promotion	23
3.1	Preference Data for Mobile Communication Services	30
4.1	Association Rules for a Local Farmer	46
6.1	Call Center Shifts and Needs for Wednesdays	73
6.2	Call Center Problem and Solution	74
8.1	List-Based Sentiment Measures from Tom's Reviews	118
8.2	Accuracy of Text Classification for Movie Reviews (Thumbs-Up or Thumbs-Down)	122
8.3	Random Forest Text Measurement Model Applied to Tom's Movie Reviews	123
9.1	New York Mets' Early Season Games in 2007	153
9.2	New York Yankees' Early Season Games in 2007	154
10.1	Computer Choice Study: Product Attributes	175
10.2	Computer Choice Study: Data for One Individual	177
10.3	Contingency Table of Top-ranked Brands and Most Valued Attributes	180
10.4	Market Simulation: Choice Set Input	187
10.5	Market Simulation: Preference Shares in a Hypothetical Four-brand Market	189
11 1	California Housing Data: Original and Computed Variables	211
11.2	Linear Regression Fit to Selected California Block Groups	215
11.3	Comparison of Regressions on Spatially Referenced Data	218
11.0	comparison of regressions on spanning referenced Data	210

This page intentionally left blank

# **Exhibits**

1.1	R Program for the Anscombe Quartet	13
2.1	Shaking Our Bobbleheads Yes and No	26
3.1	Measuring and Modeling Individual Preferences	34
4.1	Market Basket Analysis of Grocery Store Data	50
5.1	Working with Economic Data	62
6.1	Call Center Scheduling Problem and Solution	77
7.1	Text Analytics of Movie Taglines	101
8.1	Sentiment Analysis and Classification of Movie Ratings	129
9.1	Winning Probabilities by Simulation (Negative Binomial Model)	171
10.1	Computer Choice Study: Training and Testing with Hierarchical Bayes	192
10.2	Preference, Choice, and Market Simulation	197
11.1	California Housing Values: Regression and Spatial Regression Models	220
C.1	Conjoint Analysis Spine Chart	268
C.2	Market Simulation Utilities	276
C.3	Split-plotting Utilities	277
<b>C.4</b>	Wait-time Ribbon Plot	280
C.5	Word Scoring Code for Sentiment Analysis	292
C.6	Utilities for Spatial Data Analysis	296

This page intentionally left blank

# **Analytics and Data Science**

Mr. Maguire: "I just want to say one word to you, just one word."

Ben: "Yes, sir."

Mr. Maguire: "Are you listening?"

Ben: "Yes, I am."

Mr. Maguire: "Plastics."

—Walter Brooke as Mr. Maguire and Dustin Hoffman as Ben (Benjamin Braddock) in *The Graduate* (1967)

While earning a degree in philosophy may not be the best career move (unless a student plans to teach philosophy, and few of these positions are available), I greatly value my years as a student of philosophy and the liberal arts. For my bachelor's degree, I wrote an honors paper on Bertrand Russell. In graduate school at the University of Minnesota, I took courses from one of the truly great philosophers, Herbert Feigl. I read about science and the search for truth, otherwise known as epistemology. My favorite philosophy was logical empiricism.

Although my days of "thinking about thinking" (which is how Feigl defined philosophy) are far behind me, in those early years of academic training I was able to develop a keen sense for what is real and what is just talk. When we use the word *model* in predictive analytics, we are referring to a representation of the world, a rendering or description of reality, an attempt to relate one set of variables to another. Limited, imprecise, but useful, a model helps us to make sense of the world.

Predictive analytics brings together management, information technology, and modeling. It is for today's data-intensive world. Predictive analytics is data science, a multidisciplinary skill set essential for success in business, nonprofit organizations, and government. Whether forecasting sales or market share, finding a good retail site or investment opportunity, identifying consumer segments and target markets, or assessing the potential of new products or risks associated with existing products, modeling methods in predictive analytics provide the key.

Data scientists, those working in the field of predictive analytics, speak the language of business—accounting, finance, marketing, and management. They know about information technology, including data structures, algorithms, and object-oriented programming. They understand statistical modeling, machine learning, and mathematical programming. Data scientists are methodological eclectics, drawing from many scientific disciplines and translating the results of empirical research into words and pictures that management can understand.

Predictive analytics, like much of statistics, involves searching for meaningful relationships among variables and representing those relationships in models. There are response variables—things we are trying to predict. There are explanatory variables or predictors—things we observe, manipulate, or control that could relate to the response.

Regression and classification are two common types predictive models. Regression involves predicting a response with meaningful magnitude, such as quantity sold, stock price, or return on investment. Classification involves predicting a categorical response. Which brand will be purchased? Will the consumer buy the product or not? Will the account holder pay off or default on the loan? Is this bank transaction true or fraudulent?

Predictive modeling involves searching for useful predictors. Prediction problems are defined by their width or number of potential predictors and their depth or number of observations or cases in the data set. It is the number of potential predictors in business, marketing, and investment analysis

3



Figure 1.1. Data and models for research

that causes the most difficulty. There can be thousands of potential predictors with weak relationships to the response. With the aid of computers, hundreds or thousands of models can be fit to subsets of the data and tested on other subsets of the data, providing an evaluation of each predictor.

Predictive modeling involves finding good subsets of predictors or explanatory variables. Models that fit the data well are better than models that fit the data poorly. Simple models are better than complex models. Working with a list of useful predictors, we can fit many models to the available data, then evaluate those models by their simplicity and by how well they fit the data.

Consider three general approaches to research and modeling as employed in predictive analytics: traditional, data-adaptive, and model-dependent. See figure 1.1. The traditional approach to research and modeling begins with the specification of a theory or model. Classical or Bayesian methods of statistical inference are employed. Traditional methods, such as linear regression and logistic regression, estimate parameters for linear predictors. Model building involves fitting models to data. After we have fit a model, we can check it using model diagnostics.

When we employ a data-adaptive approach, we begin with data and search through those data to find useful predictors. We give little thought to the-

ories or hypotheses prior to running the analysis. This is the world of machine learning, sometimes called statistical learning or data mining. Dataadaptive methods adapt to the available data, representing nonlinear relationships and interactions among variables. The data determine the model. Data-adaptive methods are data-driven.

Model-dependent research is the third approach. It begins with the specification of a model and uses that model to generate data, predictions, or recommendations. Simulations and mathematical programming methods, primary tools of operations research, are examples of model-dependent research. When employing a model-dependent or simulation approach, models are improved by comparing generated data with real data. We ask whether simulated consumers, firms, and markets behave like real consumers, firms, and markets.

It is often a combination of models and methods that works best. Consider an application from the field of financial research. The manager of a mutual fund is looking for additional stocks for a fund's portfolio. A financial engineer employs a data-adaptive model (perhaps a neural network) to search across thousands of performance indictors and stocks, identifying a subset of stocks for further analysis. Then, working with that subset of stocks, the financial engineer employs a theory-based approach (CAPM, the capital asset pricing model) to identify a smaller set of stocks to recommend to the fund manager. As a final step, using model-dependent research (mathematical programming), the engineer identifies the minimum-risk capital investment for each of the stocks in the portfolio.

Data may be organized by observational unit, time, and space. The observational or cross-sectional unit could be an individual consumer or business or any other basis for collecting and grouping data. Data are organized in time by seconds, minutes, hours, days, and so on. Space or location is often defined by longitude and latitude.

Consider numbers of customers entering grocery stores (units of analysis) in Glendale, California on Monday (one point in time), ignoring the spatial location of the stores—these are cross-sectional data. Suppose we work with one of those stores, looking at numbers of customers entering the store each day of the week for six months—these are time series data. Then we look at numbers of customers at all of the grocery stores in Glendale

5

across six months—these are longitudinal or panel data. To complete our study, we locate these stores by longitude and latitude, so we have spatial or spatio-temporal data. For any of these data structures we could consider measures in addition to the number of customers entering stores. We look at store sales, consumer or nearby resident demographics, traffic on Glendale streets, and so doing move to multiple time series and multivariate methods. The organization of the data we collect affects the structure of the models we employ.

As we consider business problems in this book, we touch on many types of models, including cross-sectional, time series, and spatial data models. Whatever the structure of the data and associated models, prediction is the unifying theme. We use the data we have to predict data we do not yet have, recognizing that prediction is a precarious enterprise. It is the process of extrapolating and forecasting.

To make predictions, we may employ classical or Bayesian methods. Or we may dispense with parametric formulations entirely and rely upon machine learning algorithms. We do what works.<sup>1</sup> Our approach to predictive analytics is based upon a simple premise:

#### The value of a model lies in the quality of its predictions.

We learn from statistics that we should quantify our uncertainty. On the one hand, we have confidence intervals, point estimates with associated standard errors, and significance tests—that is the classical way. On the other hand, we have probability intervals, prediction intervals, Bayes factors, subjective (perhaps diffuse) priors, and posterior probability distributions the path of Bayesian statistics. Indices like the Akaike information criterion (AIC) or the Bayes information criterion (BIC) help us to to judge one model against another, providing a balance between goodness-of-fit and parsimony.

Central to our approach is a *training-and-test regimen*. We partition sample data into training and test sets. We build our model on the training set and

<sup>&</sup>lt;sup>1</sup> Within the statistical literature, Seymour Geisser (1929–2004) introduced an approach best described as *Bayesian predictive inference* (Geisser 1993). Bayesian statistics is named after Reverend Thomas Bayes (1706–1761), the creator of Bayes Theorem. In our emphasis upon the success of predictions, we are in agreement with Geisser. Our approach, however, is purely empirical and in no way dependent upon classical or Bayesian thinking.



Figure 1.2. Training-and-Test Regimen for Model Evaluation

evaluate it on the test set. Simple two- and three-way data partitioning are shown in figure 1.2.

A random splitting of a sample into training and test sets could be fortuitous, especially when working with small data sets, so we sometimes conduct statistical experiments by executing a number of random splits and averaging performance indices from the resulting test sets. There are extensions to and variations on the training-and-test theme.

One variation on the training-and-test theme is multi-fold cross-validation, illustrated in figure 1.3. We partition the sample data into *M* folds of approximately equal size and conduct a series of tests. For the five-fold cross-validation shown in the figure, we would first train on sets *B* through *E* and test on set *A*. Then we would train on sets *A* and *C* through *E*, and test on *B*. We continue until each of the five folds has been utilized as a test set. We assess performance by averaging across the test sets. In leave-one-out cross-valuation, the logical extreme of multi-fold cross-validation, there are as many test sets as there are observations in the sample.

7

Randomly divide the sample into folds of approximately equal size: С В D Е А Each fold serves once as a test fold: Iteration 1 Test Train Train Train Train Iteration 2 Train Train Train Train Test **Iteration 3** Train Train Test Train Train Iteration 4 Train Train Train Test Train Iteration 5 Train Train Train Train Test

Figure 1.3. Training-and-Test Using Multi-fold Cross-validation



Figure 1.4. Training-and-Test with Bootstrap Resampling

Another variation on the training-and-test regimen is the class of bootstrap methods. If a sample approximates the population from which it was drawn, then a sample from the sample (what is known as a resample) also approximates the population. A bootstrap procedure, as illustrated in figure 1.4, involves repeated resampling with replacement. That is, we take many random samples with replacement from the sample, and for each of these resamples, we compute a statistic of interest. The bootstrap distribution of the statistic approximates the sampling distribution of that statistic. What is the value of the bootstrap? It frees us from having to make assumptions about the population distribution. We can estimate standard errors and make probability statements working from the sample data alone. The bootstrap may also be employed to improve estimates of prediction error within a leave-one-out cross-validation process. Cross-validation and bootstrap methods are reviewed in Davison and Hinkley (1997), Efron and Tibshirani (1993), and Hastie, Tibshirani, and Friedman (2009).

9

Set I			Set II		Set III		Set IV	
x	у	x	У	x	У	x	у	
10	8.04	10	9.14	10	7.46	8	6.58	
8	6.95	8	8.14	8	6.77	8	5.76	
13	7.58	13	8.74	13	12.74	8	7.71	
9	8.81	9	8.77	9	7.11	8	8.84	
11	8.33	11	9.26	11	7.81	8	8.47	
14	9.96	14	8.10	14	8.84	8	7.04	
6	7.24	6	6.13	6	6.08	8	5.25	
4	4.26	4	3.10	4	5.39	19	12.50	
12	10.84	12	9.13	12	8.15	8	5.56	
7	4.82	7	7.26	7	6.42	8	7.91	
5	5.68	5	4.74	5	5.73	8	6.89	

Table 1.1. Data for the Anscombe Quartet

Data visualization is critical to the work of data science. Examples in this book demonstrate the importance of data visualization in discovery, diagnostics, and design. We employ tools of exploratory data analysis (discovery) and statistical modeling (diagnostics). In communicating results to management, we use presentation graphics (design).

There is no more telling demonstration of the importance of statistical graphics and data visualization than a demonstration that is affectionately known as the Anscombe Quartet. Consider the data sets in table 1.1, developed by Anscombe (1973). Looking at these tabulated data, the casual reader will note that the fourth data set is clearly different from the others. What about the first three data sets? Are there obvious differences in patterns of relationship between *x* and *y*?

When we regress y on x for the data sets, we see that the models provide similar statistical summaries. The mean of the response y is 7.5, the mean of the explanatory variable x is 9. The regression analyses for the four data sets are virtually identical. The fitted regression equation for each of the four sets is  $\hat{y} = 3 + 0.5x$ . The proportion of response variance accounted for is 0.67 for each of the four models.

Anscombe (1973) argues that statistical summaries do not tell the story of data. It is not sufficient to look at data tables, regression coefficients, and the results of statistical tests. As the plots in figure 1.5 clearly show, the four Anscombe data sets are very different from one another.



Figure 1.5. Importance of Data Visualization: The Anscombe Quartet

The Anscombe Quartet shows that we must look at data to understand data. The program for the Anscombe Quartet demonstration is provided in exhibit 1.1 at the end of this chapter. The program uses standard R graphics.

Visualization tools help us learn from data . We explore data, discover patterns in data, identify groups of observations that go together and unusual observations or outliers. We note relationships among variables, sometimes detecting underlying dimensions in the data.

Graphics for exploratory data analysis are reviewed in classic references by Tukey (1977) and Tukey and Mosteller (1977). Regression graphics are covered by Cook (1998), Cook and Weisberg (1999), and Fox and Weisberg (2011). Statistical graphics and data visualization are illustrated in the works of Tufte (1990, 1997, 2004, 2006), Few (2009), and Yau (2011, 2013). Wilkinson (2005) presents a review of human perception and graphics, as well as a conceptual structure for understanding statistical graphics. Heer, Bostock, and Ogievetsky (2010) provide a collection of contemporary visualization techniques.

The R programming environment provides a rich collection of open-source tools for data visualization, including interfaces to visualization applications on the World Wide Web. An R graphics overview is provided by Murrell (2011). R lattice graphics, discussed by Sarkar (2008, 2013), build upon the conceptual structure of an earlier system called S-Plus Trellis<sup>TM</sup> (Cleveland 1993; Becker and Cleveland 1996). Wilkinson's (2005) "grammar of graphics" approach has been implemented in the R ggplot2 package (Wickham and Chang 2013), with programming examples provided by Chang (2013).

Zeileis, Hornik, and Murrell (2009, 2013) provide advice about colors for statistical graphics. Ihaka et al. (2013) show how to specify colors in R by hue, chroma, and luminance.

Specialized techniques for data visualization may be needed when working with very large data sets as we often do in predictive analytics (Unwin, Theus, and Hofmann 2006). Partial transparency techniques can help, and hexbin plots are often better than scatter plots for showing relationships between variables (Carr, Lewin-Koh, and Maechler 2013; Lewin-Koh 2013).

These are the things that data scientists do—things that we discuss in this book:

- Finding out about. This is the first thing we do—information search, finding what others have done before, learning from the literature. We draw on the work of academics and practitioners in many fields of study, contributors to predictive analytics and data science.<sup>2</sup>
- Looking at data. We begin each modeling project with exploratory data analysis, data visualization for the purpose of discovery. We prepare data for further analysis.
- Predicting how much. We are often asked to predict how many units or dollars of product will be sold, the price of financial securities or real estate. Regression techniques are useful for making these predictions. This book begins and ends with regression problems.
- Predicting yes or no. Many business problems are classification problems. Predicting whether a person will buy or not, default on a loan or not, click on a Web page or not. We use classification techniques in predicting consumer behavior and in text classification.
- **Testing it out.** We examine models with diagnostic graphics. We see how well a model developed on one data set works on other data sets. We employ a training-and-test regimen with data partitioning, cross-validation, or bootstrap methods.
- Playing what-if. We may need to manipulate key variables to see what happens to our predictions. In pricing research we play whatif games in a simulated marketplace. In operations management we employ sensitivity testing of mathematical programming models. We see how new values for input variables affect outcomes or payoffs.
- **Explaining it all.** Data and models help us to understand the world. We turn what we have learned into an explanation that others can understand. We present project results in a clear and concise manner. These presentations benefit from well constructed data visualizations (design with a purpose).

Let us begin.

<sup>&</sup>lt;sup>2</sup> We start with the title of Richard Belew's (2000) book—*Finding Out About*—at the top of the list. This title describes well the task of information search.

Exhibit 1.1. R Program for the Anscombe Quartet

```
# The Anscombe Quartet in R
# demonstration data from
# Anscombe, F. J. 1973, February. Graphs in statistical analysis.
# The American Statistician 27: 1721.
# define the anscombe data frame
anscombe <- data.frame(</pre>
   x1 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
  x^2 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
  x3 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
   x4 = c(8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8),
  y_1 = c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68),
  y_2 = c(9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74),
  y_3 = c(7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73),
   y4 = c(6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89))
# show results from four regression analyses
with(anscombe, print(summary(lm(y1 ~ x1))))
with(anscombe, print(summary(lm(y2 ~ x2))))
with(anscombe, print(summary(lm(y3 ~ x3))))
with(anscombe, print(summary(lm(y4 ~ x4))))
# place four plots on one page using standard R graphics
# ensuring that all have the same scales
# for horizontal and vertical axes
pdf(file = "fig_more_anscombe.pdf", width = 8.5, height = 8.5)
par(mfrow=c(2,2),mar=c(3,3,3,1))
plot(x1, y1, xlim=c(2,20),ylim=c(2,14),
  pch = 19, col = "darkblue", cex = 2, las = 1)
title("Set I")
plot(x2, y2, xlim=c(2,20),ylim=c(2,14),
  pch = 19, col = "darkblue", cex = 2, las = 1)
title("Set II")
plot(x3, y3, xlim=c(2,20),ylim=c(2,14),
  pch = 19, col = "darkblue", cex = 2, las = 1)
title("Set III")
plot(x4, y4, xlim=c(2,20),ylim=c(2,14),
  pch = 19, col = "darkblue", cex = 2, las = 1)
title("Set IV")
dev.off()
par(mfrow=c(1,1),mar=c(5.1, 4.1, 4.1, 2.1)) # return to plotting defaults
# suggestions for the student
# see if you can develop a quartet of your own
# or perhaps just a duet...
# two very different data sets with the same fitted model
```

This page intentionally left blank

This page intentionally left blank

# Index

#### Α

advertising, 15–29 Akaike information criterion (AIC), 5 ARIMA model, *see* time series analysis association rule, 40–42

#### В

bag-of-words approach, *see* text analytics bar chart, *see* data visualization Bayes information criterion (BIC), 5 Bayesian statistics, 5, 178, 191, 233, 248 benchmark study, *see* simulation best-worst scaling, 260 biclustering, 248 big data, 231–235 biplot, *see* data visualization block clustering, *see* biclustering bootstrap, 8 box plot, *see* data visualization brand equity research, 173–209 bubble chart, *see* data visualization

### С

call center scheduling, *see* workforce scheduling choice study, 29 menu-based, 264 classical statistics, 5 classification, 2, 12, 113, 122, 240 accuracy, 240, 241 classification tree, *see* tree-structured model cluster analysis, 99, 248, 251 coefficient of determination, 239 collaborative filtering, 243 computational linguistics, *see* text analytics, natural language processing conjoint analysis, 33, 258 content analysis, 126 corpus, *see* text analytics correlation heat map, *see* data visualization, heat map credit scoring, 249 cross-sectional study, *see* data organization cross-validation, 6, 240

### D

data mining, see data-adaptive research data organization, 5, 58 data science, 1-14 data visualization, 8-13 bar chart, 41, 43 biplot, 90 box plot, 16, 18 bubble chart, 45 density plot, 180, 182 diagnostics, 22, 239 dot chart, 124, 217 heat map, 164, 165, 212, 213 histogram, 86, 157, 160, 162, 163 horizon plot, 54, 56, 57, 91, 92 lattice plot, 11, 16, 19, 21, 22 line graph, 72, 75 mosaic plot, 178, 179 multiple time series plot, 55 network diagram, 47 parallel coordinates, 183, 185 ribbon plot, 68-71, 280 scatter plot, 44 scatter plot matrix, 212, 214 spine chart, 30, 31, 33, 34, 268 strip plot, 16, 21 ternary plot, 180, 181 time series plot, 59, 60 tree diagram, 125, 216 data-adaptive research, 3, 4 density plot, see data visualization dependent variable, see response

discrete event simulation, *see* simulation, process simulation dot chart, *see* data visualization duration analysis, *see* survival analysis

#### Ε

economic analysis, 53–66 indexing, 54 elimination pick list, 265 empirical Bayes, 233, *see* Bayesian statistics Erlang *C*, *see* queueing model explanatory variable, 2, 3, 238 exploratory data analysis, 16

### F

financial data analysis, 249 forecasting, 58–61, 218 four Ps, *see* marketing mix model

# G

game-day simulation, *see* simulation, gameday General Inquirer, 126 generalized least-squares, 239 generalized linear model, 239 generative grammar, *see* text analytics geographically weighted regression, 216 graphics, *see* data visualization group filtering, *see* collaborative filtering

# Η

hierarchical Bayes, *see* Bayesian statistics hierarchical models, 233 histogram, *see* data visualization horizon plot, *see* data visualization hybrid model, 216

#### I

independent variable, *see* explanatory variable integer programming, *see* mathematical programming item analysis, psychometrics, 121

# Κ

Kappa, see classification, accuracy

#### L

lattice plot, *see* data visualization leading indicator, 54, 61 least-squares regression, *see* regression lexical table, *see* text analytics, terms-bydocuments matrix line graph, see data visualization linear discriminant analysis, 240 linear least-squares regression, see regression linear model, 238, 239 linear predictor, 238 linguistics, see text analytics, natural language processing log-linear models, 251 logical empiricism, 1 logistic regression, 121, 238 longitudinal study, see data organization

#### Μ

machine learning, see data-adaptive research market basket analysis, 37-52 market response model, 25 market segmentation, see segmentation market simulation, see simulation marketing mix model, 24 mathematical programming, 67, 75, 249 integer programming, 74 sensitivity testing, 75 matrix bubble chart, see data visualization, bubble chart mean-squared error (MSE), see root meansquared error (RMSE) measurement, 253-266 construct validity, 253 content validity, 127 convergent validity, 254 discriminant validity, 254 face validity, 127 multitrait-multimethod matrix, 253, 255 reliability, 253 meta-analysis, 234 metadata, see text analytics model-dependent research, 3, 4 morphology, see text analytics mosaic plot, see data visualization multidimensional scaling, 87, 89, 99, 244, 245, 251multilevel models, see hierarchical models multiple time series plot, see data visualization, time series plot multivariate methods, 99, 244

### Ν

natural language processing, *see* text analytics nearest-neighbor model, 218, 219, 243 network diagram, *see* data visualization neural network, 240

# 0

operations management, 67–83 optimization, constrained, 74 organization of data, *see* data, organization

#### Ρ

paired comparisons, 259, 262 parallel coordinates plot, see data visualization parsing, see text analytics, text parsing perceptual map, see data visualization philosophy, 1 Poisson regression, 238 predictive analytics, 1-14 definition, 2 predictor, see explanatory variable preference scaling, 245 preference study, 29 pricing research, 173-209 principal component analysis, 244 privacy, 251 probability binomial distribution, 159 negative binomial distribution, 159, 161, 164 Poisson distribution, 159, 161, 164 probability heat map, see data visualization, heat map process simulation, see simulation product positioning, 244-245 promotion, 15–29

# Q

queueing model, 67, 68, 73

### R

```
R package
   arules, 50
   arulesViz, 50
   car, 26
   caret, 129, 197
   ChoiceModelR, 197
   cluster, 101
   cvTools, 220
   e1071, 129
   forecast, 62
   ggplot2, 77, 101, 129, 197
   grid, 77, 101, 129
   lattice, 26, 197, 220
   latticeExtra, 62, 101, 129
   lmtest, 62
   lpSolve, 77
   lstringr, 129
   lubridate, 62, 77
```

lvcd, 197 mapproj, 220 maps, 220 quantmod, 62 queueing, 77 randomForest, 129, 220 RColorBrewer, 50 rpart, 129, 220 rpart.plot, 129, 220 spgwr, 220 stringr, 101 tm, 101, 129 wordcloud, 101 R-squared, 239 random forest, 122-124, 212, 217 recommender systems, 242-243 regression, 2, 12, 20, 23, 24, 121, 212, 215, 238-239 robust methods, 239 time series regression, 58 regression tree, see tree-structured model reliability, see measurement response, 2, 238 ribbon plot, see data visualization risk analytics, 249 robust methods, see regression ROC curve, see classification, accuracy root mean-squared error (RMSE), 239

# S

sales forecasting, see forecasting scatter plot, see data visualization scatter plot matrix, see data visualization segmentation, 246-248 semantics, see text analytics sentiment analysis, 113-148 simulation, 151, 152, 155, 240, 249 benchmark study, 122, 216, 240 game-day, 150, 152, 155, 156 market simulation, 183, 186, 187 process simulation, 67, 68, 75, 76 what-if analysis, 12 site selection, 218, see spatial data analysis social filtering, see collaborative filtering social network analysis, 250-251 spatial data analysis, 209-230 spatio-temporal model, 210, 219 spine chart, see data visualization sports analytics, 149-172 state space model, see time series analysis statistical experiment, see simulation statistical graphics, see data visualization statistical learning, see data-adaptive research statistical simulation, see simulation

strip plot, *see* data visualization supervised learning, 97, 248 support vector machines, 122, 240 survey research, 266 survival analysis, 238, 249 syntax, *see* text analytics

# Т

tag, see text analytics, metadata target marketing, 246-248 terms-by-documents matrix, see text analytics ternary plot, see data visualization text analytics, 83-113 bag-of-words approach, 86, 91 corpus, 87 generative grammar, 93, 94 metadata, 85 morphology, 94 natural language processing, 86, 91, 93, 128 semantics, 94 stemming, 95 syntax, 94 terms-by-documents matrix, 87, 95, 96 text parsing, 85, 93 text summarization, 97 text measure, 85, 86, 91, 126, 127 text mining, see text analytics thematic analysis, 126

time series analysis, 53 ARIMA model, 58 multiple time series, 55 state space model, 58 time series plot, *see* data visualization traditional research, 3 training-and-test regimen, 6–8, 20 tree diagram, *see* data visualization tree-structured model classification, 123, 125 regression, 212, 216 trellis plot, *see* data visualization, lattice plot triplot, *see* data visualization, ternary plot

# U

unsupervised learning, 97, 248

#### V

validity, see measurement

#### W

wait-time ribbon, *see* data visualization, ribbon plot Web analytics, 251 Weka, 49, 240 what-if analysis, *see* simulation workforce scheduling, 67–83