

THOMAS W. MILLER

FACULTY DIRECTOR OF NORTHWESTERN UNIVERSITY'S
PREDICTIVE ANALYTICS PROGRAM

MODELING
TECHNIQUES
— IN —
PREDICTIVE
ANALYTICS

BUSINESS PROBLEMS
AND SOLUTIONS WITH R

REVISED AND EXPANDED EDITION

Modeling Techniques in Predictive Analytics

Business Problems and Solutions with R

Revised and Expanded Edition

THOMAS W. MILLER

Associate Publisher: Amy Neidlinger
Executive Editor: Jeanne Glasser
Operations Specialist: Jodi Kemper
Cover Designer: Alan Clements
Managing Editor: Kristy Hart
Project Editor: Andy Beaster
Senior Compositor: Gloria Schurick
Manufacturing Buyer: Dan Uhrig

©2015 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 October 2014

ISBN-10: 0-13-388601-8

ISBN-13: 978-0-13-388601-6

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: 2014948912

Contents

Preface	v
Figures	ix
Tables	xiii
Exhibits	xv
1 Analytics and Data Science	1
2 Advertising and Promotion	14
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	107
9 Sports Analytics	143

10	Spatial Data Analysis	167
11	Brand and Price	187
12	The Big Little Data Game	221
A	Data Science Methods	225
A.1	Databases and Data Preparation	227
A.2	Classical and Bayesian Statistics	229
A.3	Regression and Classification	232
A.4	Machine Learning	237
A.5	Web and Social Network Analysis	239
A.6	Recommender Systems	241
A.7	Product Positioning	243
A.8	Market Segmentation	245
A.9	Site Selection	247
A.10	Financial Data Science	248
B	Measurement	249
C	Case Studies	263
C.1	Return of the Bobbleheads	263
C.2	DriveTime Sedans	264
C.3	Two Month's Salary	269
C.4	Wisconsin Dells	273
C.5	Computer Choice Study	278
D	Code and Utilities	283
	Bibliography	321
	Index	355

Preface

“Toto, I’ve 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 environment in which competitive advantage, however fleeting, is obtained through analytic prowess and the sharing of ideas.

Many books about predictive analytics or data science talk about strategy and management. Some focus on methods and models. Others look at information technology and code. This is a rare book does all three, appealing to business managers, modelers, and programmers 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 each passing day. Achieving competitive advantage means implementing new systems for information management and analytics. It means changing the way business is done.

Literature in the field of data science is massive, drawing from many academic disciplines and application areas. The relevant open-source code is growing quickly. Indeed, it would be a challenge to provide a comprehensive guide to predictive analytics or data science.

We look at real problems and real data. We offer a collection of vignettes with each chapter focused on a particular application area and business problem. We provide solutions that make sense. By showing modeling techniques and programming tools in action, we convert abstract concepts into concrete examples. Fully worked examples facilitate understanding.

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 have high regard for the perspective of empirical Bayesians and those working in statistical learning, which 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.

This book is possible because of the thousands of experts across the world, people who contribute time and ideas to open source. The growth of open source and the ease of growing it further ensures that developed solutions 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. Major League Baseball data for promotions and attendance 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 clients at ToutBay of Tampa, Florida, NCR Comten, Hewlett-Packard Company, Site Analytics Co. of New York, Sunseed Research of Madison, Wisconsin, and Union Cab Cooperative of Madison.

We work within open-source communities, sharing code with one another. The truth about what we do is in the programs we write. It is there for everyone to see and for some to debug. To promote student learning, each program includes step-by-step comments and suggestions for taking the analysis further. All data sets and computer programs are downloadable from the book's website at <http://www.ftpress.com/miller/>.

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.

I live in California, four miles north of Dodger Stadium, teach for Northwestern University in Evanston, Illinois, and direct product development at ToutBay, a data science firm in Tampa, Florida. Such are the benefits of a good Internet connection.

I am fortunate to be involved with graduate distance education at Northwestern University's School of Professional Studies. 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.

ToutBay is an emerging firm in the data science space. With co-founder Greg Blence, I have great hopes for growth in the coming years. Thanks

to Greg for joining me in this effort and for keeping me grounded in the practical needs of business. Academics and data science models can take us only so far. Eventually, to make a difference, we must implement our ideas and models, sharing them with one another.

Amy Hendrickson of T_EXnology 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_EX/L^AT_EX community for their contributions to this wonderful system for typesetting and publication.

Thanks to readers and reviewers of the initial edition of the book, including Suzanne Callender, Philip M. Goldfeder, Melvin Ott, and Thomas P. Ryan. In preparing this revised edition, Lorena Martin provided much needed feedback and suggestions for improving the book. Candice Bradley served dual roles as a reviewer and copyeditor. Roy L. Sanford provided technical advice about statistical models and programs. I am most grateful for their feedback and encouragement. Thanks also 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
August 2014

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	17
2.2	Dodgers Attendance by Month	17
2.3	Dodgers Weather, Fireworks, and Attendance	18
2.4	Dodgers Attendance by Visiting Team	19
2.5	Regression Model Performance: Bobbleheads and Attendance	21
3.1	Spine Chart of Preferences for Mobile Communication Services	32
4.1	Market Basket Prevalence of Initial Grocery Items	41
4.2	Market Basket Prevalence of Grocery Items by Category	43
4.3	Market Basket Association Rules: Scatter Plot	44
4.4	Market Basket Association Rules: Matrix Bubble Chart	45
4.5	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
8.1	A Few Movie Reviews According to Tom	108
8.2	A Few More Movie Reviews According to Tom	109
8.3	Fifty Words of Sentiment	111
8.4	List-Based Text Measures for Four Movie Reviews	113
8.5	Scatter Plot of Text Measures of Positive and Negative Sentiment	114
8.6	Word Importance in Classifying Movie Reviews as Thumbs-Up or Thumbs-Down	118
8.7	A Simple Tree Classifier for Thumbs-Up or Thumbs-Down	119
9.1	Predictive Modeling Framework for Picking a Winning Team	144
9.2	Game-day Simulation (offense only)	150
9.3	Mets' Away and Yankees' Home Data (offense and defense)	151
9.4	Balanced Game-day Simulation (offense and defense)	152
9.5	Actual and Theoretical Runs-scored Distributions	154
9.6	Poisson Model for Mets vs. Yankees at Yankee Stadium	156
9.7	Negative Binomial Model for Mets vs. Yankees at Yankee Stadium	157
9.8	Probability of Home Team Winning (Negative Binomial Model)	159
10.1	California Housing Data: Correlation Heat Map for the Training Data	171
10.2	California Housing Data: Scatter Plot Matrix of Selected Variables	172
10.3	Tree-Structured Regression for Predicting California Housing Values	174
10.4	Random Forests Regression for Predicting California Housing Values	175
11.1	Computer Choice Study: A Mosaic of Top Brands and Most Valued Attributes	190
11.2	Framework for Describing Consumer Preference and Choice	192

11.3	Ternary Plot of Consumer Preference and Choice	192
11.4	Comparing Consumers with Differing Brand Preferences	193
11.5	Potential for Brand Switching: Parallel Coordinates for Individual Consumers	195
11.6	Potential for Brand Switching: Parallel Coordinates for Consumer Groups	196
11.7	Market Simulation: A Mosaic of Preference Shares	199
12.1	Work of Data Science	222
A.1	Evaluating Predictive Accuracy of a Binary Classifier	234
B.1	Hypothetical Multitrait-Multimethod Matrix	251
B.2	Conjoint Degree-of-Interest Rating	254
B.3	Conjoint Sliding Scale for Profile Pairs	254
B.4	Paired Comparisons	255
B.5	Multiple-Rank-Orders	255
B.6	Best-worst Item Provides Partial Paired Comparisons	256
B.7	Paired Comparison Choice Task	258
B.8	Choice Set with Three Product Profiles	258
B.9	Menu-based Choice Task	260
B.10	Elimination Pick List	261
C.1	Computer Choice Study: One Choice Set	280
D.1	An R Programmer's Word Cloud	285

This page intentionally left blank

Tables

1.1	Data for the Anscombe Quartet	9
2.1	Bobbleheads and Dodger Dogs	16
2.2	Regression of Attendance on Month, Day of Week, and Bobblehead Promotion	22
3.1	Preference Data for Mobile Communication Services	30
4.1	Market Basket for One Shopping Trip	38
4.2	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	112
8.2	Accuracy of Text Classification for Movie Reviews (Thumbs-Up or Thumbs-Down)	116
8.3	Random Forest Text Measurement Model Applied to Tom's Movie Reviews	117
9.1	New York Mets' Early Season Games in 2007	147
9.2	New York Yankees' Early Season Games in 2007	148
10.1	California Housing Data: Original and Computed Variables	169
10.2	Linear Regression Fit to Selected California Block Groups	173
10.3	Comparison of Regressions on Spatially Referenced Data	176
11.1	Contingency Table of Top-ranked Brands and Most Valued Attributes	191
11.2	Market Simulation: Choice Set Input	198
11.3	Market Simulation: Preference Shares in a Hypothetical Four-brand Market	200
C.1	Hypothetical profits from model-guided vehicle selection	266
C.2	DriveTime Data for Sedans	267
C.3	DriveTime Sedan Color Map with Frequency Counts	268
C.4	Diamonds Data: Variable Names and Coding Rules	272

C.5	Dells Survey Data: Visitor Characteristics	276
C.6	Dells Survey Data: Visitor Activities	277
C.7	Computer Choice Study: Product Attributes	279
C.8	Computer Choice Study: Data for One Individual	281

Exhibits

1.1	Programming the Anscombe Quartet	13
2.1	Shaking Our Bobbleheads Yes and No	25
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	77
7.1	Text Analysis of Movie Taglines	100
8.1	Sentiment Analysis and Classification of Movie Ratings	123
9.1	Team Winning Probabilities by Simulation	165
10.1	Regression Models for Spatial Data	178
11.1	Training and Testing a Hierarchical Bayes Model	203
11.2	Preference, Choice, and Market Simulation	208
D.1	Conjoint Analysis Spine Chart	286
D.2	Market Simulation Utilities	294
D.3	Split-plotting Utilities	295
D.4	Wait-time Ribbon Plot	298
D.5	Movie Tagline Data Preparation Script for Text Analysis	310
D.6	Word Scoring Code for Sentiment Analysis	315
D.7	Utilities for Spatial Data Analysis	319
D.8	Making a Word Cloud	320

This page intentionally left blank

1

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.

A *model* is a representation of things, a rendering or description of reality. A typical model in data science is an attempt to relate one set of variables to another. Limited, imprecise, but useful, a model helps us to make sense of the world. A model is more than just talk because it is based on data.

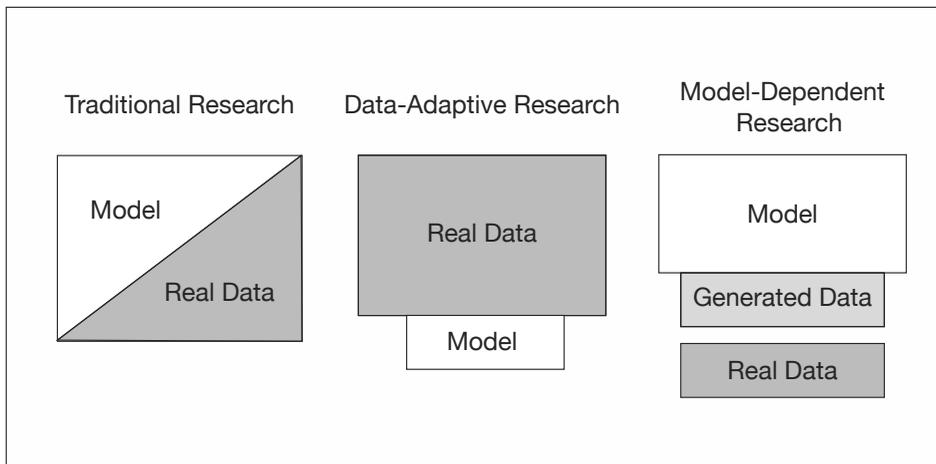
Predictive analytics brings together management, information technology, and modeling. It is designed for today's data-intensive world. Predictive analytics is data science, a multidisciplinary skill set essential for success in business, nonprofit organizations, and government. Whether forecasting sales 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, as with 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 that we observe, manipulate, or control and might relate to the response.

Regression methods help us to predict a response with meaningful magnitude, such as quantity sold, stock price, or return on investment. Classification methods help us to predict 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?

Prediction problems are defined by their width or number of potential predictors and by their depth or number of observations in the data set. It is the number of potential predictors in business, marketing, and investment analysis that causes the most difficulty. There can be thousands of potential

Figure 1.1. Data and models for research

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. Models that fit the data well are better than models that fit the data poorly. Simple models are better than complex models.

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, statistical inference, 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 and checking them with diagnostics. We validate traditional models before using them to make predictions.

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

Data-adaptive methods are data-driven. As with traditional models, we validate data-adaptive models before using them to make predictions.

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. The comparison with real data serves as a form of validation.

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 indicators 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 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. And model validation is essential to the process.

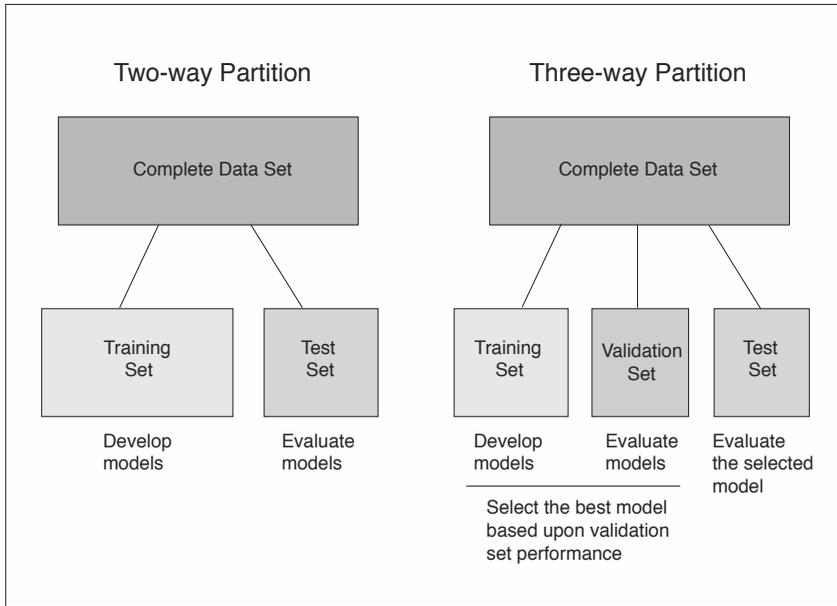
To make predictions, we may employ classical or Bayesian methods. Or we may dispense with traditional statistics entirely and rely upon machine learning algorithms. We do what works.¹ Our approach to predictive analytics is based upon a simple premise:

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, significance tests, and p -values—that is the classical way. On the other hand, we have posterior probability distributions, probability intervals, prediction intervals, Bayes factors, and subjective (perhaps diffuse) priors—the path of Bayesian statistics. Indices such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC) help us 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

¹ 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-validation, the logical extreme of multi-fold cross-validation, there are as many test sets as there are observations in the sample.

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

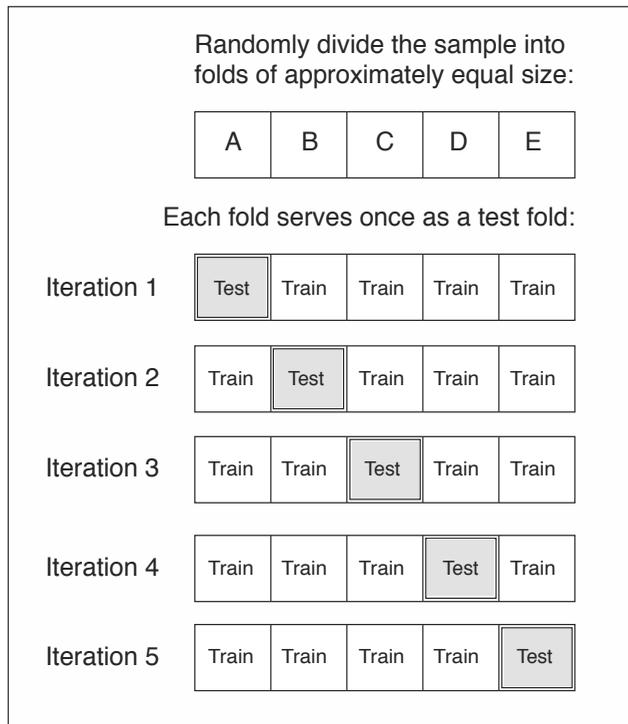
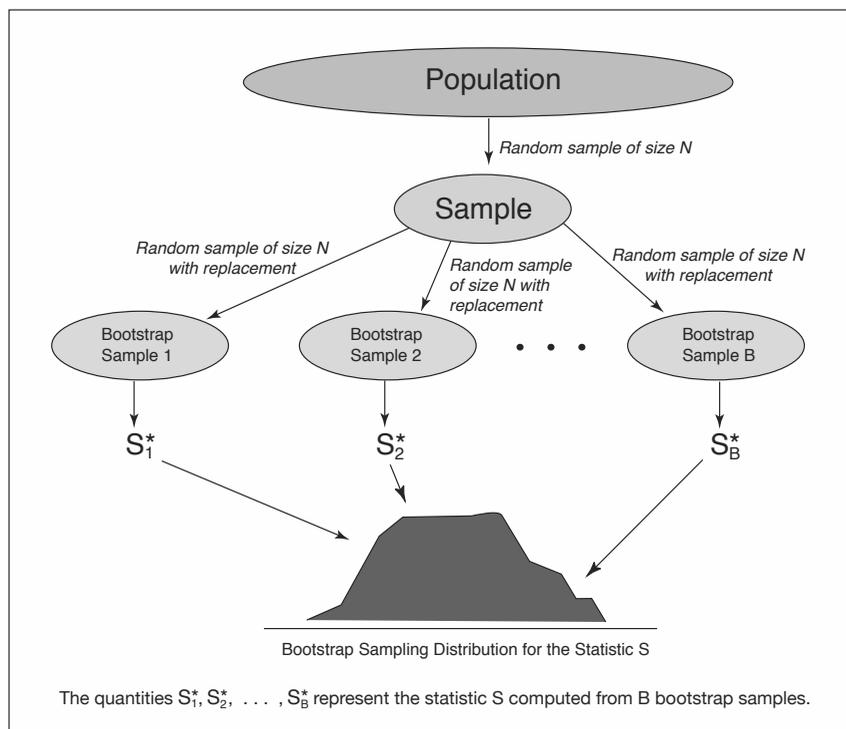


Figure 1.4. Training-and-Test with Bootstrap Resampling



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).

Table 1.1. Data for the Anscombe Quartet

Set I		Set II		Set III		Set IV	
x1	y1	x2	y2	x3	y3	x4	y4
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

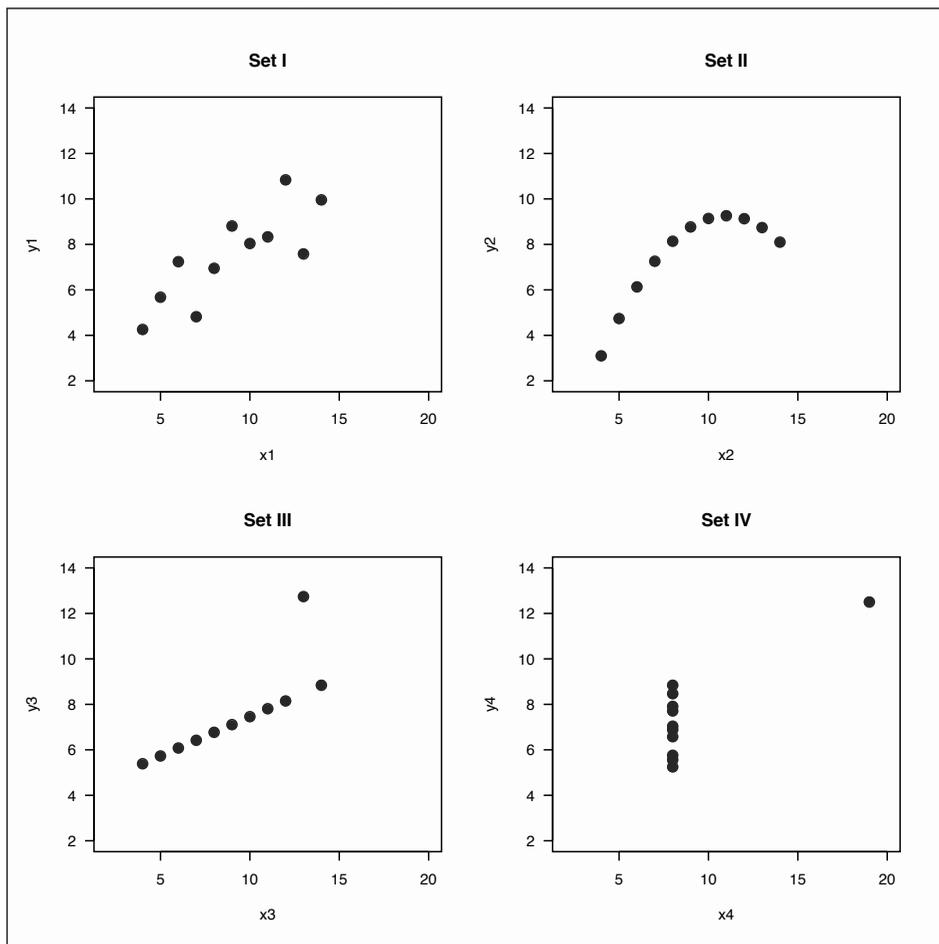
Data visualization is critical to the work of data science. Examples in this book demonstrate the importance of data visualization in discovery, diagnostics, and design. We employ tools of exploratory data analysis (discovery) and statistical modeling (diagnostics). In communicating results to management,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

Following Anscombe (1973), we would argue that statistical summaries fail to tell the story of data. We must look beyond data tables, regression coefficients, and the results of statistical tests. It is the plots in figure 1.5 that tell the story. 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 them. An R program for the Anscombe Quartet is provided at the end of this chapter in exhibit 1.1.

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. Cairo (2013) provides a general review of information graphics. Heer, Bostock, and Ogievetsky (2010) demonstrate contemporary visualization techniques for web distribution. When working with very large data sets, special methods may be needed, such as partial transparency and hexbin plots (Unwin, Theus, and Hofmann 2006; Carr, Lewin-Koh, and Maechler 2014; Lewin-Koh 2014).

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. Matloff (2011) and Lander (2014) provide useful introductions to R. A graphics overview is provided by Murrell (2011).

R lattice graphics, discussed by Sarkar (2008, 2014), build upon the conceptual structure of an earlier system called S-Plus TrellisTM (Cleveland 1993; Becker and Cleveland 1996). Wilkinson's (2005) "grammar of graphics" approach has been implemented in the R `ggplot2` and `ggvis` packages (Wickham and Chang 2014; Chang 2014), with programming examples provided by Chang (2013). Cairo (2013) and Zeileis, Hornik, and Murrell (2009, 2014) provide advice about colors for statistical graphics. Ihaka et al. (2014) show how to specify colors in R by hue, chroma, and luminance.

These are the things that data scientists do:

- **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.
- **Preparing text and data.** Text is unstructured or partially structured. Data are often messy or missing. We extract features from text. We define measures. We prepare text and data for analysis and modeling.
- **Looking at data.** We do exploratory data analysis, data visualization for the purpose of discovery. We look for groups in data. We find outliers. We identify common dimensions, patterns, and trends.
- **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.
- **Predicting yes or no.** Many business problems are classification problems. We use classification methods to predict whether or not a person will buy a product, default on a loan, or access a web page.
- **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 manipulate key variables to see what happens to our predictions. We play what-if games in simulated marketplaces. We employ sensitivity or stress testing of mathematical programming models. We see how values of input variables affect outcomes, pay-offs, and predictions. We assess uncertainty about forecasts.
- **Explaining it all.** Data and models help us 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.

Let us begin.

Exhibit 1.1. Programming the Anscombe Quartet

```
# The Anscombe Quartet (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(
  x1 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
  x2 = 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),
  y1 = c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68),
  y2 = c(9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74),
  y3 = 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, data = anscombe))))
with(anscombe, print(summary(lm(y2 ~ x2, data = anscombe))))
with(anscombe, print(summary(lm(y3 ~ x3, data = anscombe))))
with(anscombe, print(summary(lm(y4 ~ x4, data = anscombe))))

# 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_anscombe_R.pdf", width = 8.5, height = 8.5)
par(mfrow=c(2,2), mar=c(5.1, 4.1, 4.1, 2.1))
with(anscombe, plot(x1, y1, xlim=c(2,20), ylim=c(2,14), pch = 19,
  col = "darkblue", cex = 1.5, las = 1, xlab = "x1", ylab = "y1"))
title("Set I")
with(anscombe, plot(x2, y2, xlim=c(2,20), ylim=c(2,14), pch = 19,
  col = "darkblue", cex = 1.5, las = 1, xlab = "x2", ylab = "y2"))
title("Set II")
with(anscombe, plot(x3, y3, xlim=c(2,20), ylim=c(2,14), pch = 19,
  col = "darkblue", cex = 1.5, las = 1, xlab = "x3", ylab = "y3"))
title("Set III")
with(anscombe, plot(x4, y4, xlim=c(2,20), ylim=c(2,14), pch = 19,
  col = "darkblue", cex = 1.5, las = 1, xlab = "x4", ylab = "y4"))
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

Index

A

accuracy, *see* classification, predictive accuracy
advertising, 14–28
Akaike information criterion (AIC), 5
Alteryx, 237, 284
ARIMA model, *see* time series analysis
arules, *see* R package, arules
arulesViz, *see* R package, arulesViz
association rule, 40–42, 242

B

bag-of-words approach, *see* text analytics
bar chart, *see* data visualization
base rate, *see* classification, predictive accuracy
Bayes information criterion (BIC), 5
Bayes' theorem, *see* Bayesian statistics, Bayes' theorem
Bayesian statistics, 5, 177, 189, 202, 223, 230, 231, 246
 Bayes' theorem, 231
benchmark study, *see* simulation
best-worst scaling, 256
biclustering, 242
big data, 221, 227
biologically-inspired methods, 238
biplot, *see* data visualization
black box model, 237
block clustering, *see* biclustering
bootstrap method, 8
box plot, *see* data visualization
brand equity research, 187–220
bubble chart, *see* data visualization

C

call center scheduling, *see*
 scheduling, workforce scheduling
car, *see* R package, car
caret, *see* R package, caret

censoring, 170, 263
choice study, 29
 menu-based, 260
ChoiceModelR, *see* R package, ChoiceModelR
classical statistics, 5, 229, 231
 null hypothesis, 229
 power, 230
 statistical significance, 229, 230
classification, 2, 12, 107, 116, 233, 235, 237
 predictive accuracy, 234, 235
classification tree, *see* tree-structured model
cluster, *see* R package, cluster
cluster analysis, 99, 237, 238, 240
coefficient of determination, 233
collaborative filtering, 242
column-oriented database, *see* database
 system, non-relational
complexity, of model, 236
computational linguistics, *see* text analytics,
 natural language processing
confidence interval, *see* classical statistics,
 confidence interval
confusion matrix, *see* classification, predictive
 accuracy
conjoint analysis, 33, 254
content analysis, *see* text analytics, content
 analysis
corpus, *see* text analytics
correlation heat map, *see* data visualization,
 heat map
credit scoring, 248
cross-sectional study, *see* data organization
cross-validation, 6, 236
cutoff rule, *see* classification, predictive
 accuracy
cvTools, *see* R package, cvTools

D

data mining, *see* data-adaptive research
 data munging, *see* data preparation
 data organization, 5, 58
 data partitioning, 6
 data preparation, 228
 missing data, 228
 data science, 1–12, 225, 226
 data visualization, 8
 bar chart, 41, 43
 biplot, 90
 box plot, 15, 17
 bubble chart, 45
 density plot, 191, 193
 diagnostics, 21, 235
 dot chart, 118, 175
 heat map, 158, 159, 170, 171
 histogram, 86, 151, 154, 156, 157
 horizon plot, 54, 56, 57, 91, 92
 lattice plot, 11, 15, 18, 19, 21
 line graph, 72, 75
 mosaic plot, 189, 190
 multiple time series plot, 55
 network diagram, 47
 parallel coordinates, 194, 196
 ribbon plot, 68–71, 298
 scatter plot, 44
 scatter plot matrix, 170, 172
 spine chart, 31–34, 286
 strip plot, 15, 19
 ternary plot, 189, 191, 192
 time series plot, 59, 60
 tree diagram, 119, 174
 word cloud, 99, 284, 285
 data-adaptive research, 3, 4
 database system, 227, 228
 non-relational, 227, 228
 relational, 227, 228
 density plot, *see* data visualization
 dependent variable, *see* response
 discrete event simulation, *see* simulation,
 discrete event simulation
 document annotation, *see* text analytics,
 document annotation
 document database, *see* database system,
 non-relational
 dot chart, *see* data visualization
 duration analysis, *see* survival analysis

E

e1071, *see* R package, e1071
 economic analysis, 53–66
 indexing, 54

elimination pick list, 261
 empirical Bayes, 223, *see* Bayesian statistics
 Erlang C, *see* queueing model
 explanatory model, 226
 explanatory variable, 2, 3, 233
 exploratory data analysis, 15

F

false negative, *see* classification, predictive
 accuracy
 false positive, *see* classification, predictive
 accuracy
 financial data analysis, 4, 248
 forecast, *see* R package, forecast
 forecasting, 58–61, 174
 four Ps, *see* marketing mix model
 four-fold table, *see* classification, predictive
 accuracy

G

game-day simulation, *see* simulation,
 game-day
 General Inquirer, 120
 generalized linear model, 233, 236
 generative grammar, *see* text analytics
 genetic algorithms, 238
 geographically weighted regression, 174
 ggplot2, *see* R package, ggplot2
 graph database, *see* database system,
 non-relational
 graphics, *see* data visualization
 grid, *see* R package, grid
 group filtering, *see* collaborative filtering

H

heuristics, 238
 hierarchical Bayes, *see* Bayesian statistics
 hierarchical model, 177, 223
 histogram, *see* data visualization
 horizon plot, *see* data visualization

I

IBM, 237, 284
 independent variable, *see* explanatory variable
 integer programming, *see* mathematical
 programming
 interaction effect, 235
 interval estimate, *see* statistic, interval estimate
 item analysis, psychometrics, 115

K

Kappa, *see* classification, predictive accuracy
 key-value store, *see* database system,
 non-relational
 KNIME, 237

L

latent Dirichlet allocation, *see* text analytics,
 latent Dirichlet allocation
 latent semantic analysis, *see* text analytics,
 latent semantic analysis
 lattice, *see* R package, lattice
 lattice plot, *see* data visualization
 latticeExtra, *see* R package, latticeExtra
 leading indicator, 54, 61
 least-squares regression, *see* regression
 lexical table, *see* text analytics,
 terms-by-documents matrix
 line graph, *see* data visualization
 linear least-squares regression, *see* regression
 linear model, 233, 236
 linear predictor, 233
 linguistics, *see* text analytics, natural language
 processing
 lmtest, *see* R package, lmtest
 log-linear models, 240
 logical empiricism, 1
 logistic regression, 3, 115, 233
 longitudinal study, *see* data organization
 lpSolve, *see* R package, lpSolve
 lubridate, *see* R package, lubridate

M

machine learning, 237, 238, *see* data-adaptive
 research
 map-reduce, *see* database system, non-relational
 mapproj, *see* R package, mapproj
 maps, *see* R package, maps
 market basket analysis, 37–52, 242
 market response model, 24
 market segmentation, *see* segmentation
 market simulation, *see* simulation
 marketing mix model, 23
 Markov chain Monte Carlo, *see* Bayesian
 statistics, Markov chain Monte Carlo
 mathematical programming, 4, 67, 75, 248
 integer programming, 74
 sensitivity testing, 75
 matrix bubble chart, *see* data visualization,
 bubble chart
 mean-squared error (MSE), *see*
 root mean-squared error (RMSE)
 measurement, 249–262

construct validity, 249
 content validity, 121
 convergent validity, 250
 discriminant validity, 250
 face validity, 121
 multitrait-multimethod matrix, 249, 251
 reliability, 249
 meta-analysis, 223
 metadata, *see* text analytics
 Microsoft, 284
 missing data, *see* data preparation, missing
 data
 model validation, *see* training-and-test
 regimen
 model-dependent research, 3, 4
 morphology, *see* text analytics
 mosaic plot, *see* data visualization
 multicollinearity, 168, 170
 multidimensional scaling, 87, 89, 99, 240, 243,
 244
 multilevel models, *see* hierarchical models
 multiple imputation, *see* data preparation,
 missing data
 multiple time series plot, *see* data
 visualization, time series plot
 multivariate methods, 99, 243

N

natural language processing, *see* text analytics
 nearest-neighbor model, 176, 177, 242
 network diagram, *see* data visualization
 neural network, 4
 non-relational database, *see* database system,
 non-relational
 NoSQL, *see* database system, non-relational

O

operations management, 67–83
 optimization, 238
 constrained, 74
 organization of data, *see* data, organization
 over-fitting, 170, 176, 235

P

p-value, *see* statistic, p-value
 paired comparisons, 255, 258
 parallel coordinates plot, *see* data visualization
 parametric models, 235
 parsing, *see* text analytics, text parsing
 perceptual map, *see* data visualization
 philosophy, 1
 point estimate, *see* statistic, point estimate
 Poisson regression, 232
 power, *see* classical statistics, power

predictive analytics, 1–12
 definition, 2
 predictive model, 226
 predictor, *see* explanatory variable
 preference scaling, 244
 preference study, 29
 pricing research, 187–220
 principal component analysis, 238, 243
 privacy, 240
 probability
 binomial distribution, 153
 negative binomial distribution, 153, 155, 158
 Poisson distribution, 153, 155, 158
 probability cutoff, *see* classification, predictive accuracy
 probability heat map, *see* data visualization, heat map
 probability interval, *see* Bayesian statistics, probability interval
 process simulation, *see* simulation, process simulation
 product positioning, 243, 244
 promotion, 14–28
 proxy, *see* R package, proxy

Q

quantmod, *see* R package, quantmod
 queueing, *see* R package, queueing
 queueing model, 67, 68, 73

R

R package
 arules, 50
 arulesViz, 50
 car, 25
 caret, 123, 203, 208
 ChoiceModelR, 203, 208
 cluster, 100
 cvTools, 178
 e1071, 123
 forecast, 62
 ggplot2, 77, 100, 123, 208
 grid, 77, 100, 123
 lattice, 25, 165, 178, 208
 latticeExtra, 62, 100, 123
 lmtest, 62
 lpSolve, 77
 lubridate, 62, 77
 mapproj, 178
 maps, 178
 proxy, 100
 quantmod, 62

 queueing, 77
 randomForest, 123, 178
 RColorBrewer, 50
 rpart, 123, 178
 rpart.plot, 123, 178
 spgwr, 178
 stringr, 100, 123
 support.CEs, 34
 tm, 100, 123
 vcd, 208
 wordcloud, 100, 320
 R-squared, 233
 random forest, 116–118, 170, 175
 randomForest, *see* R package, randomForest
 RColorBrewer, *see* R package, RColorBrewer
 recommender systems, 241, 242
 regression, 2, 3, 12, 20, 22, 23, 115, 170, 173, 232, 236
 nonlinear regression, 236
 robust methods, 236
 time series regression, 58
 regression tree, *see* tree-structured model
 regularized regression, 236
 relational database, *see* database system, relational
 reliability, *see* measurement
 response, 2, 232
 ribbon plot, *see* data visualization
 risk analytics, 248
 robust methods, *see* regression
 ROC curve, *see* classification, predictive accuracy
 root mean-squared error (RMSE), 233
 rpart, *see* R package, rpart
 rpart.plot, *see* R package, rpart.plot
 RStudio, 284

S

sales forecasting, *see* forecasting
 sampling
 sampling variability, 230
 SAS, 237, 284
 scatter plot, *see* data visualization
 scatter plot matrix, *see* data visualization
 scheduling, 238
 workforce scheduling, 67–83
 segmentation, 245, 246
 semantics, *see* text analytics
 semi-supervised learning, 238
 sentiment analysis, 107–143
 shrinkage estimators, 236
 significance, *see* classical statistics, statistical significance
 simulation, 145, 146, 149, 236, 248

- benchmark study, 116, 174, 236, 237
 - discrete event simulation, 67, 75
 - game-day, 144, 146, 149, 150
 - market simulation, 194, 198, 200
 - process simulation, 67, 68, 76
 - what-if analysis, 12
 - site selection, 174, *see* spatial data analysis
 - smoothing methods, 236
 - splines, 236
 - social filtering, *see* collaborative filtering
 - social network analysis, 239, 240
 - spatial data analysis, 167–187
 - site selection, 247
 - spatio-temporal model, 168, 177
 - spatio-temporal model, *see* spatial data analysis, spatio-temporal model
 - spgwr, *see* R package, spgwr
 - spine chart, *see* data visualization
 - sports analytics, 143–166
 - SQL, *see* database system, relational
 - state space model, *see* time series analysis
 - statistic
 - interval estimate, 229
 - p-value, 229
 - point estimate, 229
 - test statistic, 229
 - statistical experiment, *see* simulation
 - statistical graphics, *see* data visualization
 - statistical learning, *see* data-adaptive research
 - statistical significance, *see* classical statistics, statistical significance
 - statistical simulation, *see* simulation
 - stringr, *see* R package, stringr
 - strip plot, *see* data visualization
 - supervised learning, 97, 232, 238
 - support vector machines, 116
 - support.CEs, *see* R package, support.CEs
 - survey research, 262
 - survival analysis, 248
 - syntax, *see* text analytics
- T**
- tag, *see* text analytics, metadata
 - target marketing, 245, 246
 - terms-by-documents matrix, *see* text analytics
 - ternary plot, *see* data visualization
 - test statistic, *see* statistic, test statistic
 - text analytics, 83–107
 - bag-of-words approach, 86, 91
 - content analysis, 120
 - corpus, 87
 - document annotation, 262
 - generative grammar, 93, 94
 - latent Dirichlet allocation, 238
 - latent semantic analysis, 238
 - metadata, 85
 - morphology, 94
 - natural language processing, 86, 91, 93, 122
 - semantics, 94
 - stemming, 95
 - syntax, 94
 - terms-by-documents matrix, 87, 95, 96
 - text feature, 262
 - text parsing, 85, 93
 - text summarization, 97
 - thematic analysis, 120, 238
 - text feature, *see* text analytics, text feature
 - text measure, 85, 86, 91, 120, 121, 262
 - text mining, *see* text analytics
 - thematic analysis, *see* text analytics, thematic analysis
 - time series analysis, 53
 - ARIMA model, 58
 - multiple time series, 55
 - state space model, 58
 - time series plot, *see* data visualization
 - tm, *see* R package, tm
 - traditional research, 3
 - training-and-test regimen, 5, 6, 8, 12, 20, 21, 116, 170, 174, 176, 188
 - transformation, *see* variable transformation
 - tree diagram, *see* data visualization
 - tree-structured model
 - classification, 117, 119
 - regression, 170, 174
 - trellis plot, *see* data visualization, lattice plot
 - triplot, *see* data visualization, ternary plot
- U**
- unit of analysis, 5
 - unsupervised learning, 97, 238
- V**
- validation, *see* training-and-test regimen
 - validity, *see* measurement
 - variable transformation, 168, 235
 - vcd, *see* R package, vcd
- W**
- wait-time ribbon, *see* data visualization, ribbon plot
 - web analytics, 239
 - Weka, 49
 - what-if analysis, *see* simulation
 - wordcloud, *see* R package, wordcloud and data visualization, word cloud