

# **An R Companion to Applied Regression**

**Third Edition**

**John Fox**

McMaster University

**Sanford Weisberg**

University of Minnesota



Los Angeles | London | New Delhi  
Singapore | Washington DC | Melbourne

# Contents

Preface	xiii
What Is R?	xiv
Obtaining and Installing R and RStudio	xvi
Installing R on a Windows System	xvii
Installing R on a macOS System	xvii
Installing RStudio	xviii
Installing and Using R Packages	xx
Optional: Customizing R	xxii
Optional: Installing L <sup>A</sup> T <sub>E</sub> X	xxiii
Using This Book	xxiii
Chapter Synopses	xxiv
Typographical Conventions	xxv
New in the Third Edition	xxvi
The Website for the <i>R Companion</i>	xxvii
Beyond the <i>R Companion</i>	xxviii
Acknowledgments	xxviii
About the Authors	xxx
1 Getting Started With R and RStudio	1
1.1 Projects in RStudio	2
1.2 R Basics	5
1.2.1 Interacting With R Through the Console	5
1.2.2 Editing R Commands in the Console	7
1.2.3 R Functions	7
1.2.4 Vectors and Variables	11
1.2.5 Nonnumeric Vectors	14
1.2.6 Indexing Vectors	16
1.2.7 User-Defined Functions	18
1.3 Fixing Errors and Getting Help	21
1.3.1 When Things Go Wrong	21
1.3.2 Getting Help and Information	23
1.4 Organizing Your Work in R and RStudio and Making It Reproducible	25
1.4.1 Using the RStudio Editor With R Script Files	25
1.4.2 Writing R Markdown Documents	28
1.5 An Extended Illustration: Duncan's Occupational-Prestige Regression	33
1.5.1 Examining the Data	36
1.5.2 Regression Analysis	39
1.5.3 Regression Diagnostics	40

1.6	R Functions for Basic Statistics	47
1.7	Generic Functions and Their Methods*	47
2	Reading and Manipulating Data	53
2.1	Data Input	54
2.1.1	Accessing Data From a Package	54
2.1.2	Entering a Data Frame Directly	56
2.1.3	Reading Data From Plain-Text Files	59
2.1.4	Files and Paths	63
2.1.5	Exporting or Saving a Data Frame to a File	65
2.1.6	Reading and Writing Other File Formats	66
2.2	Other Approaches to Reading and Managing Data Sets in R	67
2.3	Working With Data Frames	69
2.3.1	How the R Interpreter Finds Objects	69
2.3.2	Missing Data	72
2.3.3	Modifying and Transforming Data	79
2.3.4	Binding Rows and Columns	86
2.3.5	Aggregating Data Frames	87
2.3.6	Merging Data Frames	89
2.3.7	Reshaping Data	91
2.4	Working With Matrices, Arrays, and Lists	95
2.4.1	Matrices	96
2.4.2	Arrays	97
2.4.3	Lists	98
2.4.4	Indexing	99
2.5	Dates and Times	107
2.6	Character Data	110
2.7	Large Data Sets in R*	117
2.7.1	How Large Is “Large”?	118
2.7.2	Reading and Saving Large Data Sets	120
2.8	Complementary Reading and References	122
3	Exploring and Transforming Data	123
3.1	Examining Distributions	124
3.1.1	Histograms	124
3.1.2	Density Estimation	128
3.1.3	Quantile-Comparison Plots	130
3.1.4	Boxplots	133
3.2	Examining Relationships	134
3.2.1	Scatterplots	134
3.2.2	Parallel Boxplots	141
3.2.3	More on the <code>plot()</code> Function	144
3.3	Examining Multivariate Data	145
3.3.1	Three-Dimensional Plots	145
3.3.2	Scatterplot Matrices	146
3.4	Transforming Data	148

3.4.1	Logarithms: The Champion of Transformations	148
3.4.2	Power Transformations	154
3.4.3	Transformations and Exploratory Data Analysis	162
3.4.4	Transforming Restricted-Range Variables	167
3.4.5	Other Transformations	168
3.5	Point Labeling and Identification	169
3.5.1	The <code>identify()</code> Function	169
3.5.2	Automatic Point Labeling	170
3.6	Scatterplot Smoothing	171
3.7	Complementary Reading and References	172
4	Fitting Linear Models	173
4.1	The Linear Model	174
4.2	Linear Least-Squares Regression	176
4.2.1	Simple Linear Regression	176
4.2.2	Multiple Linear Regression	183
4.2.3	Standardized Regression Coefficients	185
4.3	Predictor Effect Plots	187
4.4	Polynomial Regression and Regression Splines	190
4.4.1	Polynomial Regression	190
4.4.2	Regression Splines*	194
4.5	Factors in Linear Models	197
4.5.1	A Linear Model With One Factor: One-Way Analysis of Variance	201
4.5.2	Additive Models With Numeric Predictors and Factors	205
4.6	Linear Models With Interactions	207
4.6.1	Interactions Between Numeric Predictors and Factors	207
4.6.2	Shortcuts for Writing Linear-Model Formulas	213
4.6.3	Multiple Factors	214
4.6.4	Interactions Between Numeric Predictors*	222
4.7	More on Factors	224
4.7.1	Dummy Coding	224
4.7.2	Other Factor Codings	224
4.7.3	Ordered Factors and Orthogonal-Polynomial Contrasts	227
4.7.4	User-Specified Contrasts*	230
4.7.5	Suppressing the Intercept in a Model With Factors*	231
4.8	Too Many Regressors*	232
4.9	The Arguments of the <code>lm()</code> Function	235
4.9.1	<code>formula</code>	235
4.9.2	<code>data</code>	238
4.9.3	<code>subset</code>	238
4.9.4	<code>weights</code>	239
4.9.5	<code>na.action</code>	239
4.9.6	<code>method, model, x, y, qr*</code>	240
4.9.7	<code>singular.ok*</code>	240
4.9.8	<code>contrasts</code>	240

4.9.9	offset	240
4.10	Complementary Reading and References	241
5	Coefficient Standard Errors, Confidence Intervals, and Hypothesis Tests	243
5.1	Coefficient Standard Errors	244
5.1.1	Conventional Standard Errors of Least-Squares Regression Coefficients	244
5.1.2	Robust Regression Coefficient Standard Errors	246
5.1.3	Using the Bootstrap to Compute Standard Errors	248
5.1.4	The Delta Method for Standard Errors of Nonlinear Functions*	252
5.2	Confidence Intervals	254
5.2.1	Wald Confidence Intervals	254
5.2.2	Bootstrap Confidence Intervals	255
5.2.3	Confidence Regions and Data Ellipses*	256
5.3	Testing Hypotheses About Regression Coefficients	258
5.3.1	Wald Tests	258
5.3.2	Likelihood-Ratio Tests and the Analysis of Variance	259
5.3.3	Sequential Analysis of Variance	260
5.3.4	The Anova () Function	262
5.3.5	Testing General Linear Hypotheses*	267
5.4	Complementary Reading and References	270
6	Fitting Generalized Linear Models	271
6.1	Review of the Structure of GLMs	272
6.2	The glm() Function in R	276
6.3	GLMs for Binary Response Data	276
6.3.1	Example: Women's Labor Force Participation	278
6.3.2	Example: Volunteering for a Psychological Experiment	282
6.3.3	Predictor Effect Plots for Logistic Regression	283
6.3.4	Analysis of Deviance and Hypothesis Tests for Logistic Regression	285
6.3.5	Fitted and Predicted Values	289
6.4	Binomial Data	289
6.5	Poisson GLMs for Count Data	296
6.6	Loglinear Models for Contingency Tables	301
6.6.1	Two-Dimensional Tables	301
6.6.2	Three-Dimensional Tables	304
6.6.3	Sampling Plans for Loglinear Models	306
6.6.4	Response Variables	307
6.7	Multinomial Response Data	309
6.8	Nested Dichotomies	314
6.9	The Proportional-Odds Model	317
6.9.1	Testing for Proportional Odds	319
6.10	Extensions	322
6.10.1	More on the Anova () Function	322

6.10.2	Gamma Models	323
6.10.3	Quasi-Likelihood Estimation	325
6.10.4	Overdispersed Binomial and Poisson Models	326
6.11	Arguments to <code>glm()</code>	330
6.11.1	<code>weights</code>	331
6.11.2	<code>start, etastart, mustart</code>	331
6.11.3	<code>offset</code>	331
6.11.4	<code>control</code>	332
6.11.5	<code>model, method, x, y</code>	332
6.12	Fitting GLMs by Iterated Weighted Least Squares*	332
6.13	Complementary Reading and References	333
7	Fitting Mixed-Effects Models	335
7.1	Background: The Linear Model Revisited	336
7.1.1	The Linear Model in Matrix Form*	336
7.2	Linear Mixed-Effects Models	336
7.2.1	Matrix Form of the Linear Mixed-Effects Model*	338
7.2.2	An Application to Hierarchical Data	339
7.2.3	Wald Tests for Linear Mixed-Effects Models	357
7.2.4	Examining the Random Effects: Computing BLUPs	358
7.2.5	An Application to Longitudinal Data	360
7.2.6	Modeling the Errors	371
7.2.7	Sandwich Standard Errors for Least-Squares Estimates	373
7.3	Generalized Linear Mixed Models	375
7.3.1	Matrix Form of the GLMM*	376
7.3.2	Example: Minneapolis Police Stops	377
7.4	Complementary Reading	382
8	Regression Diagnostics for Linear, Generalized Linear, and Mixed-Effects Models	385
8.1	Residuals	386
8.2	Basic Diagnostic Plots	388
8.2.1	Plotting Residuals	388
8.2.2	Marginal-Model Plots	391
8.2.3	Added-Variable Plots	392
8.2.4	Marginal-Conditional Plots	395
8.3	Unusual Data	396
8.3.1	Outliers and Studentized Residuals	397
8.3.2	Leverage: Hat-Values	398
8.3.3	Influence Measures	399
8.4	Transformations After Fitting a Regression Model	405
8.4.1	Transforming the Response	406
8.4.2	Predictor Transformations	410
8.5	Nonconstant Error Variance	414
8.5.1	Testing for Nonconstant Error Variance	416
8.6	Diagnostics for Generalized Linear Models	417

8.6.1	Residuals and Residual Plots	418
8.6.2	Influence Measures	421
8.6.3	Graphical Methods: Added-Variable Plots, Component-Plus-Residual Plots, and Effect Plots With Partial Residuals	422
8.7	Diagnostics for Mixed-Effects Models	425
8.7.1	Mixed-Model Component-Plus-Residual Plots	425
8.7.2	Influence Diagnostics for Mixed Models	428
8.8	Collinearity and Variance Inflation Factors	429
8.9	Additional Regression Diagnostics	434
8.10	Complementary Reading and References	435
9	Drawing Graphs	437
9.1	A General Approach to R Graphics	438
9.1.1	Defining a Coordinate System: <code>plot()</code>	439
9.1.2	Graphics Parameters: <code>par()</code>	441
9.1.3	Adding Graphical Elements: <code>axis()</code> , <code>points()</code> , <code>lines()</code> , <code>text()</code> , et al.	442
9.1.4	Specifying Colors	452
9.2	Putting It Together: Explaining Local Linear Regression	454
9.2.1	Finer Control Over Plot Layout	461
9.3	Other R Graphics Packages	467
9.3.1	The <b>lattice</b> Package	467
9.3.2	The <b>ggplot2</b> Package	469
9.3.3	Maps	472
9.3.4	Other Notable Graphics Packages	475
9.4	Complementary Reading and References	476
10	An Introduction to R Programming	477
10.1	Why Learn to Program in R?	478
10.2	Defining Functions: Preliminary Examples	479
10.2.1	Lagging a Variable	479
10.2.2	Creating an Influence Plot	482
10.3	Working With Matrices*	486
10.3.1	Basic Matrix Arithmetic	486
10.3.2	Matrix Inversion and the Solution of Linear Simultaneous Equations	488
10.3.3	Example: Linear Least-Squares Regression	489
10.3.4	Eigenvalues and Eigenvectors	491
10.3.5	Miscellaneous Matrix Computations	491
10.4	Program Control With Conditionals, Loops, and Recursion	492
10.4.1	Conditionals	492
10.4.2	Iteration (Looping)	495
10.4.3	Recursion	498
10.5	Avoiding Loops: <code>apply()</code> and Its Relatives	499
10.5.1	To Loop or Not to Loop?	505

10.6	Optimization Problems*	509
10.6.1	Zero-Inflated Poisson Regression	509
10.7	Monte-Carlo Simulations*	515
10.7.1	Testing Regression Models Using Simulation	516
10.8	Debugging R Code*	522
10.9	Object-Oriented Programming in R*	527
10.10	Writing Statistical-Modeling Functions in R*	533
10.11	Organizing Code for R Functions	536
10.12	Complementary Reading and References	537
	References	539
	Subject Index	551
	Data Set Index	567
	Package Index	569
	Index of Functions and Operators	571