
Preface

Linear models are central to the practice of statistics. They are part of the core knowledge expected of any applied statistician. Linear models are the foundation of a broad range of statistical methodologies; this book is a survey of techniques that grow from a linear model. Our starting point is the regression model with response y and predictors x_1, \dots, x_p . The model takes the form:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

where ϵ is normally distributed. This book presents three extensions to this framework. The first generalizes the y part; the second, the ϵ part; and the third, the x part of the linear model.

Generalized Linear Models: The standard linear model cannot handle nonnormal responses, y , such as counts or proportions. This motivates the development of generalized linear models that can represent categorical, binary and other response types.

Mixed Effect Models: Some data has a grouped, nested or hierarchical structure. Repeated measures, longitudinal and multilevel data consist of several observations taken on the same individual or group. This induces a correlation structure in the error, ϵ . Mixed effect models allow the modeling of such data.

Nonparametric Regression Models: In the linear model, the predictors, x , are combined in a linear way to model the effect on the response. Sometimes this linearity is insufficient to capture the structure of the data and more flexibility is required. Methods such as additive models, trees and neural networks allow a more flexible regression modeling of the response that combine the predictors in a nonparametric manner.

This book aims to provide the reader with a well-stocked toolbox of statistical methodologies. A practicing statistician needs to be aware of and familiar with the basic use of a broad range of ideas and techniques. This book will be a success if the reader is able to recognize and get started on a wide range of problems. However, the breadth comes at the expense of some depth. Fortunately, there are book-length treatments of topics discussed in every chapter of this book, so the reader will know where to go next if needed.

R is a free software environment for statistical computing and graphics. It runs on a wide variety of platforms including the Windows, Linux and Macintosh operating systems. Although there are several excellent statistical packages, only R is both free and possesses the power to perform the analyses demonstrated in this book. While it is possible in principle to learn statistical methods from purely theoretical expositions, I believe most readers learn best from the demonstrated interplay of

theory and practice. The data analysis of real examples is woven into this book and all the R commands necessary to reproduce the analyses are provided.

Prerequisites: Readers should possess some knowledge of linear models. The first chapter provides a review of these models. This book can be viewed as a sequel to *Linear Models with R*, Faraway (2004). Even so there are plenty of other good books on linear models such as Draper and Smith (1998) or Weisberg (2005), that would provide ample grounding. Some knowledge of likelihood theory is also very useful. An outline is provided in Appendix A, but this may be insufficient for those who have never seen it before. A general knowledge of statistical theory is also expected concerning such topics as hypothesis tests or confidence intervals. Even so, the emphasis in this text is on application, so readers without much statistical theory can still learn something here.

This is not a book about learning R, but the reader will inevitably pick up the language by reading through the example data analyses. Readers completely new to R will benefit from studying an introductory book such as Dalgaard (2002) or one of the many tutorials available for free at the R website. Even so, the book should be intelligible to a reader without prior knowledge of R just by reading the text and output. R skills can be further developed by modifying the examples in this book, trying the exercises and studying the help pages for each command as needed. There is a large amount of detailed help on the commands available within the software and there is no point in duplicating that here. Please refer to Appendix B for details on obtaining and installing R along with the necessary add-on packages and data necessary for running the examples in this text. S-plus derives from the same S language as R, so many of the commands in this book will work. However, there are some differences in the syntax and the availability of add-on packages, so not everything here will work in S-plus.

The website for this book is at www.stat.lsa.umich.edu/~faraway/ELM where data described in this book appears. Updates and errata will also appear there.

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

Contents

Preface	v
1 Introduction	1
2 Binomial Data	25
2.1 Challenger Disaster Example	25
2.2 Binomial Regression Model	26
2.3 Inference	29
2.4 Tolerance Distribution	31
2.5 Interpreting Odds	31
2.6 Prospective and Retrospective Sampling	34
2.7 Choice of Link Function	36
2.8 Estimation Problems	38
2.9 Goodness of Fit	40
2.10 Prediction and Effective Doses	41
2.11 Overdispersion	43
2.12 Matched Case-Control Studies	48
3 Count Regression	55
3.1 Poisson Regression	55
3.2 Rate Models	61
3.3 Negative Binomial	63
4 Contingency Tables	69
4.1 Two-by-Two Tables	69
4.2 Larger Two-Way Tables	75
4.3 Matched Pairs	79
4.4 Three-Way Contingency Tables	81
4.5 Ordinal Variables	88
5 Multinomial Data	97
5.1 Multinomial Logit Model	97
5.2 Hierarchical or Nested Responses	103
5.3 Ordinal Multinomial Responses	106

6 Generalized Linear Models	115
6.1 GLM Definition	115
6.2 Fitting a GLM	117
6.3 Hypothesis Tests	120
6.4 GLM Diagnostics	123
7 Other GLMs	135
7.1 Gamma GLM	135
7.2 Inverse Gaussian GLM	142
7.3 Joint Modeling of the Mean and Dispersion	144
7.4 Quasi-Likelihood	147
8 Random Effects	153
8.1 Estimation	154
8.2 Inference	158
8.3 Predicting Random Effects	161
8.4 Blocks as Random Effects	163
8.5 Split Plots	167
8.6 Nested Effects	170
8.7 Crossed Effects	172
8.8 Multilevel Models	174
9 Repeated Measures and Longitudinal Data	185
9.1 Longitudinal Data	186
9.2 Repeated Measures	191
9.3 Multiple Response Multilevel Models	195
10 Mixed Effect Models for Nonnormal Responses	201
10.1 Generalized Linear Mixed Models	201
10.2 Generalized Estimating Equations	204
11 Nonparametric Regression	211
11.1 Kernel Estimators	213
11.2 Splines	217
11.3 Local Polynomials	221
11.4 Wavelets	222
11.5 Other Methods	226
11.6 Comparison of Methods	227
11.7 Multivariate Predictors	228
12 Additive Models	231
12.1 Additive Models Using the <code>gam</code> Package	233
12.2 Additive Models Using <code>mgcv</code>	235
12.3 Generalized Additive Models	240
12.4 Alternating Conditional Expectations	241

CONTENTS	ix
12.5 Additivity and Variance Stabilization	244
12.6 Generalized Additive Mixed Models	246
12.7 Multivariate Adaptive Regression Splines	247
13 Trees	253
13.1 Regression Trees	253
13.2 Tree Pruning	257
13.3 Classification Trees	261
14 Neural Networks	269
14.1 Statistical Models as NNs	270
14.2 Feed-Forward Neural Network with One Hidden Layer	270
14.3 NN Application	272
14.4 Conclusion	276
A Likelihood Theory	279
A.1 Maximum Likelihood	279
A.2 Hypothesis Testing	282
B R Information	287
Bibliography	289
Index	297