Statistical Models

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 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 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 book 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 in applied statistics
- no-nonsense, direct style will appeal to both new and experienced users of statistics
- 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 (email to solutions@cambridge.org)

The author

David A. Freedman is Professor of Statistics at the University of California, Berkeley. He has also taught in Athens, Caracas, Jerusalem, Kuwait, London, Mexico City, and Stanford. He has written several previous books, including a widely used elementary text. He is one of the leading researchers in probability and statistics, with 150 papers in the professional literature.

He is a member of the American Academy of Arts and Sciences. In 2003, he received the John J. Carty Award for the Advancement of Science from the National Academy of Sciences, recognizing his "profound contributions to the theory and practice of statistics."

Freedman has consulted for the Carnegie Commission, the City of San Francisco, and the Federal Reserve, as well as several departments of the U.S. government. He has testified as an expert witness on statistics in law cases that involve employment discrimination, fair loan practices, duplicate signatures on petitions, railroad taxation, ecological inference, flight patterns of golf balls, price scanner errors, sampling techniques, and census adjustment.

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. 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

Cambridge University Press
40 West 20th Street, New York, NY 10011-4211, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521854832

First published 2005

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-13: 978-0-521-85483-2
ISBN-10: 0-521-85483-0
1. Social sciences – Statistics – Methodology. 2. Medical
statistics – Methodology. 3. Regression analysis.
4. Statistics – Methodology. I. Title.
HA29.F678 2005
300′.1′519536 – dc22 2005047097

ISBN-13 978-0-521-85483-2 hardback
ISBN-10 0-521-85483-0 hardback

ISBN-13 978-0-521-67105-7 paperback
ISBN-10 0-521-67105-1 paperback

Table of Contents

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; key papers are reprinted here for ease of reference. Some 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 knew 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 exercises in the pages that follow. Some are mathematical and some are hypothetical, but many of them are based on actual studies. That kind of exercise says, 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. (The best are presented in class, with discussion.) 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 all to be covered in one book, no matter how thick. ANOVA would be natural to discuss, but ANOVA can be viewed—with only some distortion—as a special case of regression. (The ANOVA table for regression is covered in chapter 4, along with the F-test.)

Some discussion of proportional hazards would also be natural. However, logistic regression (chapter 6) is a more common technique in the biomedical literature. Furthermore, proportional-hazard models require a substantial investment in time on risk, survival curves, and hazard rates. All tradeoffs are debatable; otherwise, they wouldn't be tradeoffs. I can only plead the finitude of semesters—never mind quarters—and the necessity of examining the logic of the enterprise as well as the mechanics.

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 1–6, and introduce simultaneity (sections 8.1–4). This usually takes 13 weeks. If things go quickly, I do the examples in chapter 8 and the bootstrap. During the last two weeks of the term, students present their projects. I often have a review period on the last day of class. On a quarter system with ten-week terms, I would skip the student presentations and chapters 7–8; the bivariate probit model in chapter 6 could also be dispensed with. For a graduate course, I supplement the material with additional case studies and discussion of technique.

Acknowledgements

I've taught graduate and undergraduate courses based on this material for many years at Berkeley, and on occasion at Stanford and Athens. I would like to thank the students in those courses for their help and support. 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, David Tranah, Don Ylvisaker, and Peng Zhao, along with several anonymous reviewers, for many useful comments. Russ Lyons was incredibly helpful, and Roger Purves was a virtual coauthor.

David A. Freedman
Berkeley, California
June 2005