APPLIED REGRESSION ANALYSIS & GENERALIZED LINEAR MODELS

3 EDITION

"The strength of this text is the **unified presentation** of several regression topics that provides the student with a global perspective on regression analysis. The student is well served with this unified approach, as it facilitates deeper research on any one topic with more advanced texts."

—E. C. Hedberg, *Arizona State University*

"This text is a **one-stop shop** for me for my first year stats sequence for students in our program. Those wanting the technical detail will be satisfied; those wanting an **excellent** explanation of these methods using **real-world examples and approachable language** will also be satisfied."

—Corey S. Sparks, *The University of Texas at San Antonio*

"I have **enjoyed** using previous editions of this text and look forward to using this edition. It covers all key topics, and quite a few advanced ones, in one **well-written text**."

—Michael S. Lynch, *University of Georgia*

PRAISE FOR THE PREVIOUS EDITIONS

"In summary, this is an **excellent** text on regression applications and methods, written with authority, lucidity, and eloquence. The second edition provides substantive and topical updates, and makes the book suitable for courses designed to emphasize both the classical and the modern aspects of regression."

—Journal of the American Statistical Association
(review of the second edition)

"Even though the book is written with social scientists as the target audience, the depth of material and how it is conveyed give it far broader appeal. Indeed, I recommend it as a **useful** learning text and resource for researchers and students in any field that applies regression or linear models (that is, most everyone), including courses for undergraduate statistics majors. The author is to be commended for giving us this book, which I trust will find a wide and enduring readership."

—Journal of the American Statistical Association
(review of the first edition)

"[T]his **wonderfully comprehensive** book focuses on regression analysis and linear models.... We enthusiastically recommend this book—having used it in class, we know that it is thorough and well-liked by students."

—Chance (review of the first edition)

SAGE
www.sagepublications.com
Los Angeles • London • New Delhi • Singapore • Washington DC • Boston

ISBN 978-1-4522-0566-3
9 781452 205663 90000

Cover illustration: John Fox

Contents

Preface xv

About the Author xxiv

1. Statistical Models and Social Science 1

1.1 Statistical Models and Social Reality 1
1.2 Observation and Experiment 4
1.3 Populations and Samples 8
Exercise 10
Summary 10
Recommended Reading 11

I. DATA CRAFT 12

2. What Is Regression Analysis? 13

2.1 Preliminaries 15
2.2 Naive Nonparametric Regression 18
2.3 Local Averaging 22
Exercise 25
Summary 26

3. Examining Data 28

3.1 Univariate Displays 30
3.1.1 Histograms 30
3.1.2 Nonparametric Density Estimation 33
3.1.3 Quantile-Comparison Plots 37
3.1.4 Boxplots 41
3.2 Plotting Bivariate Data 44
3.3 Plotting Multivariate Data 47
3.3.1 Scatterplot Matrices 48
3.3.2 Coded Scatterplots 50
3.3.3 Three-Dimensional Scatterplots 50
3.3.4 Conditioning Plots 51
Exercises 53
Summary 53
Recommended Reading 54

4. Transforming Data 55

4.1 The Family of Powers and Roots 55
4.2 Transforming Skewness 59
4.3 Transforming Nonlinearity 63

4.4 Transforming Nonconstant Spread 70
4.5 Transforming Proportions 72
4.6 Estimating Transformations as Parameters* 76
Exercises 78
Summary 79
Recommended Reading 80

II. LINEAR MODELS AND LEAST SQUARES 81

5. Linear Least-Squares Regression 82

5.1 Simple Regression 83
5.1.1 Least-Squares Fit 83
5.1.2 Simple Correlation 87
5.2 Multiple Regression 92
5.2.1 Two Explanatory Variables 92
5.2.2 Several Explanatory Variables 96
5.2.3 Multiple Correlation 98
5.2.4 Standardized Regression Coefficients 100
Exercises 102
Summary 105

6. Statistical Inference for Regression 106

6.1 Simple Regression 106
6.1.1 The Simple-Regression Model 106
6.1.2 Properties of the Least-Squares Estimator 109
6.1.3 Confidence Intervals and Hypothesis Tests 111
6.2 Multiple Regression 112
6.2.1 The Multiple-Regression Model 112
6.2.2 Confidence Intervals and Hypothesis Tests 113
6.3 Empirical Versus Structural Relations 117
6.4 Measurement Error in Explanatory Variables* 120
Exercises 123
Summary 126

7. Dummy-Variable Regression 128

7.1 A Dichotomous Factor 128
7.2 Polytomous Factors 133
7.2.1 Coefficient Quasi-Variances* 138
7.3 Modeling Interactions 140
7.3.1 Constructing Interaction Regressors 141
7.3.2 The Principle of Marginality 144
7.3.3 Interactions With Polytomous Factors 145
7.3.4 Interpreting Dummy-Regression Models With Interactions 145
7.3.5 Hypothesis Tests for Main Effects and Interactions 146
7.4 A Caution Concerning Standardized Coefficients 149
Exercises 150
Summary 151

8. Analysis of Variance 153

8.1 One-Way Analysis of Variance 153
8.1.1 Example: Duncan's Data on Occupational Prestige 155
8.1.2 The One-Way ANOVA Model 156

8.2 Two-Way Analysis of Variance 159
8.2.1 Patterns of Means in the Two-Way Classification 160
8.2.2 Two-Way ANOVA by Dummy Regression 166
8.2.3 The Two-Way ANOVA Model 168
8.2.4 Fitting the Two-Way ANOVA Model to Data 170
8.2.5 Testing Hypotheses in Two-Way ANOVA 172
8.2.6 Equal Cell Frequencies 174
8.2.7 Some Cautionary Remarks 175
8.3 Higher-Way Analysis of Variance 177
8.3.1 The Three-Way Classification 177
8.3.2 Higher-Order Classifications 180
8.3.3 Empty Cells in ANOVA 186
8.4 Analysis of Covariance 187
8.5 Linear Contrasts of Means 190
Exercises 194
Summary 200

9. Statistical Theory for Linear Models* **202**

9.1 Linear Models in Matrix Form 202
9.1.1 Dummy Regression and Analysis of Variance 203
9.1.2 Linear Contrasts 206
9.2 Least-Squares Fit 208
9.2.1 Deficient-Rank Parametrization of Linear Models 210
9.3 Properties of the Least-Squares Estimator 211
9.3.1 The Distribution of the Least-Squares Estimator 211
9.3.2 The Gauss-Markov Theorem 212
9.3.3 Maximum-Likelihood Estimation 214
9.4 Statistical Inference for Linear Models 215
9.4.1 Inference for Individual Coefficients 215
9.4.2 Inference for Several Coefficients 216
9.4.3 General Linear Hypotheses 219
9.4.4 Joint Confidence Regions 220
9.5 Multivariate Linear Models 225
9.6 Random Regressors 227
9.7 Specification Error 229
9.8 Instrumental Variables and Two-Stage Least Squares 231
9.8.1 Instrumental-Variables Estimation in Simple Regression 231
9.8.2 Instrumental-Variables Estimation in Multiple Regression 232
9.8.3 Two-Stage Least Squares 234
Exercises 236
Summary 241
Recommended Reading 243

10. The Vector Geometry of Linear Models* **245**

10.1 Simple Regression 245
10.1.1 Variables in Mean Deviation Form 247
10.1.2 Degrees of Freedom 250
10.2 Multiple Regression 252
10.3 Estimating the Error Variance 256
10.4 Analysis-of-Variance Models 258
Exercises 260

Summary 262
Recommended Reading 264

III. LINEAR-MODEL DIAGNOSTICS 265

11. Unusual and Influential Data 266
11.1 Outliers, Leverage, and Influence 266
11.2 Assessing Leverage: Hat-Values 270
11.3 Detecting Outliers: Studentized Residuals 272
11.3.1 Testing for Outliers in Linear Models 273
11.3.2 Anscombe's Insurance Analogy 274
11.4 Measuring Influence 276
11.4.1 Influence on Standard Errors 277
11.4.2 Influence on Collinearity 280
11.5 Numerical Cutoffs for Diagnostic Statistics 280
11.5.1 Hat-Values 281
11.5.2 Studentized Residuals 281
11.5.3 Measures of Influence 281
11.6 Joint Influence 282
11.6.1 Added-Variable Plots 282
11.6.2 Forward Search 286
11.7 Should Unusual Data Be Discarded? 288
11.8 Some Statistical Details* 289
11.8.1 Hat-Values and the Hat-Matrix 289
11.8.2 The Distribution of the Least-Squares Residuals 290
11.8.3 Deletion Diagnostics 290
11.8.4 Added-Variable Plots and Leverage Plots 291
Exercises 293
Summary 294
Recommended Reading 294

12. Diagnosing Non-Normality, Nonconstant Error Variance, and Nonlinearity 296
12.1 Non-Normally Distributed Errors 297
12.1.1 Confidence Envelopes by Simulated Sampling* 300
12.2 Nonconstant Error Variance 301
12.2.1 Residual Plots 301
12.2.2 Weighted-Least-Squares Estimation* 304
12.2.3 Correcting OLS Standard Errors for Nonconstant Variance* 305
12.2.4 How Nonconstant Error Variance Affects the OLS Estimator* 306
12.3 Nonlinearity 307
12.3.1 Component-Plus-Residual Plots 308
12.3.2 Component-Plus-Residual Plots for Models With Interactions 313
12.3.3 When Do Component-Plus-Residual Plots Work? 314
12.4 Discrete Data 318
12.4.1 Testing for Nonlinearity ("Lack of Fit") 318
12.4.2 Testing for Nonconstant Error Variance 322
12.5 Maximum-Likelihood Methods* 323
12.5.1 Box-Cox Transformation of Y 324
12.5.2 Box-Tidwell Transformation of the Xs 326
12.5.3 Nonconstant Error Variance Revisited 329
12.6 Structural Dimension 331

Exercises 334
Summary 338
Recommended Reading 339

13. Collinearity and Its Purported Remedies 341

13.1 Detecting Collinearity 342
13.1.1 Principal Components* 348
13.1.2 Generalized Variance Inflation* 357
13.2 Coping With Collinearity: No Quick Fix 358
13.2.1 Model Respecification 359
13.2.2 Variable Selection 359
13.2.3 Biased Estimation 361
13.2.4 Prior Information About the Regression Coefficients 364
13.2.5 Some Comparisons 365
Exercises 366
Summary 368

IV. GENERALIZED LINEAR MODELS 369

14. Logit and Probit Models for Categorical Response Variables 370

14.1 Models for Dichotomous Data 370
14.1.1 The Linear-Probability Model 372
14.1.2 Transformations of π: Logit and Probit Models 375
14.1.3 An Unobserved-Variable Formulation 379
14.1.4 Logit and Probit Models for Multiple Regression 380
14.1.5 Estimating the Linear Logit Model* 389
14.2 Models for Polytomous Data 392
14.2.1 The Polytomous Logit Model 392
14.2.2 Nested Dichotomies 399
14.2.3 Ordered Logit and Probit Models 400
14.2.4 Comparison of the Three Approaches 407
14.3 Discrete Explanatory Variables and Contingency Tables 408
14.3.1 The Binomial Logit Model* 411
Exercises 413
Summary 415
Recommended Reading 416

15. Generalized Linear Models 418

15.1 The Structure of Generalized Linear Models 418
15.1.1 Estimating and Testing GLMs 425
15.2 Generalized Linear Models for Counts 427
15.2.1 Models for Overdispersed Count Data 431
15.2.2 Loglinear Models for Contingency Tables 434
15.3 Statistical Theory for Generalized Linear Models* 443
15.3.1 Exponential Families 443
15.3.2 Maximum-Likelihood Estimation of Generalized Linear Models 445
15.3.3 Hypothesis Tests 449
15.3.4 Effect Displays 453
15.4 Diagnostics for Generalized Linear Models 453
15.4.1 Outlier, Leverage, and Influence Diagnostics 454
15.4.2 Nonlinearity Diagnostics 456

15.4.3 Collinearity Diagnostics* 459
15.5 Analyzing Data From Complex Sample Surveys 460
Exercises 464
Summary 468
Recommended Reading 471

V. EXTENDING LINEAR AND GENERALIZED LINEAR MODELS 473

16. Time-Series Regression and Generalized Least Squares* 474

16.1 Generalized Least-Squares Estimation 475
16.2 Serially Correlated Errors 476
16.2.1 The First-Order Autoregressive Process 477
16.2.2 Higher-Order Autoregressive Processes 481
16.2.3 Moving-Average and Autoregressive-Moving-Average Processes 482
16.2.4 Partial Autocorrelations 485
16.3 GLS Estimation With Autocorrelated Errors 485
16.3.1 Empirical GLS Estimation 487
16.3.2 Maximum-Likelihood Estimation 487
16.4 Correcting OLS Inference for Autocorrelated Errors 488
16.5 Diagnosing Serially Correlated Errors 489
16.6 Concluding Remarks 494
Exercises 496
Summary 499
Recommended Reading 500

17. Nonlinear Regression 502

17.1 Polynomial Regression 503
17.1.1 A Closer Look at Quadratic Surfaces* 506
17.2 Piece-wise Polynomials and Regression Splines 507
17.3 Transformable Nonlinearity 512
17.4 Nonlinear Least Squares* 515
17.4.1 Minimizing the Residual Sum of Squares 516
17.4.2 An Illustration: U.S. Population Growth 519
Exercises 521
Summary 526
Recommended Reading 527

18. Nonparametric Regression 528

18.1 Nonparametric Simple Regression: Scatterplot Smoothing 528
18.1.1 Kernel Regression 528
18.1.2 Local-Polynomial Regression 532
18.1.3 Smoothing Splines* 549
18.2 Nonparametric Multiple Regression 550
18.2.1 Local-Polynomial Multiple Regression 550
18.2.2 Additive Regression Models 563
18.3 Generalized Nonparametric Regression 572
18.3.1 Local Likelihood Estimation* 572
18.3.2 Generalized Additive Models 575
Exercises 578
Summary 580
Recommended Reading 585

19. Robust Regression* **586**

19.1 *M* Estimation 586
19.1.1 Estimating Location 586
19.1.2 *M* Estimation in Regression 592
19.2 Bounded-Influence Regression 595
19.3 Quantile Regression 597
19.4 Robust Estimation of Generalized Linear Models 600
19.5 Concluding Remarks 601
Exercises 601
Summary 603
Recommended Reading 604

20. Missing Data in Regression Models **605**

20.1 Missing Data Basics 606
20.1.1 An Illustration 607
20.2 Traditional Approaches to Missing Data 609
20.3 Maximum-Likelihood Estimation for Data Missing at Random* 613
20.3.1 The EM Algorithm 616
20.4 Bayesian Multiple Imputation 619
20.4.1 Inference for Individual Coefficients 621
20.4.2 Inference for Several Coefficients* 624
20.4.3 Practical Considerations 625
20.4.4 Example: A Regression Model for Infant Mortality 626
20.5 Selection Bias and Censoring 629
20.5.1 Truncated- and Censored-Normal Distributions 629
20.5.2 Heckman's Selection-Regression Model 632
20.5.3 Censored-Regression Models 637
Exercises 639
Summary 643
Recommended Reading 646

21. Bootstrapping Regression Models **647**

21.1 Bootstrapping Basics 647
21.2 Bootstrap Confidence Intervals 655
21.2.1 Normal-Theory Intervals 655
21.2.2 Percentile Intervals 655
21.2.3 Improved Bootstrap Intervals 656
21.3 Bootstrapping Regression Models 658
21.4 Bootstrap Hypothesis Tests* 660
21.5 Bootstrapping Complex Sampling Designs 662
21.6 Concluding Remarks 663
Exercises 664
Summary 667
Recommended Reading 668

22. Model Selection, Averaging, and Validation **669**

22.1 Model Selection 669
22.1.1 Model Selection Criteria 671
22.1.2 An Illustration: Baseball Salaries 681
22.1.3 Comments on Model Selection 683

22.2 Model Averaging* 685
22.2.1 Application to the Baseball Salary Data 687
22.2.2 Comments on Model Averaging 687
22.3 Model Validation 690
22.3.1 An Illustration: Refugee Appeals 691
22.3.2 Comments on Model Validation 693
Exercises 694
Summary 696
Recommended Reading 698

VI. MIXED-EFFECTS MODELS 699

23. Linear Mixed-Effects Models for Hierarchical and Longitudinal Data 700

23.1 Hierarchical and Longitudinal Data 701
23.2 The Linear Mixed-Effects Model 702
23.3 Modeling Hierarchical Data 704
23.3.1 Formulating a Mixed Model 708
23.3.2 Random-Effects One-Way Analysis of Variance 710
23.3.3 Random-Coefficients Regression Model 712
23.3.4 Coefficients-as-Outcomes Model 714
23.4 Modeling Longitudinal Data 717
23.5 Wald Tests for Fixed Effects 724
23.6 Likelihood-Ratio Tests of Variance and Covariance Components 726
23.7 Centering Explanatory Variables, Contextual Effects, and Fixed-Effects Models 727
23.7.1 Fixed Versus Random Effects 730
23.8 BLUPs 733
23.9 Statistical Details* 734
23.9.1 The Laird-Ware Model in Matrix Form 734
23.9.2 Wald Tests Revisited 737
Exercises 738
Summary 740
Recommended Reading 741

24. Generalized Linear and Nonlinear Mixed-Effects Models 743

24.1 Generalized Linear Mixed Models 743
24.1.1 Example: Migraine Headaches 745
24.1.2 Statistical Details* 748
24.2 Nonlinear Mixed Models* 749
24.2.1 Example: Recovery From Coma 751
24.2.2 Estimating Nonlinear Mixed Models 755
Exercises 757
Summary 757
Recommended Reading 758

Appendix A 759

References 762

Author Index 773

Subject Index 777

Data Set Index 791