ANALYTICAL METHODS FOR SOCIAL RESEARCH

Statistical Modeling and Inference for Social Science

SEAN GAILMARD

Statistical Modeling and Inference for Social Science

SEAN GAILMARD
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107003149

First published 2014

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
Gailmard, Sean.
Statistical modeling and inference for social science / Sean Gailmard, University of California, Berkeley.
pages cm. – (Analytical methods for social research)
Includes bibliographical references and index.
ISBN 978-1-107-00314-9 (hardback)
1. Social sciences – Statistical methods. I. Title.
HA29.G136 2014
519.5–dc23 2013050034

ISBN 978-1-107-00314-9 Hardback

Statistical Modeling and Inference for Social Science

This book provides an introduction to probability theory, statistical inference, and statistical modeling for social science researchers and Ph.D. students. Focusing on the connection between statistical procedures and social science theory, Sean Gailmard develops core statistical theory as a set of tools to model and assess relationships between variables – the primary aim of social scientists. Gailmard explains how social scientists express and test substantive theoretical arguments in various models. Chapter exercises require application of concepts to actual data and extend students' grasp of core theoretical concepts. Students will complete the book with the ability to read and critique statistical applications in their fields of interest.

Sean Gailmard is Associate Professor of Political Science at the University of California, Berkeley. Formerly an assistant professor at Northwestern University and at the University of Chicago, Gailmard earned his Ph.D. in social science (economics and political science) from the California Institute of Technology. He is the coauthor of *Learning While Governing: Institutions and Accountability in the Executive Branch* (2013), winner of the 2013 American Political Science Association's William H. Riker Prize for best book on political economy. His articles have been published in a variety of journals, including *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*. He currently serves as an associate editor for the *Journal of Experimental Political Science* and on the editorial boards for *Political Science Research and Methods* and *Journal of Public Policy*.

To Gina, for continuing to roll the dice with me

List of Figures

List of Tables

Acknowledgments

Various drafts of what became this book have been inflicted on various people. That I still count many of them as friends is a testament to their willingness to forgive and forget. I wish to thank them without implicating them for the results that follow. I have been fortunate to receive feedback, remarks, and critiques from Fred Boehmke, Henry Brady, Wendy Tam Cho, Doug Dion, Jamie Druckman, Chris Elmendorf, Andrew Gelman, Jeff Gill, Sandy Gordon, Shigeo Hirano, Jeff Jenkins, Joel Middleton, Benjamin I. Page, John Patty, Kevin Quinn, Jas Sekhon, Laura Stoker, Dustin Tingley, and Jonathan Wand. Miranda Yaves, Fred and Doug field-tested an early version of this manuscript with their graduate students at Iowa and produced critically important feedback as a result.

Numerous Ph.D. students at Berkeley and Northwestern used this manuscript, mostly with good cheer, as a proto-text in their introductory statistics course. I received useful feedback (and typo corrections) from many. I am especially grateful to Rachel Bernhard, Toby Bolsen, Caroline Brandt, David Broockman, Richard Hay, Chris Chambers Ju, Sebastian Karcher, Elsa Massoc, Gustav Nedergaard, Hong Min Park, David Steinberg, and Jan Vogler. My teaching assistants in various iterations at Berkeley – Sara Chatfield, Shinhye Choi, Stephen Goggin, and Adrienne Hosek – were extremely helpful as I tried to figure out how to explain the concepts in this book to graduate students unfamiliar with them.

Jamie Druckman and Jeff Gill provided essential encouragement in my decision to translate the mish-mash of notes I wrote as supplements to existing texts into this book. Mike Alvarez and Neal Beck, as co-editors of the Analytical Methods for Social Research series at Cambridge, provided extensive, much needed critique, feedback, and support that substantially improved the content and exposition. I am extremely grateful to them for their efforts, and to Mike for shepherding the manuscript and my responses to (very constructive and

helpful) anonymous reviewers through the various stages of the publication process.

I have been fortunate to work with two editors at Cambridge – Ed Parsons on the initial proposal and, for most of the review and publication process, Lew Bateman – who were judicious and supportive in considering the unique features of this manuscript and tolerant of my seemingly interminable delays in revising it.

Most tolerant and helpful of all was my family: my children, William, Aidan, and Collin, and my wife Gina. It was always a delight to try to explain to my small children what I was writing about, even more so to see my efforts partially inspire their proposed manuscripts on the finer points of Minecraft, spiders, and zombies. Our continuing collective struggle with type 1 diabetes is an object lesson in randomness. Whenever I thought I faced daunting demands, the resilience of very small people confronting the vicissitudes of this disease brought me back to earth. It is only because of Gina's graceful management of our real life that I could dedicate time to writing this book, and thus, this book is dedicated to her.

Contents

1

Introduction

PURPOSE

This text is written for a first course on statistics and quantitative methods for Ph.D. students in social science and allied fields. Anyone undertaking to write such a book must sooner or later confront the question of whether the world really needs another introductory statistics textbook. In my surveys of the market for my own classes on this subject in two social science Ph.D. programs, I clearly decided that it does.

Students in social science Ph.D. programs outside of economics have widely divergent levels of previous exposure to statistical methods, as well as comfort with mathematical expression of concepts. The typical Ph.D. program does not have the luxury of multiple "tracks" to suit different backgrounds, so one course must accommodate all of them. That course must be accessible to students with divergent levels of preparation but must also prepare them technically for future quantitative methods coursework ahead of them.

More important, I have found that students of whatever background will plunge relatively enthusiastically into methods training once they understand why it is essential for the purely substantive elements of their research. Simply put, many students, particularly those without much prior exposure to statistics, do not understand what it is or how it can help them as social scientists. Without this understanding they lack the buy-in necessary to make the technical rigors of the course seem worthwhile.

I never found a textbook that was pitched at the right technical level for introductory students and also focused on what social scientists in particular can do with statistics. Textbooks on mathematical statistics, although useful for the technically ambitious students (and entirely suitable for use in conjunction with this text), leave most students behind at a formal level. They also tend to focus, at least in their core, on univariate inference. This framework allows for

clarity of statistical theory – but it obscures what social scientists can do with statistics and leaves many students wondering why they should bother when they have social science research to do. Statistics textbooks at an intermediate level may be technically suitable for most Ph.D. students in social science but are not much better at explicating the integrative link between social science theory and statistics. These problems with the textbook market are widely enough recognized that instructors often assemble reading lists culled from various textbooks and research papers published in home-discipline journals, but this approach is inherently piecemeal and obscures the overall integration of the material from the students.

This text is my attempt to redress these limitations of current offerings. Mathematically, I assume students have had exposure to and a solid grasp of algebra for scalars. Prior exposure to calculus is helpful (especially for some proofs) but not strictly necessary; when I use calculus notation and concepts in the main text, I explain them on the fly. I essentially do not use matrix algebra in this text except for occasional use of vector notation as a shorthand. Other than this, the text is mostly self-contained, conceptually and mathematically. That is not to say it will be technically easy for most students; it is not for most of mine. But prior formal training in higher-level mathematics is not the limiting factor. What students really need to be able to do is integrate formal definitions and notations into their thinking right away when they are introduced.

This in itself is not a substitute for deep training of future methodologists or theoreticians of statistical modeling but should suffice for students who need solid intuition about the tools they are using more than anything else. The text does not skimp on formal proofs and derivations of core results, for example, of the law of iterated expectations, the expected value of ordinary least squares regression coefficients, canonical formulas for standard errors, and the like. My conviction is that any intelligent user of statistical methods should understand the connection between assumptions they make about their data and inference problem and the properties of the resulting statistical output; formal proofs are simply the arguments that we use to establish those connections. To keep the length and formal demands manageable, I also often present an important theorem or approach verbally (e.g., the Cramér-Rao theorem, derivation of the F distribution, properties of parametric bootstrapped standard errors) to alert students to the key issue and indicate important topics for work in more advanced quantitative methods courses.

This approach implies that mathematical demands of the presentation are not uniform across chapters and topics. For example, the formal presentation of the sampling distribution for least squares regression coefficients (Chapter 7) is much more intense than my presentation of the concept of maximum likelihood estimation (Chapter 9), which is rather more intuitive. Because the purpose of the book is to develop solid understanding of core ideas in statistical inference and modeling rather than to present a formally complete and precise treatment of every topic for its own sake, this seems to be a useful compromise. In