Applied Linear Models with SAS

This textbook for a second course in basic statistics for undergraduates or first-year graduate students introduces linear regression models and describes other linear models including Poisson regression, logistic regression, proportional hazards regression, and nonparametric regression. Numerous examples drawn from the news and current events with an emphasis on health issues illustrate these concepts.

Assuming only a pre-calculus background, the author keeps equations to a minimum and demonstrates all computations using SAS. Most of the programs and output are displayed in a self-contained way, with an emphasis on the interpretation of the output in terms of how it relates to the motivating example. Plenty of exercises conclude every chapter. All of the datasets and SAS programs are available from the book's Web site, along with other ancillary material.

Dr. Daniel Zelterman is Professor of Epidemiology and Public Health in the Division of Biostatistics at Yale University. His application areas include work in genetics, HIV, and cancer. Before moving to Yale in 1995, he was on the faculty of the University of Minnesota and at the State University of New York at Albany. He is an elected Fellow of the American Statistical Association. He serves as associate editor of *Biometrics* and other statistical journals. He is the author of *Models for Discrete Data* (1999), *Advanced Log-Linear Models Using SAS* (2002), *Discrete Distributions: Application in the Health Sciences* (2004), and *Models for Discrete Data: 2nd Edition* (2006). In his spare time he plays the bassoon in orchestral groups and has backpacked hundreds of miles of the Appalachian Trail.

Applied Linear Models with SAS

Daniel Zelterman
Yale University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521761598

First published 2010

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Zelterman, Daniel.
Applied linear models with SAS / Daniel Zelterman.
p. cm.
Includes index.
ISBN 978-0-521-76159-8 (hardback)
1. Linear regression. 2. Linear models (Statistics) 3. SAS (Computer program language) I. Title.
QA278.2.Z45 2010
519.5′35–dc22 2009053487

ISBN 978-0-521-76159-8 Hardback

Copyright page continues after page xiii.

Contents

Preface

Linear models are a powerful and useful set of methods in a large number of settings. Very briefly, there is some outcome measurement that is very important to us and we want to explain variations in its values in terms of other measurements in the data. The heights of several trees can be explained in terms of the trees' ages, for example. It is not a straight line relationship, of course, but knowledge of a tree's age offers us a large amount of explanatory value. We might also want to take into account the effects of measurements on the amount of light, water, nutrients, and weather conditions experienced by each tree. Some of these measurements will have greater explanatory value than others and we may want to quantify the relative usefulness of these different measures. Even after we are given all of this information, some trees will appear to thrive and others will remain stunted, when all are subjected to identical conditions. This variability is the whole reason for statistics existing as a scientific discipline. We usually try to avoid the use of the word "prediction" because this assumes that there is a cause-and-effect relationship. A tree's age does not directly cause it to grow, for example, but rather, a cumulative process associated with many environmental factors results in increasing height and continued survival. The best estimate we can make is a statement about the behavior of the average tree under identical conditions.

Many of my students go on to work in the pharmaceutical or health-care industry after graduating with a masters degree. Consequently, the choice of examples has a decidedly health/medical bias. We expect our students to be useful to their employers the day they leave our program so there is not a lot of time to spend on advanced theory that is not directly applicable. Not all of the examples are from the health sciences. Diverse examples such as the number of lottery winners and temperatures in various US cities are part of our common knowledge. Such examples do not need a lengthy explanation for the reader to appreciate many of the aspects of the data being presented.

How is this book different? The mathematical content and notation are kept to an absolute minimum. To paraphrase the noted physicist Steven Hawking, who

has written extensively for the popular audience, every equation loses half of your audience. There is really no need for formulas and their derivations in a book of this type if we rely on the computer to calculate quantities of interest. Long gone are the days of doing statistics with calculators or on the back of an envelope. Students of mathematical statistics should be able to provide the derivations of the formulas but they represent a very different audience. All of the important formulas are programmed in software so there is no need for the general user to know these.

The three important skills needed by a well-educated student of applied statistics are

1. Recognize the appropriate method needed in a given setting.
2. Have the necessary computer skills to perform the analysis.
3. Be able to interpret the output and draw conclusions in terms of the original data.

This book gives examples to introduce the reader to a variety of commonly encountered settings and provides guidance through these to complete these three goals. Not all possible situations can be described, of course, but the chosen settings include a broad survey of the type of problems the student of applied statistics is likely to run into.

What do I ask of my readers? We still need to use a lot of mathematical concepts such as the connection between a linear equation and drawing the line on $X - Y$ coordinates. There will be algebra and special functions such as square roots and logarithms. Logarithms, while we are on the subject, are always to the base e (=2.718) and not base 10.

We will also need a nodding acquaintance with the concepts of calculus. Many of us may have taken calculus in college, a long time ago, and not had much need to use it in the years since then. Perhaps we intentionally chose a course of study that avoided abstract mathematics. Even so, calculus represents an important and useful tool. The definition of the derivative of a function (What does this new function represent?) and integral (What does *this* new function represent?) are needed although we will never need to actually find a derivative or an integral. The necessary refresher to these important concepts is given in Section 1.4.

Also helpful is a previous course in statistics. The reader should be familiar with the mean and standard deviation, normal and binomial distributions, and hypothesis tests in general and the chi-squared and t-tests specifically. These important concepts are reviewed in Chapter 2 but an appreciation of these important ideas is almost a full course in itself. There is a large reliance on p-values in scientific research so it is important to know exactly what these represent.

There are a number of excellent general-purpose statistical packages available. We have chosen to illustrate our examples using SAS because of its wide acceptance and use in many industries but especially health care and pharmaceutical. Most of the examples given here are small, to emphasize interpretation and encourage practice. These datasets could be examined by most software packages. SAS, however, is

capable of handling huge datasets so the skills learned here can easily be used if and when much larger projects are encountered later.

The reader should already have some familiarity with running SAS on a computer. This would include using the editor to change the program, submitting the program, and retrieving and then printing the output. There are also popular point-and-click approaches to data analysis. While these are quick and acceptable, their ease of use comes with the price of not always being able to repeat the analysis because of the lack of a printed record of the steps that were taken. Data analysis, then, should be reproducible.

We will review some of the basics of SAS but a little hand-holding will prevent some of the agonizing frustrations that can occur when first starting out. Running the computer and, more generally, doing the exercises in this book are a very necessary part of learning statistics. Just as you cannot learn to play the piano simply by reading a book, statistical expertise, and the accompanying computer skills, can only be obtained through hours of active participation in the relevant act. Again, much like the piano, the instrument is not damaged by playing a wrong note. Nobody will laugh at you if you try something truly outlandish on the computer either. Perhaps something better will come of a new look at a familiar setting. Similarly, the reader is encouraged to look at the data and try a variety of different ways of looking, plotting, modeling, transforming, and manipulating. Unlike a mathematical problem with only one correct solution (contrary to many of our preconceived notions), there is often a lot of flexibility in the way statistics can be applied to summarize a set of data. As with yet another analogy to music, there are many ways to play the same song.

Acknowledgments

Thanks to the many students and teaching assistants who have provided useful comments and suggestions to the exposition as well as the computer assignments. Also to Chang Yu, Steven Schwager, and Amelia Dziengeleski for their careful readings of early drafts of the manuscript. Lauren Cowles and her staff at Cambridge University Press provided innumerable improvements and links to useful Web sites.

The DASL (pronounced "dazzle") StatLib library maintained at Carnegie Mellon University is a great resource and provided data for many examples and exercises contained here. Ed Tufte's books on graphics have taught me to look at data more carefully. His books are highly recommended.

I am grateful to *The New York Times* for their permission to use many graphic illustrations.

Finally, thanks to my wife Linda who provided important doses of encouragement and kept me on task. This work is dedicated to her memory.

The Pennsylvania State University Department of Meteorology supplied the graphics for the weather map in Fig. 1.1.

DANIEL ZELTERMAN
Hamden, CT
August 25, 2009

The following figures are copyright *The New York Times* and used with permission: Figure 1.4 (June 23, 2007); Figure 1.6 (August 15, 2008); Figure 1.7 (August 4, 2008); Figure 1.8 (August 23, 2008); Figure 1.9 (January 8, 2009); Figure 1.10 (October 14, 2008); Figure 4.4 (April 17, 2008); Figure 6.1 (January 17, 2008); Figure 7.1 (June 13, 2007); Figure 10.3 (May 30, 2007). All rights are reserved by *The New York Times*.

Table 2.1: From J. P. Frisby and J. L. Clatworthy, "Learning to see complex random-dot stereograms," *Perception* 4(2), pp. 173–78. Copyright 1975 Pion Limited, London.

Figure 2.1: Courtesy of John Huchra.

Table 2.8: From Marcello Pagano and Kimberlee Gauvreau. *Principles of Biostatistics*, 2E. Copyright 2000 Brooks/Cole, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions

Table 2.10: From N. Teasdale, C. Bard, J. Larue, et al., "On the cognitive penetrability of posture control," *Experimental Aging Research*. Copyright 1993 Taylor & Francis, reprinted by permission of the publisher (Taylor & Francis Group, http://www.informaworld.com).

Tables 5.1 and 6.15: From Frederick Mosteller and John W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*, Table 5.1 pp. 73–74 and Table 6.15 pp. 549–51. Copyright 1977 Addison-Wesley Publishing Company, Inc. Reproduced by permission of Pearson Education, Inc.

Table 6.10: From Douglas Bates and Donald Watts, *Nonlinear Regression Analysis and Its Applications*. Copyright 2007 John Wiley and Sons, Inc. Reprinted with permission of John Wiley and Sons, Inc.

Table 6.11: From James A. Koziol and Donna A. Maxwell, "A distribution-free test for tumor-growth curve analysis with application to an animal tumor immunotherapy experiment," *Biometrics* 37, pp. 383–90. Reprinted with permission of Oxford University Press.

Table 7.1: From A. J. Dobson, *Introduction to Generalized Linear Models*, 2E. Copyright 2001 Taylor & Francis Group LLC – Books. Reproduced with permission of Taylor & Francis Group LLC – Books in the format textbook via Copyright Clearance Center.

Table 9.1: From R. G. Miller et al., *Biostatistics Casebook 318*. Copyright 1980 John Wiley and Sons, Inc. Reprinted with permission of John Wiley and Sons, Inc.

Table 10.7: From P. J. Antsaklis and K. M. Passino, *Introduction to Intelligent and Autonomous Control*, Chapter 2, pp. 27–56; Figure 2 of James S. Albus, "A Reference Model Architecture for Intelligent Systems Design." Copyright 1993 Kluwer Academic Publishers with kind permission of Springer Science and Business Media.

Table 10.11: From Pieter Joost van Wattum et al., "Patterns of response to acute Naxolone infusion in Tourette's Syndrome," *Movement Disorders* 15 (2000) pp. 1252–54. Reprinted with permission of Oxford University Press.

Table 11.2: From Jennifer Wheler et al., "Survival of patients in a phase 1 clinic," *Cancer* 115 (5), pp. 1091–99. Copyright 2009 American Cancer Society.

Table 11.4: From Thomas R. Fleming et al., "Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data," *Biometrics* 36 (1980), pp. 607–25. Reprinted with permission of Oxford University Press.

Table 12.2: From Nicholas Lange et al., *Case Studies in Biometry*. Copyright 1994 John Wiley and Sons, Inc. Reprinted with permission of John Wiley and Sons, Inc.

Table 12.4: From J. M. Krall, V. A. Uthoff, and J. B. Harley. "A step-up procedure for selecting variables associated with survival," *Biometrics* 31 (1975), pp. 49–57. Reprinted with permission of Oxford University Press.