Contents

1 Statistical Models 1
1.1 Introduction and Overview 1
1.2 Conventions for Describing Data 1
1.3 Plotting Data 5
1.4 Coding for Factors 10
1.5 Statistical Models Describe Both Random and Systematic Features of Data 11
1.6 Regression Models 12
1.7 Interpreting Regression Models 16
1.8 All Models Are Wrong, but Some Are Useful 17
1.9 The Purpose of a Statistical Model Affects How It Is Developed 18
1.10 Accuracy vs Parsimony 19
1.11 Experiments vs Observational Studies: Causality vs Association 21
1.12 Data Collection and Generalizability 22
1.13 Using R for Statistical Modelling 23
1.14 Summary 24
Problems 25
References 29

2 Linear Regression Models 31
2.1 Introduction and Overview 31
2.2 Linear Regression Models Defined 31
2.3 Simple Linear Regression 35
2.3.1 Least-Squares Estimation 35
2.3.2 Coefficient Estimates 36
2.3.3 Estimating the Variance σ^2 38
2.3.4 Standard Errors of the Coefficients 39
2.3.5 Standard Errors of Fitted Values 39

2.4 Estimation for Multiple Regression 40
2.4.1 Coefficient Estimates 40
2.4.2 Estimating the Variance σ^2 42
2.4.3 Standard Errors 42
* 2.5 Matrix Formulation of Linear Regression Models 43
* 2.5.1 Matrix Notation 43
* 2.5.2 Coefficient Estimates 44
* 2.5.3 Estimating the Variance σ^2 46
* 2.5.4 Estimating the Variance of $\hat{\boldsymbol{\beta}}$ 47
* 2.5.5 Estimating the Variance of Fitted Values 47
2.6 Fitting Linear Regression Models Using R 48
2.7 Interpreting the Regression Coefficients 52
2.8 Inference for Linear Regression Models: t-Tests 53
2.8.1 Normal Linear Regression Models 53
2.8.2 The Distribution of $\hat{\beta}_j$ 53
2.8.3 Hypothesis Tests for β_j 54
2.8.4 Confidence Intervals for β_j 55
2.8.5 Confidence Intervals for μ 56
2.9 Analysis of Variance for Regression Models 58
2.10 Comparing Nested Models 61
2.10.1 Analysis of Variance to Compare Two Nested Models 61
2.10.2 Sequential Analysis of Variance 63
2.10.3 Parallel and Independent Regressions 66
2.10.4 The Marginality Principle 70
2.11 Choosing Between Non-nested Models: AIC and BIC 70
2.12 Tools to Assist in Model Selection 72
2.12.1 Adding and Dropping Variables 72
2.12.2 Automated Methods for Model Selection 73
2.12.3 Objections to Using Stepwise Procedures 76
2.13 Case Study 76
2.14 Using R for Fitting Linear Regression Models 79
2.15 Summary 82
Problems 83
References 90

3 Linear Regression Models: Diagnostics and Model-Building 93
3.1 Introduction and Overview 93
3.2 Assumptions from a Practical Point of View 94
3.2.1 Types of Assumptions 94
3.2.2 The Linear Predictor 94
3.2.3 Constant Variance 94
3.2.4 Independence 95
3.2.5 Normality 96

3.2.6 Measurement Scales . . . 96
3.2.7 Approximations and Consequences . . . 96
3.3 Residuals for Normal Linear Regression Models . . . 97
3.4 The Leverages for Linear Regression Models . . . 98
3.4.1 Leverage and Extreme Covariate Values . . . 98
* 3.4.2 The Leverages Using Matrix Algebra . . . 100
3.5 Residual Plots . . . 101
3.5.1 Plot Residuals Against x_j: Linearity . . . 101
3.5.2 Partial Residual Plots . . . 102
3.5.3 Plot Residuals Against $\hat{\mu}$: Constant Variance . . . 104
3.5.4 Q–Q Plots and Normality . . . 105
3.5.5 Lag Plots and Dependence over Time . . . 106
3.6 Outliers and Influential Observations . . . 108
3.6.1 Introduction . . . 108
3.6.2 Outliers and Studentized Residuals . . . 109
3.6.3 Influential Observations . . . 110
3.7 Terminology for Residuals . . . 115
3.8 Remedies: Fixing Identified Problems . . . 115
3.9 Transforming the Response . . . 116
3.9.1 Symmetry, Constraints and the Ladder of Powers . . . 116
3.9.2 Variance-Stabilizing Transformations . . . 117
3.9.3 Box–Cox Transformations . . . 120
3.10 Simple Transformations of Covariates . . . 121
3.11 Polynomial Trends . . . 127
3.12 Regression Splines . . . 131
3.13 Fixing Identified Outliers . . . 134
3.14 Collinearity . . . 135
3.15 Case Studies . . . 138
3.15.1 Case Study 1 . . . 138
3.15.2 Case Study 2 . . . 141
3.16 Using R for Diagnostic Analysis of Linear Regression Models . . . 146
3.17 Summary . . . 147
Problems . . . 149
References . . . 162

4 Beyond Linear Regression: The Method of Maximum Likelihood . . . 165
4.1 Introduction and Overview . . . 165
4.2 The Need for Non-normal Regression Models . . . 165
4.2.1 When Linear Models Are a Poor Choice . . . 165
4.2.2 Binary Outcomes and Binomial Counts . . . 166
4.2.3 Unrestricted Counts: Poisson or Negative Binomial . . . 168
4.2.4 Continuous Positive Observations . . . 169
4.3 Generalizing the Normal Linear Model . . . 171

4.4 The Idea of Likelihood Estimation 172
4.5 Maximum Likelihood for Estimating One Parameter 176
4.5.1 Score Equations 176
4.5.2 Information: Observed and Expected 177
4.5.3 Standard Errors of Parameters 179
4.6 Maximum Likelihood for More Than One Parameter 180
4.6.1 Score Equations 180
4.6.2 Information: Observed and Expected 182
4.6.3 Standard Errors of Parameters 183
* 4.7 Maximum Likelihood Using Matrix Algebra 183
* 4.7.1 Notation 183
* 4.7.2 Score Equations 183
* 4.7.3 Information: Observed and Expected 184
* 4.7.4 Standard Errors of Parameters 186
* 4.8 Fisher Scoring for Computing MLEs 186
4.9 Properties of MLEs 189
4.9.1 Introduction 189
4.9.2 Properties of MLEs for One Parameter 189
* 4.9.3 Properties of MLEs for Many Parameters 190
4.10 Hypothesis Testing: Large Sample Asymptotic Results 191
4.10.1 Introduction 191
* 4.10.2 Global Tests 194
* 4.10.3 Tests About Subsets of Parameters 196
4.10.4 Tests About One Parameter in a Set of Parameters . 197
4.10.5 Comparing the Three Methods 199
4.11 Confidence Intervals 200
* 4.11.1 Confidence Regions for More Than One Parameter .. 200
4.11.2 Confidence Intervals for Single Parameters 200
4.12 Comparing Non-nested Models: The AIC and BIC 202
4.13 Summary 204
* 4.14 Appendix: R Code to Fit Models to the Quilpie Rainfall Data 204
Problems 206
References 209

5 Generalized Linear Models: Structure 211
5.1 Introduction and Overview 211
5.2 The Two Components of Generalized Linear Models 211
5.3 The Random Component: Exponential Dispersion Models ... 212
5.3.1 Examples of EDMs 212
5.3.2 Definition of EDMs 212
5.3.3 Generating Functions 214
5.3.4 The Moment Generating and Cumulant Functions for EDMs 215
5.3.5 The Mean and Variance of an EDM 216

5.3.6 The Variance Function 217
5.4 EDMs in Dispersion Model Form 218
5.4.1 The Unit Deviance and the Dispersion Model Form ... 218
5.4.2 The Saddlepoint Approximation 223
5.4.3 The Distribution of the Unit Deviance 224
5.4.4 Accuracy of the Saddlepoint Approximation 225
5.4.5 Accuracy of the χ_1^2 Distribution for the Unit Deviance 226
5.5 The Systematic Component 229
5.5.1 Link Function 229
5.5.2 Offsets 229
5.6 Generalized Linear Models Defined 230
5.7 The Total Deviance 231
5.8 Regression Transformations Approximate GLMs 232
5.9 Summary 234
Problems 235
References 240

6 **Generalized Linear Models: Estimation** 243
6.1 Introduction and Overview 243
6.2 Likelihood Calculations for β 243
6.2.1 Differentiating the Probability Function 243
6.2.2 Score Equations and Information for β 244
6.3 Computing Estimates of β 245
6.4 The Residual Deviance 248
6.5 Standard Errors for $\hat{\beta}$ 250
* 6.6 Estimation of β: Matrix Formulation 250
6.7 Estimation of GLMs Is Locally Like Linear Regression 252
6.8 Estimating ϕ 252
6.8.1 Introduction 252
6.8.2 The Maximum Likelihood Estimator of ϕ 253
6.8.3 Modified Profile Log-Likelihood Estimator of ϕ 253
6.8.4 Mean Deviance Estimator of ϕ 254
6.8.5 Pearson Estimator of ϕ 255
6.8.6 Which Estimator of ϕ Is Best? 255
6.9 Using R to Fit GLMs 257
6.10 Summary 259
Problems 261
References 262

7 **Generalized Linear Models: Inference** 265
7.1 Introduction and Overview 265
7.2 Inference for Coefficients When ϕ Is Known 265
7.2.1 Wald Tests for Single Regression Coefficients 265
7.2.2 Confidence Intervals for Individual Coefficients 266

7.2.3 Confidence Intervals for μ . . . 267
7.2.4 Likelihood Ratio Tests to Compare Nested Models: χ^2 Tests . . . 269
7.2.5 Analysis of Deviance Tables to Compare Nested Models . . . 270
7.2.6 Score Tests . . . 271
* 7.2.7 Score Tests Using Matrices . . . 272
7.3 Large Sample Asymptotics . . . 273
7.4 Goodness-of-Fit Tests with ϕ Known . . . 274
7.4.1 The Idea of Goodness-of-Fit . . . 274
7.4.2 Deviance Goodness-of-Fit Test . . . 275
7.4.3 Pearson Goodness-of-Fit Test . . . 275
7.5 Small Dispersion Asymptotics . . . 276
7.6 Inference for Coefficients When ϕ Is Unknown . . . 278
7.6.1 Wald Tests for Single Regression Coefficients . . . 278
7.6.2 Confidence Intervals for Individual Coefficients . . . 280
* 7.6.3 Confidence Intervals for μ . . . 281
7.6.4 Likelihood Ratio Tests to Compare Nested Models: F-Tests . . . 282
7.6.5 Analysis of Deviance Tables to Compare Nested Models . . . 284
7.6.6 Score Tests . . . 286
7.7 Comparing Wald, Score and Likelihood Ratio Tests . . . 287
7.8 Choosing Between Non-nested GLMs: AIC and BIC . . . 288
7.9 Automated Methods for Model Selection . . . 289
7.10 Using R to Perform Tests . . . 290
7.11 Summary . . . 292
Problems . . . 293
References . . . 296

8 Generalized Linear Models: Diagnostics . . . 297
8.1 Introduction and Overview . . . 297
8.2 Assumptions of GLMs . . . 297
8.3 Residuals for GLMs . . . 298
8.3.1 Response Residuals Are Insufficient for GLMs . . . 298
8.3.2 Pearson Residuals . . . 299
8.3.3 Deviance Residuals . . . 300
8.3.4 Quantile Residuals . . . 300
8.4 The Leverages in GLMs . . . 304
8.4.1 Working Leverages . . . 304
* 8.4.2 The Hat Matrix . . . 304
8.5 Leverage Standardized Residuals for GLMs . . . 305
8.6 When to Use Which Type of Residual . . . 306
8.7 Checking the Model Assumptions . . . 306
8.7.1 Introduction . . . 306

8.7.2 Independence: Plot Residuals Against Lagged Residuals 307
8.7.3 Plots to Check the Systematic Component 307
8.7.4 Plots to Check the Random Component 311
8.8 Outliers and Influential Observations 312
8.8.1 Introduction 312
8.8.2 Outliers and Studentized Residuals 312
8.8.3 Influential Observations 313
8.9 Remedies: Fixing Identified Problems 315
8.10 Quasi-Likelihood and Extended Quasi-Likelihood 318
8.11 Collinearity 321
8.12 Case Study 322
8.13 Using R for Diagnostic Analysis of GLMs 325
8.14 Summary 326
Problems 327
References 330

9 Models for Proportions: Binomial GLMs 333
9.1 Introduction and Overview 333
9.2 Modelling Proportions 333
9.3 Link Functions 336
9.4 Tolerance Distributions and the Probit Link 338
9.5 Odds, Odds Ratios and the Logit Link 340
9.6 Median Effective Dose, ED50 343
9.7 The Complementary Log-Log Link in Assay Analysis 344
9.8 Overdispersion 347
9.9 When Wald Tests Fail 351
9.10 No Goodness-of-Fit for Binary Responses 354
9.11 Case Study 354
9.12 Using R to Fit GLMs to Proportion Data 360
9.13 Summary 360
Problems 361
References 367

10 Models for Counts: Poisson and Negative Binomial GLMs 371
10.1 Introduction and Overview 371
10.2 Summary of Poisson GLMs 371
10.3 Modelling Rates 373
10.4 Contingency Tables: Log-Linear Models 378
10.4.1 Introduction 378
10.4.2 Two Dimensional Tables: Systematic Component 378
10.4.3 Two-Dimensional Tables: Random Components 380
10.4.4 Three-Dimensional Tables 385
10.4.5 Simpson's Paradox 389
10.4.6 Equivalence of Binomial and Poisson GLMs 392

10.4.7 Higher-Order Tables 393
10.4.8 Structural Zeros in Contingency Tables 395
10.5 Overdispersion 397
10.5.1 Overdispersion for Poisson GLMs 397
10.5.2 Negative Binomial GLMs 399
10.5.3 Quasi-Poisson Models 402
10.6 Case Study 404
10.7 Using R to Fit GLMs to Count Data 411
10.8 Summary 411
Problems 412
References 422

11 Positive Continuous Data: Gamma and Inverse Gaussian GLMs 425
11.1 Introduction and Overview 425
11.2 Modelling Positive Continuous Data 425
11.3 The Gamma Distribution 427
11.4 The Inverse Gaussian Distribution 431
11.5 Link Functions 433
11.6 Estimating the Dispersion Parameter 436
11.6.1 Estimating ϕ for the Gamma Distribution 436
11.6.2 Estimating ϕ for the Inverse Gaussian Distribution . . 439
11.7 Case Studies 440
11.7.1 Case Study 1 440
11.7.2 Case Study 2 442
11.8 Using R to Fit Gamma and Inverse Gaussian GLMS 445
11.9 Summary 445
Problems 446
References 454

12 Tweedie GLMs 457
12.1 Introduction and Overview 457
12.2 The Tweedie EDMs 457
12.2.1 Introducing Tweedie Distributions 457
12.2.2 The Structure of Tweedie EDMs 460
12.2.3 Tweedie EDMs for Positive Continuous Data 461
12.2.4 Tweedie EDMs for Positive Continuous Data with Exact Zeros 463
12.3 Tweedie GLMs 464
12.3.1 Introduction 464
12.3.2 Estimation of the Index Parameter ξ 465
12.3.3 Fitting Tweedie GLMs 469
12.4 Case Studies 473
12.4.1 Case Study 1 473
12.4.2 Case Study 2 475

12.5 Using R to Fit Tweedie GLMs 478
12.6 Summary 479
Problems 480
References 488

13 Extra Problems 491
13.1 Introduction and Overview 491
Problems 491
References 500

Using R for Data Analysis 503
A.1 Introduction and Overview 503
A.2 Preparing to Use R 503
A.2.1 Introduction to R 503
A.2.2 Important R Websites 504
A.2.3 Obtaining and Installing R 504
A.2.4 Downloading and Installing R Packages 504
A.2.5 Using R Packages 505
A.2.6 The R Packages Used in This Book 506
A.3 Introduction to Using R 506
A.3.1 Basic Use of R as an Advanced Calculator 506
A.3.2 Quitting R 508
A.3.3 Obtaining Help in R 508
A.3.4 Variable Names in R 508
A.3.5 Working with Vectors in R 509
A.3.6 Loading Data into R 511
A.3.7 Working with Data Frames in R 513
A.3.8 Using Functions in R 514
A.3.9 Basic Statistical Functions in R 515
A.3.10 Basic Plotting in R 516
A.3.11 Writing Functions in R 518
* A.3.12 Matrix Arithmetic in R 520
References 523

The GLMsData package 525
References 527

Selected Solutions 529
Solutions from Chap. 1 529
Solutions from Chap. 2 530
Solutions from Chap. 3 532
Solutions from Chap. 4 534
Solutions from Chap. 5 536
Solutions from Chap. 6 537
Solutions from Chap. 7 537
Solutions from Chap. 8 539

Solutions from Chap. 9 539
Solutions from Chap. 10 541
Solutions from Chap. 11 544
Solutions from Chap. 12 547
Solutions from Chap. 13 548
References 550

Index: Data sets 551

Index: R commands 553

Index: General topics 557