Springer Texts in Statistics

Series Editors:

G. Casella
S. Fienberg
I. Olkin

Springer Texts in Statistics

For other titles published in this series, go to
www.springer.com/series/417

Simon J. Sheather

A Modern Approach to Regression with R

Springer

Simon J. Sheather
Department of Statistics
Texas A&M University
College Station, TX, USA

ISBN: 978-0-387-09607-0 e-ISBN: 978-0-387-09608-7
DOI: 10.1007/978-0-387-09608-7

Library of Congress Control Number: 2008940909

Printed on acid-free paper

springer.com

Dedicated to
My mother, Margaret,
and my wife, Filomena

Preface

This book focuses on tools and techniques for building regression models using real-world data and assessing their validity. A key theme throughout the book is that ***it makes sense to base inferences or conclusions only on valid models***.

Plots are shown to be an important tool for both building regression models and assessing their validity. We shall see that deciding what to plot and how each plot should be interpreted will be a major challenge. In order to overcome this challenge we shall need to understand the mathematical properties of the fitted regression models and associated diagnostic procedures. As such this will be an area of focus throughout the book. In particular, we shall carefully study the properties of residuals in order to understand when patterns in residual plots provide direct information about model misspecification and when they do not.

The regression output and plots that appear throughout the book have been generated using R. The output from R that appears in this book has been edited in minor ways. On the book web site you will find the R code used in each example in the text. You will also find SAS-code and Stata-code to produce the equivalent output on the book web site. Primers containing expanded explanation of R, SAS and Stata and their use in this book are also available on the book web site. Purpose-built functions have been written in SAS and Stata to cover some of the regression procedures discussed in this book. Examples include a multivariate version of the Box-Cox transformation method, inverse response plots and marginal model plots.

The book contains a number of new real data sets from applications ranging from rating restaurants, rating wines, predicting newspaper circulation and magazine revenue, comparing the performance of NFL kickers and comparing finalists in the Miss America pageant across states. In addition, a number of real data sets that have appeared in other books are also considered. The practice of considering contemporary real data sets was begun based on questions from students about how regression can be used in real life. One of the aspects of the book that sets it apart from many other regression books is that complete details are provided for each example. This completeness helps students better understand how regression is used in practice to build different models and assess their validity.

Included in the Exercises are two different types of problems involving data. In the first, a situation is described and it is up to the students to develop a valid regression model. In the second type of problem a situation is described and then output

from one or models is provided and students are asked to comment and provide conclusions. This has been a conscious choice as I have found that both types of problems enhance student learning.

Chapters 2, 3 and 4 look at the case when there is a single predictor. This again has been a conscious choice as it enables students to look at many aspects of regression in the simplest possible setting. Chapters 5, 6, 7 and 9 focus on regression models with multiple predictors. In Chapter 8 we consider logistic regression. Chapter 9 considers regression models with correlated errors. Finally, Chapter 10 provides an introduction to random effects and mixed models.

Throughout the book specific suggestions are given on how to proceed when performing a regression analysis. Flow charts providing step-by-step instructions are provided first for regression problems involving a single predictor and later for multiple regression problems. The flow charts were first produced in response to requests from students when this material was first taught. They have been used with great success ever since.

Chapter 1 contains a discussion of four real examples. The first example highlights a key message of the book, namely, it is only sensible to base decisions of inferences on a valid regression model. The other three examples provide an indication of the practical problems one can solve using the regression methods discussed in the book.

In Chapter 2 we consider problems involving modeling the relationship between two variables. Throughout this chapter we assume that the model under consideration is a valid model (i.e., correctly specified.)

In Chapter 3 we will see that when we use a regression model we implicitly make a series of assumptions. We then consider a series of tools known as regression diagnostics to check each assumption. Having used these tools to diagnose potential problems with the assumptions, we look at how to first identify and then overcome or deal with problems with assumptions due to nonconstant variance or nonlinearity. A primary aim of Chapter 3 is to understand what actually happens when the standard assumptions associated with a regression model are violated, and what should be done in response to each violation.

In Chapter 3, we show that it is sometimes possible to overcome nonconstant error variance by transforming the response and/or the predictor variables. In Chapter 4 we consider an alternative way of coping with nonconstant error variance, namely weighted least squares.

Chapter 5 considers multiple linear regression problems involving modeling the relationship between a dependent variable and two or more predictor variables. Throughout Chapter 5, we assume that the multiple linear regression model under consideration is a valid model for the data. Chapter 6 considers regression diagnostics to check each of these assumptions associated with having a valid multiple regression model.

In Chapter 7 we consider methods for choosing the "best" model from a class of multiple regression models, using what are called variable selection methods. We discuss the consequences of variable selection on subsequent inferential procedures, (i.e., tests and confidence intervals).

Chapter 8 considers the situation in which the response variable follows a binomial distribution rather than a continuous distribution. We show that an appropriate model in this circumstance is a logistic regression model. We consider both inferential and diagnostic procedures for logistic regression models.

In many situations data are collected over time. It is common for such data sets to exhibit serial correlation, that is, results from the current time period are correlated with results from earlier time periods. Thus, these data sets violate the assumption that the errors are independent, an important assumption necessary for the validity of least squares based regression methods. Chapter 9 considers regression models when the errors are correlated over time. Importantly, we show how to re-specify a regression model with correlated errors as a different but equivalent regression model with uncorrelated errors. We shall discover that this allows us to use the diagnostic methods discussed in earlier chapters on problems with correlated errors.

Chapter 10 contains an introduction to random effects and mixed models. We again stress the use of re-specifying such models to obtain equivalent models with uncorrelated errors.

Finally, the Appendix discusses two nonparametric smoothing techniques, namely, kernel density estimation and nonparametric regression for a single predictor.

The book is aimed at first-year graduate students in statistics. It could also be used for a senior undergraduate class. The text grew out of a set of class notes, used for both a graduate and a senior undergraduate semester-long regression course at Texas A&M University. I am grateful to the students who took these courses. I would like to make special mention of Brad Barney, Dana Bergstresser, Charles Lindsey, Andrew Redd and Elizabeth Young. Charles Lindsey wrote the Stata code that appears in the Stata primer that accompanies the book. Elizabeth Young, along with Brad Barney and Charles Lindsey, wrote the SAS code that appears in the SAS primer that accompanies the book. Brad Barney kindly provided the analyses of the NFL kicker data in Chapter 1. Brad Barney and Andrew Redd contributed some of the R code used in the book.

Readers of this book will find that the work of Cook and Weisberg has had a profound influence on my thinking about regression. In particular, this book contains many references to the books by Cook and Weisberg (1999b) and Weisberg (2005).

The content of the book has also been influenced by a number of people. Robert Kohn and Geoff Eagleson, my colleagues for more than 10 years at the University of New South Wales, taught me a lot about regression but more importantly about the importance of thoroughness when it comes to scholarship. My long-time collaborators on nonparametric statistics, Tom Hettmansperger and Joe McKean have helped me enormously both professionally and personally for more than 20 years. Lively discussions with Mike Speed about valid models and residual plots lead to dramatic changes to the examples and the discussion of this subject in Chapter 6. Mike Longnecker, kindly acted as my teaching mentor when I joined Texas A&M University in 2005. A number of reviewers provided valuable comments and

suggestions. I would like to especially acknowledge Larry Wasserman, Bruce Brown and Fred Lombard in this regard. Finally, I am grateful to Jennifer South who painstakingly proofread the whole manuscript.

The web site that accompanies the book contains R, SAS and Stata code and primers, along with all the data sets from the book can be found at www.stat.tamu.edu/~sheather/book. Also available at the book web site are online tutorials on matrices, R and SAS.

College Station, Texas Simon Sheather
October 2008

Contents