

Practical Regression and Anova using R

Julian J. Faraway

July 2002

Copyright ©1999, 2000, 2002 Julian J. Faraway

Permission to reproduce individual copies of this book for personal use is granted. Multiple copies may be created for nonprofit academic purposes — a nominal charge to cover the expense of reproduction may be made. Reproduction for profit is prohibited without permission.

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 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

1	Introduction	8
1.1	Before you start	8
1.1.1	Formulation	8
1.1.2	Data Collection	9
1.1.3	Initial Data Analysis	9
1.2	When to use Regression Analysis	13
1.3	History	14
2	Estimation	16
2.1	Example	16
2.2	Linear Model	16
2.3	Matrix Representation	17
2.4	Estimating β	17
2.5	Least squares estimation	18
2.6	Examples of calculating $\hat{\beta}$	19
2.7	Why is $\hat{\beta}$ a good estimate?	19
2.8	Gauss-Markov Theorem	20
2.9	Mean and Variance of $\hat{\beta}$	21
2.10	Estimating σ^2	21
2.11	Goodness of Fit	21
2.12	Example	23
3	Inference	26
3.1	Hypothesis tests to compare models	26
3.2	Some Examples	28
3.2.1	Test of all predictors	28
3.2.2	Testing just one predictor	30
3.2.3	Testing a pair of predictors	31
3.2.4	Testing a subspace	32
3.3	Concerns about Hypothesis Testing	33
3.4	Confidence Intervals for β	36
3.5	Confidence intervals for predictions	39
3.6	Orthogonality	41
3.7	Identifiability	44
3.8	Summary	46
3.9	What can go wrong?	46
3.9.1	Source and quality of the data	46

3.9.2	Error component	47
3.9.3	Structural Component	47
3.10	Interpreting Parameter Estimates	48
4	Errors in Predictors	55
5	Generalized Least Squares	59
5.1	The general case	59
5.2	Weighted Least Squares	62
5.3	Iteratively Reweighted Least Squares	64
6	Testing for Lack of Fit	65
6.1	σ^2 known	66
6.2	σ^2 unknown	67
7	Diagnostics	72
7.1	Residuals and Leverage	72
7.2	Studentized Residuals	74
7.3	An outlier test	75
7.4	Influential Observations	78
7.5	Residual Plots	80
7.6	Non-Constant Variance	83
7.7	Non-Linearity	85
7.8	Assessing Normality	88
7.9	Half-normal plots	91
7.10	Correlated Errors	92
8	Transformation	95
8.1	Transforming the response	95
8.2	Transforming the predictors	98
8.2.1	Broken Stick Regression	98
8.2.2	Polynomials	100
8.3	Regression Splines	102
8.4	Modern Methods	104
9	Scale Changes, Principal Components and Collinearity	106
9.1	Changes of Scale	106
9.2	Principal Components	107
9.3	Partial Least Squares	113
9.4	Collinearity	117
9.5	Ridge Regression	120
10	Variable Selection	124
10.1	Hierarchical Models	124
10.2	Stepwise Procedures	125
10.2.1	Forward Selection	125
10.2.2	Stepwise Regression	126
10.3	Criterion-based procedures	128

10.4 Summary	133
11 Statistical Strategy and Model Uncertainty	134
11.1 Strategy	134
11.2 Experiment	135
11.3 Discussion	136
12 Chicago Insurance Redlining - a complete example	138
13 Robust and Resistant Regression	150
14 Missing Data	156
15 Analysis of Covariance	160
15.1 A two-level example	161
15.2 Coding qualitative predictors	164
15.3 A Three-level example	165
16 ANOVA	168
16.1 One-Way Anova	168
16.1.1 The model	168
16.1.2 Estimation and testing	168
16.1.3 An example	169
16.1.4 Diagnostics	171
16.1.5 Multiple Comparisons	172
16.1.6 Contrasts	177
16.1.7 Scheffé's theorem for multiple comparisons	177
16.1.8 Testing for homogeneity of variance	179
16.2 Two-Way Anova	179
16.2.1 One observation per cell	180
16.2.2 More than one observation per cell	180
16.2.3 Interpreting the interaction effect	180
16.2.4 Replication	184
16.3 Blocking designs	185
16.3.1 Randomized Block design	185
16.3.2 Relative advantage of RCBD over CRD	190
16.4 Latin Squares	191
16.5 Balanced Incomplete Block design	195
16.6 Factorial experiments	200
A Recommended Books	204
A.1 Books on R	204
A.2 Books on Regression and Anova	204
B R functions and data	205

C Quick introduction to R	207
C.1 Reading the data in	207
C.2 Numerical Summaries	207
C.3 Graphical Summaries	209
C.4 Selecting subsets of the data	209
C.5 Learning more about R	210