Statistical Models: Theory and Practice

This lively and engaging textbook explains the things you have to know in order to read empirical papers in the social and health sciences, as well as the techniques you need to build statistical models of your own. The author, David A. Freedman, explains the basic ideas of association and regression, and takes you through the current models that link these ideas to causality.

The focus is on applications of linear models, including generalized least squares and two-stage least squares, with probits and logits for binary variables. The bootstrap is developed as a technique for estimating bias and computing standard errors. Careful attention is paid to the principles of statistical inference. There is background material on study design, bivariate regression, and matrix algebra. To develop technique, there are computer labs with sample computer programs. The book is rich in exercises, most with answers.

Target audiences include advanced undergraduates and beginning graduate students in statistics, as well as students and professionals in the social and health sciences. The discussion in the book is organized around published studies, as are many of the exercises. Relevant journal articles are reprinted at the back of the book. Freedman makes a thorough appraisal of the statistical methods in these papers and in a variety of other examples. He illustrates the principles of modeling, and the pitfalls. The discussion shows you how to think about the critical issues—including the connection (or lack of it) between the statistical models and the real phenomena.

Features of the book

- Authoritative guide by a well-known author with wide experience in teaching, research, and consulting
- Will be of interest to anyone who deals with applied statistics
- No-nonsense, direct style
- Careful analysis of statistical issues that come up in substantive applications, mainly in the social and health sciences
- Can be used as a text in a course or read on its own
- Developed over many years at Berkeley, thoroughly class tested
- Background material on regression and matrix algebra
- Plenty of exercises
- Extra material for instructors, including data sets and MATLAB code for lab projects (send email to solutions@cambridge.org)

The author

David A. Freedman (1938–2008) was Professor of Statistics at the University of California, Berkeley. He was a distinguished mathematical statistician whose theoretical research ranged from the analysis of martingale inequalities, Markov processes, de Finetti's theorem, consistency of Bayes estimators, sampling, the bootstrap, and procedures for testing and evaluating models to methods for causal inference.

Freedman published widely on the application—and misapplication—of statistics in the social sciences, including epidemiology, demography, public policy, and law. He emphasized exposing and checking the assumptions that underlie standard methods, as well as understanding how those methods behave when the assumptions are false—for example, how regression models behave when fitted to data from randomized experiments. He had a remarkable talent for integrating carefully honed statistical arguments with compelling empirical applications and illustrations, as this book exemplifies.

Freedman was a member of the American Academy of Arts and Sciences, and in 2003 received the National Academy of Science's John J. Carty Award, for his "profound contributions to the theory and practice of statistics."

Cover illustration

The ellipse on the cover shows the region in the plane where a bivariate normal probability density exceeds a threshold level. The correlation coefficient is 0.50. The means of x and y are equal. So are the standard deviations. The dashed line is both the major axis of the ellipse and the SD line. The solid line gives the regression of y on x. The normal density (with suitable means and standard deviations) serves as a mathematical idealization of the Pearson-Lee data on heights, discussed in chapter 2. Normal densities are reviewed in chapter 3.

Statistical Models: Theory and Practice

David A. Freedman
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

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

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

www.cambridge.org
Information on this title: www.cambridge.org/9780521743853

First published 2009

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

Freedman, David, 1938–
Statistical models : theory and practice / David A. Freedman.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-74385-3 (pbk.)
1. Linear models (Statistics) 2. Bootstrap (Statistics)
3. Mathematical statistics. I. Title.
QA279.F74 2009
519.5–dc22 2009002062

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate. Information regarding prices, travel timetables, and other factual information given in this work are correct at the time of first printing, but Cambridge University Press does not guarantee the accuracy of such information thereafter.

Table of Contents

Reprints

Foreword to the Revised Edition

Some books are correct. Some are clear. Some are useful. Some are entertaining. Few are even two of these. This book is all four. *Statistical Models: Theory and Practice* is lucid, candid and insightful, a joy to read. We are fortunate that David Freedman finished this new edition before his death in late 2008. We are deeply saddened by his passing, and we greatly admire the energy and cheer he brought to this volume—and many other projects—during his final months.

This book focuses on half a dozen of the most common tools in applied statistics, presenting them crisply, without jargon or hyperbole. It dissects real applications: a quarter of the book reprints articles from the social and life sciences that hinge on statistical models. It articulates the assumptions necessary for the tools to behave well and identifies the work that the assumptions do. This clarity makes it easier for students and practitioners to see where the methods will be reliable; where they are likely to fail, and how badly; where a different method might work; and where no inference is possible—no matter what tool somebody tries to sell them.

Many texts at this level are little more than bestiaries of methods, presenting dozens of tools with scant explication or insight, a cookbook, numbers-are-numbers approach. “If the left hand side is continuous, use a linear model; fit by least-squares. If the left hand side is discrete, use a logit or probit model; fit by maximum likelihood.” Presenting statistics this way invites students to believe that the resulting parameter estimates, standard errors, and tests of significance are meaningful—perhaps even untangling complex causal relationships. They teach students to think scientific inference is purely algorithmic. Plug in the numbers; out comes science. This undervalues both substantive and statistical knowledge.

To select an appropriate statistical method actually requires careful thought about how the data were collected and what they measure. Data are not “just numbers.” Using statistical methods in situations where the underlying assumptions are false can yield gold or dross—but more often dross.

Statistical Models brings this message home by showing both good and questionable applications of statistical tools in landmark research: a study of political intolerance during the McCarthy period, the effect of Catholic schooling on completion of high school and entry into college, the relationship between fertility and education, and the role of government institutions in shaping social capital. Other examples are drawn from medicine and epidemiology, including John Snow's classic work on the cause of cholera—a shining example of the success of simple statistical tools when paired with substantive knowledge and plenty of shoe leather. These real applications bring the theory to life and motivate the exercises.

The text is accessible to upper-division undergraduates and beginning graduate students. Advanced graduate students and established researchers will also find new insights. Indeed, the three of us have learned much by reading it and teaching from it.

And those who read this textbook have not exhausted Freedman's approachable work on these topics. Many of his related research articles are collected in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (Cambridge University Press, 2009), a useful companion to this text. The collection goes further into some applications mentioned in the textbook, such as the etiology of cholera and the health effects of Hormone Replacement Therapy. Other applications range from adjusting the census for undercount to quantifying earthquake risk. Several articles address theoretical issues raised in the textbook. For instance, randomized assignment in an experiment is not enough to justify regression: without further assumptions, multiple regression estimates of treatment effects are biased. The collection also covers the philosophical foundations of statistics and methods the textbook does not, such as survival analysis.

Statistical Models: Theory and Practice presents serious applications and the underlying theory without sacrificing clarity or accessibility. Freedman shows with wit and clarity how statistical analysis can inform and how it can deceive. This book is unlike any other, a treasure: an introductory book that conveys some of the wisdom required to make reliable statistical inferences. It is an important part of Freedman's legacy.

David Collier, Jasjeet Singh Sekhon, and Philip B. Stark
University of California, Berkeley

Preface

This book is primarily intended for advanced undergraduates or beginning graduate students in statistics. It should also be of interest to many students and professionals in the social and health sciences. Although written as a textbook, it can be read on its own. The focus is on applications of linear models, including generalized least squares, two-stage least squares, probits and logits. The bootstrap is explained as a technique for estimating bias and computing standard errors.

The contents of the book can fairly be described as what you have to know in order to start reading empirical papers that use statistical models. The emphasis throughout is on the connection—or lack of connection—between the models and the real phenomena. Much of the discussion is organized around published studies; the key papers are reprinted for ease of reference. Some observers may find the tone of the discussion too skeptical. If you are among them, I would make an unusual request: suspend belief until you finish reading the book. (Suspension of disbelief is all too easily obtained, but that is a topic for another day.)

The first chapter contrasts observational studies with experiments, and introduces regression as a technique that may help to adjust for confounding in observational studies. There is a chapter that explains the regression line, and another chapter with a quick review of matrix algebra. (At Berkeley, half the statistics majors need these chapters.) The going would be much easier with students who know such material. Another big plus would be a solid upper-division course introducing the basics of probability and statistics.

Technique is developed by practice. At Berkeley, we have lab sessions where students use the computer to analyze data. There is a baker's dozen of these labs at the back of the book, with outlines for several more, and there are sample computer programs. Data are available to instructors from the publisher, along with source files for the labs and computer code: send email to solutions@cambridge.org.

A textbook is only as good as its exercises, and there are plenty of them in the pages that follow. Some are mathematical and some are hypothetical, providing the analogs of lemmas and counter-examples in a more conventional treatment. On the other hand, many of the exercises are based on actual studies. Here is a summary of the data and the analysis; here is a specific issue: where do you come down? Answers to most of the exercises are at the back of the book. Beyond exercises and labs, students at Berkeley write papers during the semester. Instructions for projects are also available from the publisher.

A text is defined in part by what it chooses to discuss, and in part by what it chooses to ignore; the topics of interest are not to be covered in one book, no matter how thick. My objective was to explain how practitioners infer causation from association, with the bootstrap as a counterpoint to the usual asymptotics. Examining the logic of the enterprise is crucial, and that takes time. If a favorite technique has been slighted, perhaps this reasoning will make amends.

There is enough material in the book for 15–20 weeks of lectures and discussion at the undergraduate level, or 10–15 weeks at the graduate level. With undergraduates on the semester system, I cover chapters 1–7, and introduce simultaneity (sections 9.1–4). This usually takes 13 weeks. If things go quickly, I do the bootstrap (chapter 8), and the examples in chapter 9. On a quarter system with ten-week terms, I would skip the student presentations and chapters 8–9; the bivariate probit model in chapter 7 could also be dispensed with.

During the last two weeks of a semester, students present their projects, or discuss them with me in office hours. I often have a review period on the last day of class. For a graduate course, I supplement the material with additional case studies and discussion of technique.

The revised text organizes the chapters somewhat differently, which makes the teaching much easier. The exposition has been improved in a number of other ways, without (I hope) introducing new difficulties. There are many new examples and exercises.

Acknowledgements

I've taught graduate and undergraduate courses based on this material for many years at Berkeley, and on occasion at Stanford and Athens. The students in those courses were helpful and supportive. I would also like to thank Dick Berk, Máire Ní Bhrolcháin, Taylor Boas, Derek Briggs, David Collier, Persi Diaconis, Thad Dunning, Mike Finkelstein, Paul Humphreys, Jon McAuliffe, Doug Rivers, Mike Roberts, Don Ylvisaker, and PengZhao, along with several anonymous reviewers, for many useful comments. Russ Lyons and Roger Purves were virtual coauthors; David Tranah was an outstanding editor.