Applied Linear Models with SAS

DANIEL ZELTERMAN

CAMBRIDGE

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.

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.

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

Contents

Preface		*page* ix
Acknowledgments		xiii
1	Introduction	1
1.1	What Is Statistics?	1
1.2	Statistics in the News: The Weather Map	4
1.3	Mathematical Background	6
1.4	Calculus	7
1.5	Calculus in the News: New Home Sales	9
1.6	Statistics in the News: IMF Loans and Tuberculosis	11
1.7	Exercises	13
2	Principles of Statistics	21
2.1	Binomial Distribution	21
2.2	Confidence Intervals and the Hubble Constant	25
2.3	Normal Distribution	26
2.4	Hypothesis Tests	30
2.5	The Student t-Test	34
2.6	The Chi-Squared Test and 2 × 2 Tables	42
2.7	What Are Degrees of Freedom?	47
2.8	SAS, in a Nutshell	49
2.9	Survey of the Rest of the Book	51
2.10	Exercises	52
3	Introduction to Linear Regression	58
3.1	Low-Birth-Weight Infants	58
3.2	The Least Squares Regression Line	59
3.3	Regression in SAS	63
3.4	Statistics in the News: Future Health Care Costs	65
3.5	Exercises	66

4 Assessing the Regression 75
4.1 Correlation 75
4.2 Statistics in the News: Correlations of the Global Economy 77
4.3 Analysis of Variance 78
4.4 Model Assumptions and Residual Plots 81
4.5 Exercises 84

5 Multiple Linear Regression 90
5.1 Introductory Example: Maximum January Temperatures 90
5.2 Graphical Displays of Multivariate Data 94
5.3 Leverage and the Hat Matrix Diagonal 96
5.4 Jackknife Diagnostics 99
5.5 Partial Regression Plots and Correlations 102
5.6 Model-Building Strategies 105
5.7 Exercises 110

6 Indicators, Interactions, and Transformations 120
6.1 Indicator Variables 120
6.2 Synergy in the News: Airline Mergers 127
6.3 Interactions of Explanatory Variables 128
6.4 Transformations 132
6.5 Additional Topics: Longitudinal Data 137
6.6 Exercises 138

7 Nonparametric Statistics 150
7.1 A Test for Medians 150
7.2 Statistics in the News: Math Achievement Scores 153
7.3 Rank Sum Test 155
7.4 Nonparametric Methods in SAS 156
7.5 Ranking and the Healthiest State 157
7.6 Nonparametric Regression: LOESS 160
7.7 Exercises 163

8 Logistic Regression 169
8.1 Example 169
8.2 The Logit Transformation 170
8.3 Logistic Regression in SAS 173
8.4 Statistics in the News: The New York Mets 177
8.5 Key Points 178
8.6 Exercises 179

9 Diagnostics for Logistic Regression 187
9.1 Some Syntax for `proc logistic` 188
9.2 Residuals for Logistic Regression 190

9.3 Influence in Logistic Regression 193
9.4 Exercises 197

10 Poisson Regression 204
10.1 Statistics in the News: Lottery Winners 204
10.2 Poisson Distribution Basics 204
10.3 Regression Models for Poisson Data 206
10.4 Statistics in the News: Attacks in Iraq 208
10.5 Poisson Regression in SAS 209
10.6 Exercises 215

11 Survival Analysis 225
11.1 Censoring 225
11.2 The Survival Curve and Its Estimate 227
11.3 The Log-Rank Test and SAS Program 232
11.4 Exercises 235

12 Proportional Hazards Regression 237
12.1 The Hazard Function 237
12.2 The Model of Proportional Hazards Regression 239
12.3 Proportional Hazards Regression in SAS 241
12.4 Exercises 243

13 Review of Methods 247
13.1 The Appropriate Method 247
13.2 Other Review Questions 249

Appendix: Statistical Tables 255
A.1 Normal Distribution 255
A.2 Chi-squared Tables 257

References 259
Selected Solutions and Hints 263
Index 269

1

Introduction

We are surrounded by data. With a tap at a computer keyboard, we have access to more than we could possibly absorb in a lifetime. But is this data the same as information? How do we get from numbers to understanding? How do we identify simplifying trends – but also find exceptions to the rule? The computers that provide access to the data also provide the tools to answer these questions. Unfortunately, owning a hammer does not enable us to build a fine house. It takes experience using the tools, knowing when they are appropriate, and also knowing their limitations.

The study of statistics provides the tools to create understanding out of raw data. Expertise comes with experience, of course. We need equal amounts of theory (in the form of statistical tools), technical skills (at the computer), and critical analysis (identifying the limitations of various methods for each setting). A lack of one of these cannot be made up by the other two.

This chapter provides a review of statistics in general, along with the mathematical and statistical prerequisites that will be used in subsequent chapters. Even more broadly, the reader will be reminded of the larger picture. It is very easy to learn many statistical methods only to lose sight of the point of it all.

1.1 What Is Statistics?

In an effort to present a lot of mathematical formulas, we sometimes lose track of the central idea of the discipline. It is important to remember the big picture when we get too close to the subject.

Let us consider a vast wall that separates our lives from the place where the information resides. It is impossible to see over or around this wall, but every now and then we have the good fortune of having some pieces of data thrown over to us. On the basis of this fragmentary sampled data, we are supposed to infer the composition of the remainder on the other side. This is the aim of *statistical inference*.

The population is usually vast and infinite, whereas the sample is just a handful of numbers.

> In statistical inference we infer properties of the population from the sample.

There is an enormous possibility for error, of course. If all of the left-handed people I know also have artistic ability, am I allowed to generalize this to a statement that all left-handed people are artistic? I may not know very many left-handed people. In this case I do not have much data to make my claim, and my statement should reflect a large possibility of error. Maybe most of my friends are also artists. In this case we say that the sampled data is *biased* because it contains more artists than would be found in a representative sample of the population.

The population in this example is the totality of all left-handed people. Maybe the population should be *all* people, if we also want to show that artistic ability is greater in left-handed people than in right-handed people. We can't possibly measure such a large group. Instead, we must resign ourselves to the observed or *empirical* data made up of the people we know. This is called a *convenience sample* because it is not really random and may not be representative.

Consider next the separate concepts of sample and population for numerically valued data. The sample *average* is a number that we use to infer the value of the population *mean*. The average of several numbers is itself a number that we obtain. The population mean, however, is on the other side of the imaginary wall and is not observable. In fact, the population mean is almost an unknowable quantity that could not be observed even after a lifetime of study. Fortunately, statistical inference allows us to make statements about the population mean on the basis of the sample average. Sometimes we forget that this inference is taking place and will confuse the sample statistic with the population attribute.

> Statistics are functions of the sampled data. Parameters are properties of the population.

Often the sampled data comes at great expense and through personal hardship, as in the case of clinical trials of new therapies for life-threatening diseases. In a clinical trial involving cancer, for example, costs are typically many thousands of dollars per patient enrolled. Innovative therapies can easily cost ten times that amount. Sometimes the most important data consists of a single number, such as how long the patient lived, recorded only after the patient loses the fight with his or her disease.

Sometimes we attempt to collect all of the data, as in the case of a *census*. The U.S. Constitution specifically mandates that a complete census of the population be performed every ten years.[1] The writers of the Constitution knew that in order to

[1] Article 1, Section 2 reads, in part: "Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be

have a representative democracy and a fair tax system, we also need to know where the people live and work. The composition of the House of Representatives is based on the decennial census. Locally, communities need to know about population shifts to plan for schools and roads. Despite the importance of the census data, there continues to be controversy on how to identify and count certain segments of the population, including the homeless, prison inmates, migrant workers, college students, and foreign persons living in the country without appropriate documentation.

Statistical inference is the process of generalizing from a sample of data to the larger population. The sample average is a simple statistic that immediately comes to mind. The Student t-test is the principal method used to make inferences about the population mean on the basis of the sample average. We review this method in Section 2.5. The sample *median* is the value at which half of the sample is above and half is below. The median is discussed in Chapter 7.

The standard deviation measures how far individual observations deviate from their average.

The sample *standard deviation* allows us to estimate the scale of variability in the population. On the basis of the normal distribution (Section 2.3), we usually expect about 68% of the population to appear within one standard deviation (above or below) of the mean. Similarly, about 95% of the population should occur within two standard deviations of the population mean.

The standard error measures the sampling variability of the mean.

A commonly used measure related to the standard deviation is the *standard error*, also called the *standard error of the mean* and often abbreviated SEM. These two similar-sounding quantities refer to very different measures. The standard error estimates the standard deviation associated with the sample average. As the sample size increases, the standard deviation (which refers to individuals in the population) should not appreciably change. On the other hand, a large sample size is associated with a precise estimate of the population mean as a consequence of a small standard error. This relationship provides the incentive for larger sample sizes, allowing us to estimate the population mean more accurately. The relationship is

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{\text{Sample size}}}$$

determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct."

Consider a simple example. We want to measure the heights of a group of people. There will always be tall people, and there will always be short people, so changing the sample size does not appreciably alter the standard deviation of the data. Individual variations will always be observed. If we were interested in estimating the average height, then the standard error will decrease with an increase in the sample size (at a rate of $1/\sqrt{\text{sample size}}$), motivating the use of ever-larger samples. The average will be measured with greater precision, and this precision is described in terms of the standard error. Similarly, if we want to measure the average with twice the precision, then we will need a sample size four times larger.

Another commonly used term associated with the standard deviation is *variance.* The relationship between the variance and the standard deviation is

Variance = (Standard deviation)^2

The standard deviation and variance are obtained in SAS using `proc univariate`, for example. The formula appears often, and the reader should be familiar with it, even though its value will be calculated using a computer.

Given observed sample values $x_1, x_2, \ldots, x_n$, we compute the *sample variance* from

$$s^2 = \text{sample variance} = \frac{1}{n-1} \sum_i (x_i - \overline{x})^2, \qquad (1.1)$$

where $\overline{x}$ is the average of the observed values.

This estimate is often denoted by the symbol s^2. Similarly, the estimated sample standard deviation s is the square root of this estimator. Intuitively, we see that (1.1) averages the squared difference between each observation and the sample average, except that the denominator is one less than the sample size. The "$n - 1$" term is the degrees of freedom for this expression and is described in Sections 2.5 and 2.7.

1.2 Statistics in the News: The Weather Map

Sometimes it is possible to be overwhelmed with too much information. The business section of the newspaper is filled with stock prices, and the sports section has a wealth of scores and data on athletic endeavors. The business section frequently has several graphs and charts illustrating trends, rates, and prices. The sports pages have yet to catch up with the business section in terms of aids for the reader.

As an excellent way to summarize and display a huge amount of information, we reproduce the U.S. weather map from October 27, 2008, in Figure 1.1. There are several levels of information depicted here, all overlaid on top of one another. First we recognize the geographic-political map indicating the shorelines and state boundaries. The large map at the top provides the details of that day's weather. The large Hs indicate the locations of high barometric pressure centers. Regions with

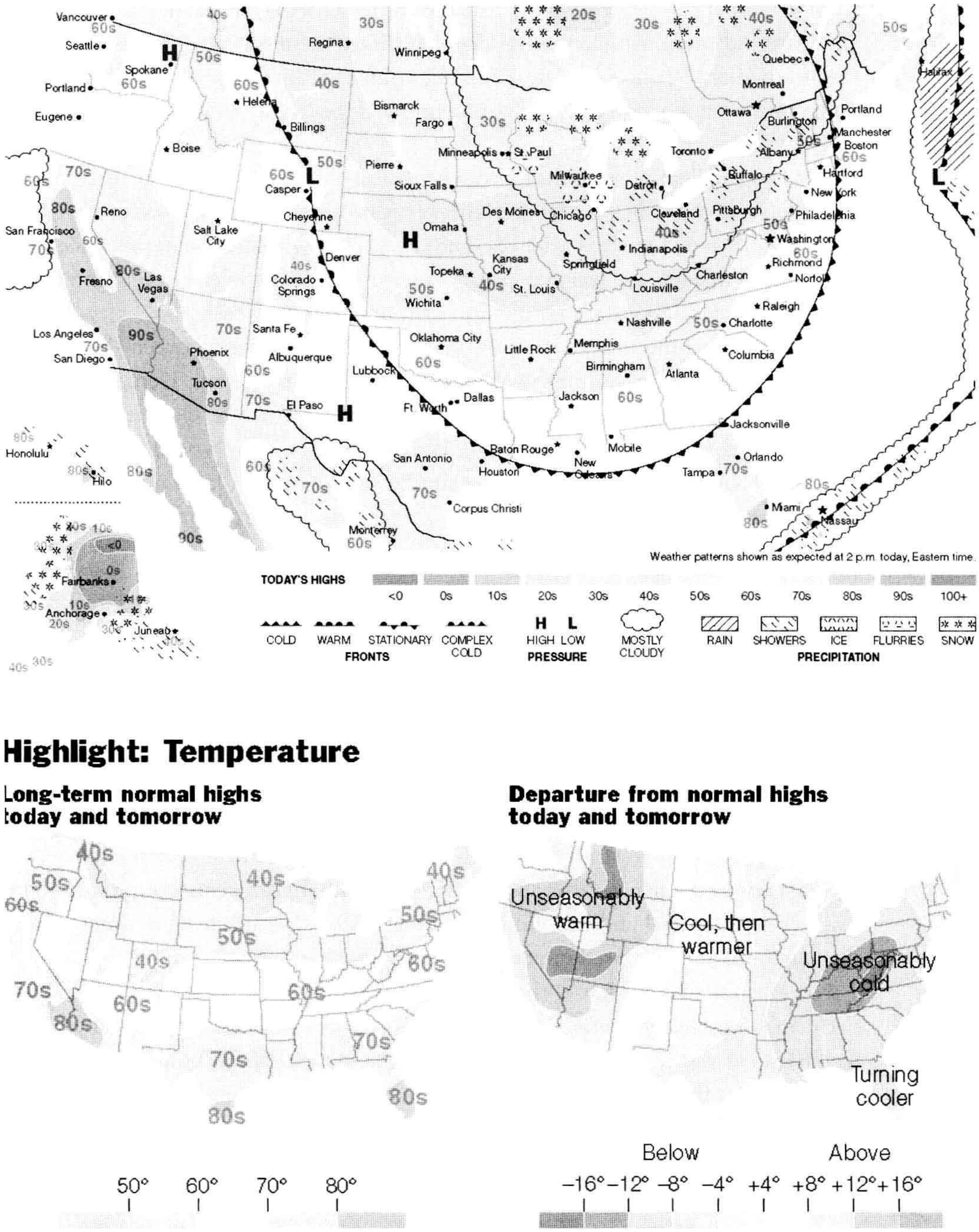

Figure 1.1 The U.S. weather map for October 27, 2008: Observed, expected, and residual data. Courtesy of Pennsylvania State University, Department of Meteorology.