Contents

Preface **ix**

1 Introduction **1**
1.1 Before You Start 1
1.2 Initial Data Analysis 2
1.3 When to Use Regression Analysis 6
1.4 History 7

2 Estimation **11**
2.1 Linear Model 11
2.2 Matrix Representation 12
2.3 Estimating β 12
2.4 Least Squares Estimation 13
2.5 Examples of Calculating $\hat{\beta}$ 14
2.6 Gauss–Markov Theorem 15
2.7 Goodness of Fit 16
2.8 Example 18
2.9 Identifiability 21

3 Inference **25**
3.1 Hypothesis Tests to Compare Models 25
3.2 Testing Examples 27
3.3 Permutation Tests 32
3.4 Confidence Intervals for β 34
3.5 Confidence Intervals for Predictions 36
3.6 Designed Experiments 39
3.7 Observational Data 43
3.8 Practical Difficulties 47

4 Diagnostics **53**
4.1 Checking Error Assumptions 53
4.2 Finding Unusual Observations 64
4.3 Checking the Structure of the Model 72

5 Problems with the Predictors **77**
5.1 Errors in the Predictors 77

5.2 Changes of Scale 81
5.3 Collinearity 83

6 Problems with the Error 89
6.1 Generalized Least Squares 89
6.2 Weighted Least Squares 92
6.3 Testing for Lack of Fit 94
6.4 Robust Regression 98

7 Transformation 109
7.1 Transforming the Response 109
7.2 Transforming the Predictors 112

8 Variable Selection 121
8.1 Hierarchical Models 121
8.2 Testing-Based Procedures 122
8.3 Criterion-Based Procedures 125
8.4 Summary 130

9 Shrinkage Methods 133
9.1 Principal Components 133
9.2 Partial Least Squares 140
9.3 Ridge Regression 143

10 Statistical Strategy and Model Uncertainty 147
10.1 Strategy 147
10.2 An Experiment in Model Building 148
10.3 Discussion 149

11 Insurance Redlining — A Complete Example 151
11.1 Ecological Correlation 151
11.2 Initial Data Analysis 153
11.3 Initial Model and Diagnostics 156
11.4 Transformation and Variable Selection 158
11.5 Discussion 161

12 Missing Data 163

13 Analysis of Covariance 167
13.1 A Two-Level Example 168
13.2 Coding Qualitative Predictors 172
13.3 A Multilevel Factor Example 174

14 One-Way Analysis of Variance 181
14.1 The Model 181

14.2 An Example 182
14.3 Diagnostics 185
14.4 Pairwise Comparisons 186

15 Factorial Designs 189
15.1 Two-Way ANOVA 189
15.2 Two-Way ANOVA with One Observation per Cell 190
15.3 Two-Way ANOVA with More than One Observation per Cell 193
15.4 Larger Factorial Experiments 197

16 Block Designs 203
16.1 Randomized Block Design 204
16.2 Latin Squares 208
16.3 Balanced Incomplete Block Design 212

A R Installation, Functions and Data 217

B Quick Introduction to R 219
B.1 Reading the Data In 219
B.2 Numerical Summaries 219
B.3 Graphical Summaries 220
B.4 Selecting Subsets of the Data 221
B.5 Learning More about R 222

Bibliography 223

Index 227

Preface

There are many books on regression and analysis of variance. These books expect different levels of preparedness and place different emphases on the material. This book is not introductory. It presumes some knowledge of basic statistical theory and practice. Readers are expected to know the essentials of statistical inference such as estimation, hypothesis testing and confidence intervals. A basic knowledge of data analysis is presumed. Some linear algebra and calculus are also required.

The emphasis of this text is on the practice of regression and analysis of variance. The objective is to learn what methods are available and more importantly, when they should be applied. Many examples are presented to clarify the use of the techniques and to demonstrate what conclusions can be made. There is relatively less emphasis on mathematical theory, partly because some prior knowledge is assumed and partly because the issues are better tackled elsewhere. Theory is important because it guides the approach we take. I take a wider view of statistical theory. It is not just the formal theorems. Qualitative statistical concepts are just as important in statistics because these enable us to actually do it rather than just talk about it. These qualitative principles are harder to learn because they are difficult to state precisely but they guide the successful experienced statistician.

Data analysis cannot be learned without actually doing it. This means using a statistical computing package. There is a wide choice of such packages. They are designed for different audiences and have different strengths and weaknesses. I have chosen to use R (Ref. Ihaka and Gentleman (1996) and R Development Core Team (2003)). Why have I used R? There are several reasons.

1. *Versatility*. R is also a programming language, so I am not limited by the procedures that are preprogrammed by a package. It is relatively easy to program new methods in R.
2. *Interactivity*. Data analysis is inherently interactive. Some older statistical packages were designed when computing was more expensive and batch processing of computations was the norm. Despite improvements in hardware, the old batch processing paradigm lives on in their use. R does one thing at a time, allowing us to make changes on the basis of what we see during the analysis.
3. *Freedom*. R is based on S from which the commercial package S-plus is derived. R itself is open-source software and may be obtained free of charge to all. Linux, Macintosh, Windows and other UNIX versions are maintained and can be obtained from the R-project at `www.r-project.org`. R is mostly compatible with S-plus, meaning that S-plus could easily be used for most of the examples provided in this book.

4. *Popularity*. SAS is the most common statistics package in general use but R or S is most popular with researchers in statistics. A look at common statistical journals confirms this popularity. R is also popular for quantitative applications in finance.

Getting Started with R

R requires some effort to learn. Such effort will be repaid with increased productivity. You can learn how to obtain R in Appendix A along with instructions on the installation of additional software and data used in this book.

This book is not an introduction to R. Appendix B provides a brief introduction to the language, but alone is insufficient. I have intentionally included in the text all the commands used to produce the output seen in this book. This means that you can reproduce these analyses and experiment with changes and variations before fully understanding R. You may choose to start working through this text before learning R and pick it up as you go. Free introductory guides to R may be obtained from the R project Web site at `www.r-project.org`. Introductory books have been written by Dalgaard (2002) and Maindonald and Braun (2003). Venables and Ripley (2002) also have an introduction to R along with more advanced material. Fox (2002) is intended as a companion to a standard regression text. You may also find Becker, Chambers, and Wilks (1998) and Chambers and Hastie (1991) to be useful references to the S language. Ripley and Venables (2000) wrote a more advanced text on programming in S or R.

The Web site for this book is at `www.stat.lsa.umich.edu/~faraway/LMR` where data described in this book appear. Updates and errata will appear there also.

Thanks to the builders of R without whom this book would not have been possible.