Practical Regression and Anova using R

Julian J. Faraway

July 2002

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. Students are expected to know the essentials of statistical inference like estimation, hypothesis testing and confidence intervals. A basic knowledge of data analysis is presumed. Some linear algebra and calculus is 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 learnt 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)). Why do I use R ? The are several reasons.

1. Versatility. R is a 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. R is based on S from which the commercial package S-plus is derived. R itself is open-source software and may be freely redistributed. 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 the examples given in this book.

4. Popularity. SAS is the most common statistics package in general 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.

The greatest disadvantage of R is that it is not so easy to learn. Some investment of effort is required before productivity gains will be realized. This book is not an introduction to R . There is a short introduction

in the Appendix but readers are referred to the R-project web site at `www.r-project.org` where you can find introductory documentation and information about books on R . 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 . The reader may choose to start working through this text before learning R and pick it up as you go.

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

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

Contents