Contents

1 Preliminaries **15**

1.1 Investigation and Exploration Prior to Analysis 15

1.1.1 A 'simple' straight line regression example 16

1.1.2 Data that cry out for transformation 17

1.1.3 Classification of forensic glass 19

1.2 Data Analysis Aims and Purposes 20

1.2.1 Data-based Generalization 22

1.2.2 The source/target distinction 22

1.2.3 Data – some further considerations 25

1.3 Models and Model Assumptions 26

1.3.1 What criterion should be used to add a line? 26

1.3.2 Beyond least squares 29

1.4 Model Implications and Model Scrutiny 29

1.4.1 The iid assumption in action – simulation 30

1.4.2 Diagnostics – checks on the fitted model 32

1.4.3 Key questions for the use of models 39

1.5 Some Further Issues for Data Analysis 40

1.5.1 When are rough and ready methods enough? 40

1.5.2 Automation 40

1.5.3 The 'Learning from Data' Perspective 41

1.6 Weighting Effects in Summary Data 41

1.6.1 Simple examples 42

1.6.2 The UCB Admissions Data 42

1.6.3 Summary of continuous outcome data - an example 44

1.6.4 Biases from omission of features – further comments 46

1.7 Exercises 46

2 The Limits of Statistical Learning **49**

2.1 A clearly inappropriate use of least squares modeling 50

2.2 Variation between patterns of wage change 51

2.3 Variation within a site, and variation between sites 52

2.4 Record times for athletic races 53

2.5 Global average temperature – a time series 55

2.6 Analysis of a substantial dataset – US accident data 57

2.6.1 Alternative and better approach – a ratio of ratios 59

3 Data-Based Generalization **63**

3.1 Populations and Samples 63

3.2 Simulation from the Normal, versus Bootstrap Resampling 64
3.2.1 Samples from an empirical distribution 65
3.3 How accurate is the density curve or boxplot? 65
3.3.1 Samples from a theoretical normal population 65
3.3.2 How Accurate is the Density Curve – Bootstrap Resampling 67
3.4 Theoretical probability distributions 69
3.4.1 Normal and other probability plots 71
3.5 Sample Statistics and Sampling Distributions 71
3.5.1 Variance and Standard Deviation: 71
3.5.2 The Standard Error of the Mean (SEM): 72
3.5.3 The sampling distribution of the mean: 72
3.6 Accuracy Assessment 75
3.6.1 Mechanisms for assessing predictive accuracy 75
3.6.2 Methods for assessing accuracy of parameter estimates 78
3.6.3 Examples 78
3.6.4 Male cats – regression of head weight on body weight 81

4 Linear Models 85

4.1 Linear Models in R – Basic Ideas 85
4.1.1 Straight line regression – algebraic details 85
4.1.2 Syntax – model, graphics and table formulae: 86
4.1.3 The technicalities of linear models 87
4.1.4 A note on the least squares methodology 88
4.1.5 Two explanatory variables –the Northern Ireland hill race data . . . 88
4.2 Empirical Methods for Model Assessment 93
4.2.1 The sampling distribution of the regression estimates 95
4.3 Linear Models – Exploiting the Magic of Model Formulae 97
4.3.1 Linear models – general definition 97
4.3.2 Modeling qualitative effects 98
4.3.3 Interaction terms 101
4.4 Regression Examples 102
4.4.1 The `meuse` data set – heavy metal concentrations 102
4.4.2 1970 cost for US electricity producers 104
4.5 Variable Selection – Stepwise and Related Methods 110
4.5.1 Use of simulation to check out selection effects: 111
4.5.2 Variable selection – further points 112
4.5.3 Model selection more generally: 112

5 Generalized Linear Models (GLMs) – Logistic and Poisson Regression 113

5.1 An introduction to logistic regression 113
5.1.1 Logistic regression for the US accident data 115
5.2 Poisson regression 117
5.2.1 Numbers of moths, by habitat type & transect length 117
5.2.2 Not all count data is Poisson 120

6 Generalized Additive Models (GAMs) 121

6.1 Linear models, fitted using `lm()`, can be curvilinear . . . 121
6.1.1 Polynomials and orthogonal polynomials . . . 122
6.1.2 Regression splines versus smoothing or penalized splines . . . 122
6.2 Regression Spline Basis Terms . . . 123
6.2.1 Cubic regression splines . . . 123
6.2.2 Alternative choices of regression spline basis functions . . . 126
6.3 Generalized Additive Models . . . 126
6.3.1 Smoothing splines and penalized splines . . . 127
6.3.2 A GAM model with a single smoothing term . . . 127
6.3.3 Lead concentrations – joint effects of `elev` and `dist` . . . 131
6.3.4 Can the anova *p*-values be trusted? . . . 135
6.3.5 Fitting a GAM model to climate data . . . 137
6.3.6 Smoothing terms with time series data – issues of interpretation . . 138
6.4 Logistic regression with GAM smoothing term . . . 141
6.5 Poisson regression with GAM smoothing term . . . 143
6.6 Exercises . . . 144

7 Discrimination and Classification 147

7.1 Linear Methods for Discrimination . . . 147
7.1.1 `lda()` and `qda()` . . . 147
7.1.2 Canonical discriminant scores (`lda()` only), and associated plots . . 149
7.1.3 Linear Discriminant Analysis – practocal details . . . 149
7.1.4 Accuracy comparisons . . . 156
7.2 Tree-based Methods . . . 157
7.3 Random forests . . . 160
7.3.1 The virtues of random forests . . . 162
7.3.2 Prior probabilities . . . 162
7.3.3 Who is likely to buy an insurance? . . . 165
7.4 Issues for Comparing Different Methods . . . 166
7.4.1 Further methods – Support Vector Machines, and Boosting . . . 169
7.5 Multiple Levels of Variation – the `Vowel` Dataset . . . 170

8 Regression Modeling – Further Notes 175

8.1 Errors in *x* . . . 175
8.1.1 Measurement of dietary intake . . . 175
8.1.2 A simulation of the effect of measurement error . . . 176
8.1.3 Errors in variables – multiple regression . . . 177
8.2 Data collection biases . . . 178
8.2.1 Model and/or variable selection bias . . . 178
8.3 Further Examples and Commentary . . . 178
8.3.1 Do the left-handed die young . . . 178
8.3.2 Hormone replacement therapy . . . 178
8.3.3 Does screening reduce deaths from gastric cancer? . . . 179
8.3.4 Alcohol consumptions and risk of coronary heart disease . . . 180
8.3.5 Freakonomics . . . 182
8.3.6 Further reading . . . 182

9 Ordination 183
9.1 Distance measures 184
9.1.1 Euclidean distances 184
9.1.2 Non-Euclidean distance measures 184
9.2 From distances to a configuration in Euclidean space 185
9.3 Non-metric scaling 186
9.3.1 Examples 186

10 Text Mining Using the tm Package 193
10.1 The Ovid documents – from Text Files to Corpus 193
10.2 Text Mining of Spam Email 194
10.2.1 Download and untar data from the Spamassassin site 195
10.2.2 Creation of volatile corpora 196
10.2.3 Creation of a Document Term matrix 197
10.2.4 Weighting of terms 198
10.2.5 Discriminant approaches 199

11 Technical Mathematical Results 203
11.1 Probability distributions – Definition and Theory 203
11.1.1 Expectation Algebra – Means and Variances 204
11.2 Least Squares Estimates 205
11.2.1 The mean is a least squares estimator 205
11.2.2 Beyond Least Squares – Maximum Likelihood 205
11.3 The Likelihood Principle 206
11.4 Linear models – matrix derivations & extensions 207
11.5 Generalized Linear Models – theory & computation 208
11.5.1 Maximum likelihood parameter estimates 208
11.6 Linear and Quadratic Discriminant Analysis – Theory 209
11.7 From Distances to a Representation in Euclidean Space 212
11.7.1 An exact representation? 213

12 Graphical User Interfaces – the R Commander and Rattle 215
12.1 The R Commander GUI – A Guided to Getting Started 215
12.2 The Rattle GUI 216

Bibliography 219

13 Color Versions of Selected Graphs 225

A Warning Note:

A big computer, a complex algorithm and a long time does not equal science.
[Robert Gentleman, SSC 2003, Halifax, June 2003]