

PREFACE

Linear regression analysis consists of a collection of techniques used to explore relationships between variables. It is interesting both theoretically because of the elegance of the underlying theory, and from an applied point of view, because of the wide variety of uses of regression that have appeared, and continue to appear every day. In this book, regression methods, used to fit models for a dependent variable as a function of one or more independent variables, are discussed for the reader who wants to learn to apply them to data. The central themes are building models, assessing fit and reliability, and drawing conclusions. If used as a textbook, it is intended as a second or third course in statistics. The only definite prerequisites are familiarity with the ideas of significance tests, p -values, confidence intervals, random variables, estimation of parameters, and also with the normal distribution, and distributions derived from it, such as Student's t , and the F , and χ^2 . Of course, additional knowledge of statistical methods or linear algebra will be of value.

The book is divided into 11 chapters. Chapters 1 and 2 provide fairly standard results for least squares estimation in simple and multiple regression, respectively. The third chapter is called "Drawing Conclusions" and is about interpreting the results of the methods from the first two chapters. Also, a discussion of the effects of independent variables that are imperfectly measured is given. Chapter 4 presents additional results on least squares estimation. Chapters 5 and 6 cover methods for studying the lack of fit of a model, checking for failures of assumptions, and assessing the reliability of a fitted model. In Chapter 5, theoretical results for the necessary statistics are given, since these will be unfamiliar to many readers, while Chapter 6 covers graphical and other procedures based on these statistics, as well as possible remedies for the problems they uncover. In Chapter 7, the topics covered are relevant to problems of model building, including dummy variables, polynomial regression, and principal components. Then, Chapter 8 provides methods for selecting a model based on a subset of variables. In Chapter 9, special considerations when regression methods are to be used to make predictions are discussed. In

viii Preface

each of these chapters, the methods discussed are illustrated by examples using real data.

The last two chapters are shorter than the earlier ones. Chapter 10 gives guidelines for analysis of partially observed or incomplete data. Finally, in Chapter 11, alternatives to least squares estimates are discussed.

Several of the chapters have associated appendixes that have been collected at the end of the text, but are numbered to correspond to the chapters. For example: Appendix 1A.2 is the second appendix for Chapter 1. The chapters are ordered for a semester or quarter course on linear regression, and Chapters 1 to 8 make up a rigorous one-quarter course.

Homework problems are provided for each of the first nine chapters. The theoretical problems are intended only for students with the necessary statistical background. Problems that require analysis of data are intended for everyone. Some of these have been left vague in their requirements, so that they can be varied according to the interests of the students. Most of the problems use real data and can be approached in many ways.

Computers. The growth of the use of regression methods can be traced directly to wider availability of computers. While this book is not intended as a manual for any specific computer program, it is oriented for the reader who expects to use computers to apply the techniques learned. High quality software for regression calculations is available, and references to the necessary sources are in the text, in the homework problems, and in the appendixes.

Acknowledgments. I am grateful to the many people who have commented on early drafts of the book, supplied examples, or through discussion have clarified my own thoughts on the topics covered. Included in this group are Christopher Bingham, Morton Brown, Cathy Campbell, Dennis Cook, Stephen Fienberg, James Frane, Seymour Geisser, John Hartigan, David Hinkley, Alan Izenman, Soren Johansen, Kenneth Koehler, David Lane, Kinley Larntz, John Rice, Donald Rubin, Wei-Chung Shih, G. W. Stewart, Douglas Tiffany, Carol Weisberg, Howard Weisberg, and an anonymous reader. Also, I wish to thank the production staff at the University of Minnesota, Naomi Miner, Sue Hangege, Therese Therrien, and especially Marianne O'Brien, whose expert assistance made completion of this work a reality.

During the writing of this book, I have benefited from partial support from a grant from the U.S. National Institute of General Medical Sciences. Additional support for computations has been provided by the University Computer Center, University of Minnesota.

SANFORD WEISBERG