Applied Regression
Including Computing
and Graphics

R. DENNIS COOK

SANFORD WEISBERG
The University of Minnesota
St. Paul, Minnesota
1.8 Complements, 22
Problems, 22

2 Introduction to Regression

2.1 Using Boxplots to Study Length | Age, 28
2.2 Using a Scatterplot to Study Length | Age, 31
2.3 Mouse Modes, 31
   2.3.1 Show Coordinates Mouse Mode, 32
   2.3.2 Slicing Mode, 32
   2.3.3 Brushing Mode, 33
2.4 Characterizing Length | Age, 33
2.5 Mean and Variance Functions, 35
   2.5.1 Mean Function, 35
   2.5.2 Variance Function, 36
2.6 Highlights, 37
2.7 Complements, 37
Problems, 37

3 Introduction to Smoothing

3.1 Slicing a Scatterplot, 40
3.2 Estimating E(y | x) by Slicing, 42
3.3 Estimating E(y | x) by Smoothing, 42
3.4 Checking a Theory, 45
3.5 Boxplots, 45
3.6 Snow Geese, 48
   3.6.1 Snow Goose Regression, 49
   3.6.2 Mean Function, 51
   3.6.3 Variance Function, 51
3.7 Complements, 53
Problems, 53

4 Bivariate Distributions

4.1 General Bivariate Distributions, 56
   4.1.1 Bivariate Densities, 58
   4.1.2 Connecting with Regression, 59
   4.1.3 Independence, 59
   4.1.4 Covariance, 60
   4.1.5 Correlation Coefficient, 62
CONTENTS

4.2 Bivariate Normal Distribution, 63
  4.2.1 Correlation Coefficient in Normal Populations, 64
  4.2.2 Correlation Coefficient in Non-normal Populations, 68
4.3 Regression in Bivariate Normal Populations, 69
  4.3.1 Mean Function, 70
  4.3.2 Mean Function in Standardized Variables, 70
  4.3.3 Mean Function as a Straight Line, 72
  4.3.4 Variance Function, 74
4.4 Smoothing Bivariate Normal Data, 76
4.5 Complements, 78
  4.5.1 Confidence Interval for a Correlation, 78
  4.5.2 References, 78
Problems, 78

5 Two-Dimensional Plots
  5.1 Aspect Ratio and Focusing, 81
  5.2 Power Transformations, 84
  5.3 Thinking about Power Transformations, 86
  5.4 Log Transformations, 87
  5.5 Showing Labels and Coordinates, 88
  5.6 Linking Plots, 89
  5.7 Point Symbols and Colors, 90
  5.8 Brushing, 90
  5.9 Name Lists, 90
  5.10 Probability Plots, 90
  5.11 Complements, 92
Problems, 93

PART II TOOLS

6 Simple Linear Regression
  6.1 Simple Linear Regression, 98
  6.2 Least Squares Estimation, 101
    6.2.1 Notation, 101
    6.2.2 The Least Squares Criterion, 102
    6.2.3 Ordinary Least Squares Estimators, 105
    6.2.4 More on Sample Correlation, 106
CONTENTS

6.2.5 Some Properties of Least Squares Estimates, 106
6.2.6 Estimating the Common Variance, $\sigma^2$, 107
6.2.7 Summary, 107

6.3 Using $\text{Arc}$, 107
6.3.1 Interpreting the Intercept, 110

6.4 Inference, 112
6.4.1 Inferences about Parameters, 112
6.4.2 Estimating Population Means, 115
6.4.3 Prediction, 117

6.5 Forbes' Experiments, Revisited, 118

6.6 Model Comparison, 120
6.6.1 Models, 120
6.6.2 Analysis of Variance, 122

6.7 Complements, 125
6.7.1 Derivation of Estimates, 125
6.7.2 Means and Variances of Estimates, 126
6.7.3 Why Least Squares?, 128
6.7.4 Alternatives to Least Squares, 129
6.7.5 Accuracy of Estimates, 130
6.7.6 Role of Normality, 130
6.7.7 Measurement Error, 130
6.7.8 References, 132

Problems, 132

7 Introduction to Multiple Linear Regression

7.1 The Scatterplot Matrix, 140
7.1.1 Pairs of Variables, 141
7.1.2 Separated Points, 142
7.1.3 Marginal Response Plots, 143
7.1.4 Extracting Plots, 145

7.2 Terms and Predictors, 145

7.3 Examples, 147
7.3.1 Simple Linear Regression, 147
7.3.2 Polynomial Mean Functions with One Predictor, 148
7.3.3 Two Predictors, 150
7.3.4 Polynomial Mean Functions with Two Predictors, 151
7.3.5 Many Predictors, 151
7.4 Multiple Linear Regression, 152
7.5 Estimation of Parameters, 153
7.6 Inference, 158
  7.6.1 Tests and Confidence Statements about Parameters, 159
  7.6.2 Prediction, 160
  7.6.3 Leverage and Extrapolation, 161
  7.6.4 General Linear Combinations, 163
  7.6.5 Overall Analysis of Variance, 164
  7.6.6 The Coefficient of Determination, 165
7.7 The Lake Mary Data, 166
7.8 Regression Through the Origin, 167
7.9 Complements, 168
  7.9.1 An Introduction to Matrices, 168
  7.9.2 Random Vectors, 172
  7.9.3 Correlation Matrix, 173
  7.9.4 Applications to Multiple Linear Regression, 173
  7.9.5 Ordinary Least Squares Estimates, 174
  7.9.6 References, 178

Problems, 178

8 Three-Dimensional Plots 185
8.1 Viewing a Three-Dimensional Plot, 185
  8.1.1 Rotation Control, 187
  8.1.2 Recalling Views, 187
  8.1.3 Rocking, 187
  8.1.4 Show Axes, 188
  8.1.5 Depth Cuing, 188
  8.1.6 Zooming, 188
8.2 Adding a Polynomial Surface, 188
  8.2.1 Parametric Smoother Slidebar, 188
  8.2.2 Extracting Fitted Values, 189
  8.2.3 Adding a Function, 189
  8.2.4 Residuals, 190
8.3 Scaling and Centering, 190
8.4 2D Plots from 3D Plots, 191
  8.4.1 Saving a Linear Combination, 192
  8.4.2 Rotation in 2D, 192
8.4.3 Extracting a 2D Plot, 194
8.4.4 Summary, 194
8.5 Removing a Linear Trend in 3D Plots, 194
8.6 Using Uncorrelated Variables, 196
8.7 Complements, 198
Problems, 199

9 Weights and Lack-of-Fit 202

9.1 Snow Geese, 202
  9.1.1 Visually Assessing Lack-of-Fit, 202
  9.1.2 Nonconstant Variances, 204
9.2 Weighted Least Squares, 204
  9.2.1 Particle Physics Example, 206
  9.2.2 Predictions, 209
9.3 Lack-of-Fit Methods, 210
  9.3.1 Visual Lack-of-Fit with Smooths, 211
  9.3.2 Lack-of-Fit Based on Variance, 212
  9.3.3 Variance Known, 213
  9.3.4 External Estimates of Variation, 214
  9.3.5 Replicate Observations, 214
  9.3.6 Subsampling, 217
9.4 Fitting with Subpopulation Averages, 217
9.5 Complements, 219
  9.5.1 Weighted Least Squares, 219
  9.5.2 The lowess Smoother, 220
  9.5.3 References, 220
Problems, 221

10 Understanding Coefficients 230

10.1 Interpreting Coefficients, 230
  10.1.1 Rescaling, 230
  10.1.2 Rate of Change, 231
  10.1.3 Reparameterization, 232
  10.1.4 Nonlinear Functions of Terms, 234
  10.1.5 Variances of Coefficient Estimates, 234
  10.1.6 Standardization of Terms, 235
10.2 The Multivariate Normal Distribution, 235
10.3 Sampling Distributions, 237
10.4 Correlation Versus Causation and the Sleep Data, 238
   10.4.1 Missing Data, 239
   10.4.2 The Mean Function, 240
   10.4.3 The Danger Indicator, 240
   10.4.4 Interpretation, 242
10.5 2D Added-Variable Plots, 243
   10.5.1 Adding a Predictor to Simple Regression, 244
   10.5.2 Added-Variable Plots in Arc, 247
10.6 Properties of 2D Added-Variable Plots, 247
   10.6.1 Intercept, 247
   10.6.2 Slope, 247
   10.6.3 Residuals, 248
   10.6.4 Sample Partial Correlation, 248
   10.6.5 t-Statistics, 248
   10.6.6 Three Extreme Cases, 248
10.7 3D Added-Variable Plots, 250
10.8 Confidence Regions, 250
   10.8.1 Confidence Regions for Two Coefficient Estimates, 251
   10.8.2 Bivariate Confidence Regions When the Mean Function Has Many Terms, 254
   10.8.3 General Confidence Regions, 255
10.9 Complements, 256
   10.9.1 Missing Data, 256
   10.9.2 Causation, Association, and Experimental Designs, 256
   10.9.3 Net Effects Plots, 256
   10.9.4 References, 257

Problems, 257

11 Relating Mean Functions

11.1 Removing Terms, 263
   11.1.1 Marginal Mean Functions, 264
   11.1.2 Marginal Variance Functions, 265
   11.1.3 Example, 266
11.2 Tests to Compare Models, 266
11.3 Highway Accident Data, 267
   11.3.1 Testing Equality of Coefficients, 269
   11.3.2 Offsets, 270
11.4 Sequential Fitting, 271
11.5 Selecting Terms, 272
  11.5.1 Criteria for Selecting Submodels, 274
  11.5.2 Stepwise Methods, 275
  11.5.3 Highway Accident Data, 276
11.6 Complements, 283
Problems, 283

12 Factors and Interactions 287

12.1 Factors, 287
  12.1.1 Two Levels, 287
  12.1.2 Many Levels, 288
12.2 Twin Data, 288
12.3 One-Way Analysis of Variance, 290
12.4 Models with Categorical and Continuous Predictors, 292
  12.4.1 Fitting, 294
  12.4.2 Tests, 296
12.5 Turkey Diets, 297
  12.5.1 The Zero Dose, 298
  12.5.2 Adapting to Curvature, 298
12.6 Casuarina Data, 299
  12.6.1 Effect Through the Intercept, 301
  12.6.2 Effect Through Intercept and Slope, 304
12.7 Factorial Experiments, 305
12.8 Complements, 308
  12.8.1 Alternate Definitions of Factors, 308
  12.8.2 Comparing Slopes from Separate Fits, 309
  12.8.3 References, 309
Problems, 310

13 Response Transformations 316

13.1 Response Transformations, 316
  13.1.1 Variance Stabilizing Transformations, 317
  13.1.2 Transforming to Linearity with One Predictor, 317
  13.1.3 Inverse Fitted Value Plot, 320
  13.1.4 Numerical Choice of Transformation, 321
13.2 Transformations to Normality, 324
  13.2.1 Visual Choice of Transformation, 324
13.2.2 Automatic Choice of Transformations, 326
13.2.3 Possible Routes, 329
13.3 Complements, 329
13.3.1 The Box–Cox Method, 329
13.3.2 Profile Log-Likelihoods and Confidence Curves, 330
13.3.3 Transformation Families, 330
13.3.4 References, 331

Problems, 332

14 Diagnostics I: Curvature and Nonconstant Variance 334

14.1 The Residuals, 336
14.1.1 Definitions and Rationale, 336
14.1.2 Residual Plots, 337
14.1.3 Choosing Residual Plots, 339
14.1.4 Examples of Residual Plots, 340
14.1.5 A Note of Caution, 342
14.2 Testing for Curvature, 343
14.3 Testing for Nonconstant Variance, 346
14.3.1 Transactions Data, 347
14.3.2 Caution Data, 349
14.4 Complements, 350

Problems, 350

15 Diagnostics II: Influence and Outliers 354

15.1 Adaptive Score Data, 356
15.2 Influential Cases and Cook’s Distance, 357
15.3 Residuals, 360
15.3.1 Studentized Residuals, 360
15.3.2 Cook’s Distance Again, 360
15.4 Outliers, 361
15.4.1 Testing for a Single Outlier, 362
15.4.2 Checking Every Case, 364
15.4.3 Adaptive Score Data, 364
15.5 Fuel Data, 365
15.6 Complements, 368
15.6.1 Updating Formula, 368
15.6.2 Local Influence, 368
15.6.3 References, 369
Problems, 369

16 Predictor Transformations 373

16.1 Regression Through Transformation, 373
  16.1.1 Power Curves and Polynomial Fits, 373
  16.1.2 Transformations via Smoothing, 375
  16.1.3 General Formulation, 375

16.2 Ceres Plots, 376
  16.2.1 Constant $E(u_{1j}|u_2)$, No Augmentation, 377
  16.2.2 Linear $E(u_{1j}|u_2)$, Linear Augmentation, 377
  16.2.3 Quadratic $E(u_{1j}|u_2)$, Quadratic Augmentation, 377
  16.2.4 General $E(u_{1j}|u_2)$, Smooth Augmentation, 378

16.3 Berkeley Guidance Study, 378
16.4 Haystack Data, 380
16.5 Transforming Multiple Terms, 383
  16.5.1 Estimating Additive Transformations of Several Terms, 383
  16.5.2 Assessing the Transformations, 384

16.6 Ceres Plots with Smooth Augmentation, 384
16.7 Transforming Two Terms Simultaneously, 388
  16.7.1 Models for Transforming Two Terms, 388
  16.7.2 Example: Plant Height, 389

16.8 Complements, 392
  16.8.1 Mixed Forms of $E(u_{1j}|u_2)$, 392
  16.8.2 References, 393

Problems, 393

17 Model Assessment 396

17.1 Model Checking Plots, 397
  17.1.1 Checking Mean Functions, 399
  17.1.2 Checking Variance Functions, 401

17.2 Relation to Residual Plots, 403
17.3 Sleep Data, 404
17.4 Complements, 406

Problems, 407
PART III  REGRESSION GRAPHICS  

18  Visualizing Regression  

18.1 Pine Trees, 411  
18.2 The Estimated 2D Summary Plot, 412  
18.3 Structural Dimension, 413  
  18.3.1 Zero-Dimensional Structure, 413  
  18.3.2 One-Dimensional Structure, 413  
  18.3.3 Two-Dimensional Structure, 416  
18.4 Checking an Estimated Summary Plot, 417  
18.5 More Examples and Refinements, 419  
  18.5.1 Visualizing Linear Regression in 3D Plots, 419  
  18.5.2 Linear Regression Without Linearly Related Predictors, 422  
  18.5.3 More on Ordinary Least Squares Summary Views, 423  
18.6 Complements, 425  
Problems, 425  

19  Visualizing Regression with Many Predictors  

19.1 Linearly Related Predictors, 430  
19.2 Checking Linearly Related Predictors, 431  
19.3 Linearly Related Predictors and the 1D Model, 432  
19.4 Transforming to Get Linearly Related Predictors, 433  
19.5 Finding Dimension Graphically, 434  
  19.5.1 The Inverse Regression Curve, 434  
  19.5.2 Inverse Marginal Response Plots, 436  
19.6 Australian Athletes Data, 438  
19.7 Complements, 441  
  19.7.1 Sliced Inverse Regression, 441  
  19.7.2 References, 442  
Problems, 442  

20  Graphical Regression  

20.1 Overview of Graphical Regression, 446  
20.2 Mussels’ Muscles, 447  
  20.2.1 The GREG Predictors, 447  
  20.2.2 Graphical Regression, 448
20.3 Reaction Yield, 452
  20.3.1 Linearly Related Predictors, 453
  20.3.2 Graphical Regression, 454
  20.3.3 Continuing the Analysis, 454

20.4 Variations, 457
  20.4.1 Standardizing the Linear Predictors, 457
  20.4.2 Improving Resolution in 3D Added-Variable Plots, 457
  20.4.3 Model Checking, 458
  20.4.4 Using the Linearly Related Predictors, 460

20.5 Complements, 461
  20.5.1 GREG Predictors and Principal Hessian Directions, 461
  20.5.2 References, 462

Problems, 462

PART IV LOGISTIC REGRESSION AND GENERALIZED LINEAR MODELS

21 Binomial Regression

  21.1 Recumbent Cows, 467
     21.1.1 Categorical Predictors, 468
     21.1.2 Continuous Predictors, 470
  21.2 Probability Models for Counted Data, 471
     21.2.1 The Bernoulli Distribution, 471
     21.2.2 Binomial Random Variables, 472
     21.2.3 Inference, 474
  21.3 Binomial Regression, 475
     21.3.1 Mean Functions for Binomial Regression, 476
     21.3.2 Summary, 477
  21.4 Fitting Logistic Regression, 478
     21.4.1 Understanding Coefficients, 480
     21.4.2 Many Terms, 482
     21.4.3 Deviance, 483
     21.4.4 Goodness-of-Fit Tests, 485
  21.5 Weevil Preferences, 486
  21.6 Complements, 489
     21.6.1 Normal Approximation to the Binomial, 489
     21.6.2 Smoothing a Binary Response, 490
22 Graphical and Diagnostic Methods for Logistic Regression

22.1 One-Predictor Methods, 497
22.1.1 Jittering to See Relative Density, 498
22.1.2 Using the Conditional Density of $x | y$, 498
22.1.3 Logistic Regression from Conditional Densities, 500
22.1.4 Specific Conditional Densities, 500
22.1.5 Implications for the Recumbent Cow Data, 501

22.2 Visualizing Logistic Regression with Two or More Predictors, 504
22.2.1 Assessing the Predictors, 505
22.2.2 Assessing a Logistic Model with Two Predictors, 506
22.2.3 Assessing a Logistic Model with Three Predictors, 507

22.3 Transforming Predictors, 509
22.3.1 Guidelines, 509
22.3.2 Transforming $x | y$ to Multivariate Normality, 510

22.4 Diagnostic Methods, 512
22.4.1 Residual Plots, 512
22.4.2 Influence, 513
22.4.3 Model Checking Plots, 514

22.5 Adding Factors, 517

22.6 Extending Predictor Transformations, 519
22.6.1 Power Transformations with a Binomial Response, 519
22.6.2 Ceres Plots, 519

22.7 Complements, 519
22.7.1 Marginal Odds Ratio, 519
22.7.2 Relative Density, 520
22.7.3 Deviance Residuals, 520
22.7.4 Outliers, 520
22.7.5 Overdispersion, 521
22.7.6 Graphical Regression, 521
22.7.7 References, 522

Problems, 522

23 Generalized Linear Models 525

23.1 Components of a Generalized Linear Model, 525
23.2 Normal Models, 527
  23.2.1 Transformation of Parameters, 531
  23.2.2 Transformation to Simple Linear Regression, 531
23.3 Poisson Regression, 532
  23.3.1 Log-Linear Models, 538
23.4 Gamma Regression, 539
23.5 Complements, 540
  23.5.1 Poisson Distribution, 540
  23.5.2 Gamma Distribution, 542
  23.5.3 References, 542

Problems, 542

Appendix A Arc 545

A.1 Getting the Software, 545
  A.1.1 Macintosh OS, 545
  A.1.2 Windows OS, 547
  A.1.3 Unix, 548
  A.1.4 What You Get, 548
  A.1.5 Data Files, 548
A.2 The Text Window, 549
  A.2.1 Typing in the Text Window, 549
  A.2.2 Typing Data, 549
  A.2.3 Working with Lists, 551
  A.2.4 Calculating the Slope and Intercept, 552
A.3 Saving and Printing, 553
  A.3.1 Text, 553
  A.3.2 Graphics, 554
A.4 Quitting, 554
A.5 Data Files, 554
  A.5.1 Plain Data, 554
  A.5.2 Plain Data File with Variable Labels, 555
CONTENTS

A.5.3 Importing Data from a Spreadsheet, 555
A.5.4 Special Characters, 555
A.5.5 Getting into Trouble with Plain Data Files, 555
A.5.6 Formatted Data File, 556
A.5.7 Creating a Data Set from the Text Window, 558
A.5.8 Old-Style Data Files, 558
A.5.9 Missing Values, 559

A.6 The Arc Menu, 559

A.7 The Data Set Menu, 560
  A.7.1 Description of Data, 560
  A.7.2 Modifying Data, 561

A.8 Graphics from the Graph&Fit Menu, 562
  A.8.1 Histograms and Plot Controls, 563
  A.8.2 Two-Dimensional Plots and Plot Controls, 564
  A.8.3 Three-Dimensional Plots, 566
  A.8.4 Boxplots, 566
  A.8.5 Scatterplot Matrices, 566

A.9 Fitting Models, 566

A.10 Model Menüs, 567

A.11 Adding Statistics to a Data Set, 567

A.12 Some Useful Functions, 567
  A.12.1 Getting Help, 570

References 571

Author Index 579

Subject Index 583