Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes
Atkinson/Riani: Robust Diagnostic Regression Analysis
Atkinson/Riani/Ceriloi: Exploring Multivariate Data with the Forward Search
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition
Berk: Statistical Learning from a Regression Perspective
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition
Bucklew: Introduction to Rare Event Simulation
Cappé/Moulines/Rydén: Inference in Hidden Markov Models
Chan/Tong: Chaos: A Statistical Perspective
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation
Coles: An Introduction to Statistical Modeling of Extreme Values
Devroye/Lugosi: Combinatorial Methods in Density Estimation
Diggle/Ribeiro: Model-based Geostatistics
Dudoit/Van der Laan: Multiple Testing Procedures with Applications to Genomics
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation
Fahrmeir/Tutz: Multivariate Statistical Modeling Based on Generalized Linear Models, 2nd edition
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods
Ferraty/Vieu: Nonparametric Functional Data Analysis: Theory and Practice
Ferreira/Lee: Multiscale Modeling: A Bayesian Perspective
Fienberg/Hoaglin: Selected Papers of Frederick Mosteller
Frühwirth-Schnatter: Finite Mixture and Markov Switching Models
Ghosh/Ramamoorthi: Bayesian Nonparametrics
Glaz/Naus/Wallenstein: Scan Statistics
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition
Gouriéroux: ARCH Models and Financial Applications
Gu: Smoothing Spline ANOVA Models
Gyöfi/Kohler/Krzyżak/Walk: A Distribution-Free Theory of Nonparametric Regression
Haberman: Advanced Statistics, Volume I: Description of Populations
Hall: The Bootstrap and Edgeworth Expansion
Härdle: Smoothing Techniques: With Implementation in S
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis
Hart: Nonparametric Smoothing and Lack-of-Fit Tests
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation
Huet/Bouvier/Poursat/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2nd edition

(continued after index)

Richard A. Berk

Statistical Learning from a Regression Perspective

Springer

Richard A. Berk
Department of Criminology, School of Arts and Sciences
Department of Statistics, The Wharton School
The University of Pennsylvania
Philadelphia, PA
USA

ISBN: 978-0-387-77500-5 e-ISBN: 978-0-387-77501-2
DOI: 10.1007/978-0-387-77501-2

Library of Congress Control Number: 2008926886

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

In memory of Peter H. Rossi, a mentor, colleague, and friend.

"In God we trust. All others must have data."

(W. Edwards Deming)

Preface

As I was writing my recent book on regression analysis (Berk, 2003), I was struck by how few alternatives to conventional regression there were. In the social sciences, for example, one either did causal modeling econometric style or largely gave up quantitative work. The life sciences did not seem quite so driven by causal modeling, but causal modeling was a popular tool. As I argued at length in my book, causal modeling as commonly undertaken is a loser.

There also seemed to be a more general problem. Across a range of scientific disciplines there was too often little interest in statistical tools emphasizing induction and description. With the primary goal of getting the "right" model and its associated p-values, the older and interesting tradition of exploratory data analysis had largely become an under-the-table activity; the approach was in fact commonly used, but rarely discussed in polite company. How could one be a real scientist, guided by "theory" and engaged in deductive model testing, while at the same time snooping around in the data to determine which models to test? In the battle for prestige, model testing had won.

Around the same time, I became aware of some new developments in applied mathematics, computer sciences, and statistics making data exploration a virtue. And with the virtue came a variety of new ideas and concepts, coupled with the very latest in statistical computing. These new approaches, variously identified as "data mining," "statistical learning," "machine learning," and other names, were being tried in a number of the natural and biomedical sciences, and the initial experience looked promising.

As I started to read more deeply, however, I was struck by how difficult it was to work across writings from such disparate disciplines. Even when the material was essentially the same, it was very difficult to tell if it was. Each discipline brought it own goals, concepts, naming conventions, and (maybe worst of all) notation to the table.

In the midst of trying to impose some of my own order on the material, I came upon *The Elements of Statistical Learning*, by Trevor Hastie, Robert

Tibshirani, and Jerome Friedman (Springer-Verlag, 2001). I saw in the book a heroic effort to integrate a very wide variety of data analysis tools. I learned from the book and was then able to approach more primary material within a useful framework.

This book is my attempt to integrate some of the same material and some new developments of the past six years. Its intended audience is practitioners in the social, biomedical, and ecological sciences. Applications to real data addressing real empirical questions are emphasized. Although considerable effort has gone into providing explanations of why the statistical procedures work the way they do, the required mathematical background is modest. A solid course or two in regression analysis and some familiarity with resampling procedures should suffice. A good benchmark for regression is Freedman's *Statistical Models: Theory and Practice* (2005). A good benchmark for resampling is Manly's *Randomization, Bootstrap, and Monte Carlo Methods in Biology* 1997. Matrix algebra and calculus are used only as languages of exposition, and only as needed. There are no proofs to be followed.

The procedures discussed are limited to those that can be viewed as a form of regression analysis. As explained more completely in the first chapter, this means concentrating on statistical tools for which the conditional distribution of a response variable is the defining interest and for which characterizing the relationships between predictors and the response is undertaken in a serious and accessible manner.

Regression analysis provides a unifying theme that will ease translations across disciplines. It will also increase the comfort level for many scientists and policy analysts for whom regression analysis is a key data analysis tool. At the same time, a regression framework will highlight how the approaches discussed can be seen as alternatives to conventional causal modeling.

Because the goal is to convey how these procedures can be (and are being) used in practice, the material requires relatively in-depth illustrations and rather detailed information on the context in which the data analysis is being undertaken. The book draws heavily, therefore, on datasets with which I am very familiar. The same point applies to the software used and described.

The regression framework comes at a price. A 2005 announcement for a conference on data mining sponsored by the Society for Industrial and Applied Mathematics (SIAM) listed the following topics: query/constraint-based data mining, trend and periodicity analysis, mining data streams, data reduction/preprocessing, feature extraction and selection, postprocessing, collaborative filtering/personalization, cost-based decision making, visual data mining, privacy-sensitive data mining, and lots more. Many of these topics cannot be considered a form of regression analysis. For example, procedures used for edge detection (e.g., determining the boundaries of different kinds of land use from remote sensing data) are basically a filtering process to remove noise from the signal.

Another class of problems makes no distinction between predictors and responses. The relevant techniques can be closely related, at least in spirit, to procedures such as factor analysis and cluster analysis. One might explore, for example, the interaction patterns among children at school: who plays with whom. These too are not discussed.

Other topics can be considered regression analysis only as a formality. For example, a common data mining application in marketing is to extract from the purchasing behavior of individual shoppers patterns that can be used to forecast future purchases. But there are no predictors in the usual regression sense. The conditioning is on each individual shopper. The question is not what features of shoppers predict what they will purchase, but what a given shopper is likely to purchase.

Finally, there are a large number of procedures that focus on the conditional distribution of the response, much as with any regression analysis, but with little attention to how the predictors are related to the response (Horváth and Yamamoto, 2006; Camacho et al., 2006). Such procedures neglect a key feature of regression analysis, at least as discussed in this book, and are not considered. That said, there is no principled reason in many cases why the role of each predictor could not be better represented, and perhaps in the near future that shortcoming will be remedied.

In short, although using a regression framework implies a big-tent approach to the topics included, it is not an exhaustive tent. Many interesting and powerful tools are not discussed. Where appropriate, however, references to that material are provided.

I may have gone a bit overboard with the number of citations I provide. The relevant literatures are changing and growing rapidly. Today's breakthrough can be tomorrow's bust, and work that by current thinking is uninteresting can be the spark for dramatic advances in the future. At any given moment, it can be difficult to determine which is which. In response, I have attempted to provide a rich mix of background material, even at the risk of not being sufficiently selective. (And I have probably missed some useful papers nevertheless.)

In the material that follows, I have tried to use consistent notation. This has proved to be very difficult because of important differences in the conceptual traditions represented and the complexity of statistical tools discussed. For example, it is common to see the use of the expected value operator even when the data cannot be characterized as a collection of random variables and when the sole goal is description.

I draw where I can from the notation used in *The Elements of Statistical Learning* (Hastie et al., 2001). Thus, the symbol X is used for an input variable, or predictor in statistical parlance. When X is a set of inputs to be treated as a vector, each component is indexed by a subscript (e.g., X_j). Quantitative outputs, also called response variables, are represented by Y, and categorical outputs, another kind of response variable, are represented by G with K categories. Upper case letters are used to refer to variables in a general way, with details to follow as needed. Sometimes these variables are treated as random variables, and sometimes not. I try to make that clear in context.

Observed values are shown in lower case, usually with a subscript. Thus x_i is the ith observed value for the variable X. Sometimes these observed values are nothing more than the data on hand. Sometimes they are realizations of random variables. Again, I try to make this clear in context.

Matrices are represented in bold uppercase. For example, in matrix form the usual set of p predictors, each with N observations, is an $N \times p$ matrix $\mathbf{X}$. The subscript i is generally used for observations and the subscript j for variables. Bold lowercase letters are used for vectors with N elements, commonly columns of $\mathbf{X}$. Other vectors are generally not represented in boldface fonts, but again, I try to make this clear in context.

If one treats Y as a random variable, its observed values y are either a random sample from a population or a realization of a stochastic process. The conditional means of the random variable Y for various configurations of $\mathbf{X}$-values are commonly referred to as "expected values," and are either the conditional means of Y for different configurations of $\mathbf{X}$-values in the population or for the stochastic process by which the data were generated. A common notation is $\mathrm{E}(\mathrm{Y}|\mathbf{X})$. The $\mathrm{E}(\mathrm{Y}|\mathbf{X})$ is also often called a "parameter." The conditional means computed from the data are often called "sample statistics," or in this case, "sample means." In the regression context, the sample means are commonly referred to as the fitted values, often written as $\hat{y}|\mathbf{X}$. Subscripting can follow as already described.

Unfortunately, after that it gets messier. First, I often have to decipher the intent in the notation used by others. No doubt I sometimes get it wrong. For example, it is often unclear if a computer algorithm is formally meant to be an estimator or a descriptor.

Second, there are some complications in representing nested realizations of the same variable (as in the bootstrap), or model output that is subject to several different chance processes. There is a practical limit to the number and types of bars, asterisks, hats, and tildes one can effectively use. I try to provide warnings (and apologies) when things get cluttered.

There are also some labeling issues. When I am referring to the general linear model (i.e., linear regression, analysis of variance, and analysis of covariance), I use the terms classical linear regression, or conventional linear regression. All regressions in which the functional forms are determined before the fitting process begins, I call parametric. All regressions in which the functional forms are determined as part of the fitting process, I call nonparametric. When there is some of both, I call the regressions semiparametric. Sometimes the lines among parametric, nonparametric, and semiparametric are fuzzy, but I try to make clear what I mean in context. Although these naming conventions are roughly consistent with much common practice, they are not universal.

All of the computing done for this book was undertaken in R. R is a programming language designed for statistical computing and graphics. It has become a major vehicle for developmental work in statistics and is increasingly being used by practitioners. A key reason for relying on R for this book is that most of the newest developments in statistical learning and related fields can be found in R. Another reason is that it is free.

Readers familiar with S or S-plus will immediately feel at home; R is basically a "dialect" of S. For others, there are several excellent books providing a good introduction to data analysis using R. Dalgaard (2002), Crawley (2007), and Maindonald and Braun (2007) are all very accessible. Readers who are especially interested in graphics should consult Murrell (2006). The most useful R website can be found at http://www.r-project.org/.

The use of R raises the question of how much R-code to include. The R-code used to construct all of the applications in the book could be made available. However, detailed code is largely not shown. Many of the procedures used are somewhat in flux. Code that works one day may need some tweaking the next. As an alternative, the procedures discussed are identified as needed so that detailed information about how to proceed in R can be easily obtained from R help commands or supporting documentation. In addition, there is a web site where many examples, including the data, can be found (WEB ADDRESS TO BE ADDED). When the data used in this book are proprietary or otherwise not publicly available, similar data and appropriate R-code are substituted.

There are exercises at the end of each chapter. They are meant to be hands-on data analyses built around R. As such, they require some facility with R. However, the goals of each problem are reasonably clear so that other software and datasets can be used. Often the exercises can be usefully repeated with different datasets.

The book has been written so that later chapters depend substantially on earlier chapters. For example, because classification and regression trees (CART) can be an important component of boosting, it may be difficult to follow the discussion of boosting without having read the earlier chapter on CART. However, readers who already have a solid background in material covered earlier should have little trouble skipping ahead. The notation and terms used are reasonably standard or can be easily figured out. In addition, the final chapter can be read at almost any time. One reviewer suggested that much of the material could be usefully brought forward to Chapter 1.

Finally, there is the matter of tone. The past several decades have seen the development of a dizzying array of new statistical procedures, sometimes introduced with the hype of a big-budget movie. Advertising from major statistical software providers has typically made things worse. Although there have been genuine and useful advances, none of the techniques have ever lived up to their most optimistic billing. Widespread misuse has further increased the gap between promised performance and actual performance. In this book, therefore, the tone will be cautious, some might even say dark. I hope this will not discourage readers from engaging seriously with the material. The intent is to provide a balanced discussion of the limitations as well as the strengths of the statistical learning procedures.

While working on this book, I was able to rely on support from several sources. Much of the work was funded by a grant from the National Science Foundation: SES-0437169, “Ensemble Methods for Data Analysis in the Behavioral, Social and Economic Sciences.” The first draft was completed while I was on sabbatical at the Department of Earth, Atmosphere, and Oceans, at the Ecole Normale Supérieur in Paris. The second draft was completed after I moved from UCLA to the University of Pennsylvania. All three locations provided congenial working environments. Most important, I benefited enormously from discussions about statistical learning with colleagues at UCLA, Penn and elsewhere: Larry Brown, Andreas Buja, Jan de Leeuw, David Freedman, Mark Hansen, Andy Liaw, Greg Ridgeway, Bob Stine, Mikhail Traskin and Adi Wyner. Each is knowledgeable, smart and constructive. I also learned a great deal from several very helpful, anonymous reviews. Dick Koch was enormously helpful and patient when I had problems making TeXShop perform properly. Finally, I have benefited over the past several years from interacting with talented graduate students: Yan He, Weihua Huang, Brian Kriegler, and Jie Shen. Brian Kriegler deserves a special thanks for working through the exercises at the end of each chapter.

Certain datasets and analyses were funded as part of research projects undertaken for the California Policy Research Center, The Inter-America Tropical Tuna Commission, the National Institute of Justice, the County of Los Angeles, the California Department of Correction and Rehabilitation, the Los Angeles Sheriff’s Department, and the Philadelphia Department of Adult Probation and Parole. Support from all of these sources is gratefully acknowledged.

Contents