Data Analysis Using Regression and Multilevel/Hierarchical Models

Data Analysis Using Regression and Multilevel/Hierarchical Models is a comprehensive manual for the applied researcher who wants to perform data analysis using linear and nonlinear regression and multilevel models. The book introduces and demonstrates a wide variety of models, at the same time instructing the reader in how to fit these models using freely available software packages. The book illustrates the concepts by working through scores of real data examples that have arisen in the authors' own applied research, with programming code provided for each one. Topics covered include causal inference, including regression, poststratification, matching, regression discontinuity, and instrumental variables, as well as multilevel logistic regression and missing-data imputation. Practical tips regarding building, fitting, and understanding are provided throughout.

Andrew Gelman is Professor of Statistics and Professor of Political Science at Columbia University. He has published more than 150 articles in statistical theory, methods, and computation and in applications areas including decision analysis, survey sampling, political science, public health, and policy. His other books are *Bayesian Data Analysis* (1995, second edition 2003) and *Teaching Statistics: A Bag of Tricks* (2002).

Jennifer Hill is Assistant Professor of Public Affairs in the Department of International and Public Affairs at Columbia University. She has coauthored articles that have appeared in the *Journal of the American Statistical Association*, *American Political Science Review*, *American Journal of Public Health*, *Developmental Psychology*, the *Economic Journal*, and the *Journal of Policy Analysis and Management*, among others.

Analytical Methods for Social Research

Analytical Methods for Social Research presents texts on empirical and formal methods for the social sciences. Volumes in the series address both the theoretical underpinnings of analytical techniques and their application in social research. Some series volumes are broad in scope, cutting across a number of disciplines. Others focus mainly on methodological applications within specific fields such as political science, sociology, demography, and public health. The series serves a mix of students and researchers in the social sciences and statistics.

Other Titles in the Series:

Event History Modeling: A Guide for Social Scientists, by Janet M. Box-Steffensmeier and Bradford S. Jones
Ecological Inference: New Methodological Strategies, edited by Gary King, Ori Rosen, and Martin A. Tanner
Spatial Models of Parliamentary Voting, by Keith T. Poole
Essential Mathematics for Political and Social Research, by Jeff Gill
Political Game Theory: An Introduction, by Nolan McCarty and Adam Meirowitz

Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN
Columbia University

JENNIFER HILL
Columbia University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521686891

First published 2007

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Gelman, Andrew.
Data analysis using regression and multilevel/hierarchical models / Andrew Gelman,
Jennifer Hill.
p. cm. – (Analytical methods for social research)
Includes bibliographical references.
ISBN 0-521-86706-1 (hardcover) – ISBN 0-521-68689-X (pbk.)
1. Regression analysis. 2. Multilevel models (Statistics) I. Hill, Jennifer, 1969–
II. Title. III. Series.
HA31.3.G45 2006
519.536–dc22 2006040566

ISBN 978-0-521-86706-1 Hardback
ISBN 978-0-521-68689-1 Paperback

For Zacky and for Audrey

Cambridge University Press & Assessment
978-0-521-68689-1 — Data Analysis Using Regression and Multilevel/Hierarchical Models
Andrew Gelman , Jennifer Hill
Frontmatter
More Information

Contents

Cambridge University Press & Assessment
978-0-521-68689-1 — Data Analysis Using Regression and Multilevel/Hierarchical Models
Andrew Gelman , Jennifer Hill
Frontmatter
More Information

List of examples

Preface

Aim of this book

This book originated as lecture notes for a course in regression and multilevel modeling, offered by the statistics department at Columbia University and attended by graduate students and postdoctoral researchers in social sciences (political science, economics, psychology, education, business, social work, and public health) and statistics. The prerequisite is statistics up to and including an introduction to multiple regression.

Advanced mathematics is not assumed—it is important to understand the linear model in regression, but it is not necessary to follow the matrix algebra in the derivation of least squares computations. It is useful to be familiar with exponents and logarithms, especially when working with generalized linear models.

After completing Part 1 of this book, you should be able to fit classical linear and generalized linear regression models—and do more with these models than simply look at their coefficients and their statistical significance. Applied goals include causal inference, prediction, comparison, and data description. After completing Part 2, you should be able to fit regression models for multilevel data. Part 3 takes you from data collection, through model understanding (looking at a table of estimated coefficients is usually not enough), to model checking and missing data. The appendixes include some reference materials on key tips, statistical graphics, and software for model fitting.

What you should be able to do after reading this book and working through the examples

This text is structured through models and examples, with the intention that after each chapter you should have certain skills in fitting, understanding, and displaying models:

- *Part 1A:* Fit, understand, and graph classical regressions and generalized linear models.
  - *Chapter 3:* Fit linear regressions and be able to interpret and display estimated coefficients.
  - *Chapter 4:* Build linear regression models by transforming and combining variables.
  - *Chapter 5:* Fit, understand, and display logistic regression models for binary data.
  - *Chapter 6:* Fit, understand, and display generalized linear models, including Poisson regression with overdispersion and ordered logit and probit models.
- *Part 1B:* Use regression to learn about quantities of substantive interest (not just regression coefficients).
  - *Chapter 7:* Simulate probability models and uncertainty about inferences and predictions.

  - *Chapter 8:* Check model fits using fake-data simulation and predictive simulation.
  - *Chapter 9:* Understand assumptions underlying causal inference. Set up regressions for causal inference and understand the challenges that arise.
  - *Chapter 10:* Understand the assumptions underlying propensity score matching, instrumental variables, and other techniques to perform causal inference when simple regression is not enough. Be able to use these when appropriate.

- *Part 2A:* Understand and graph multilevel models.
  - *Chapter 11:* Understand multilevel data structures and models as generalizations of classical regression.
  - *Chapter 12:* Understand and graph simple varying-intercept regressions and interpret as partial-pooling estimates.
  - *Chapter 13:* Understand and graph multilevel linear models with varying intercepts and slopes, non-nested structures, and other complications.
  - *Chapter 14:* Understand and graph multilevel logistic models.
  - *Chapter 15:* Understand and graph multilevel overdispersed Poisson, ordered logit and probit, and other generalized linear models.

- *Part 2B:* Fit multilevel models using the software packages R and Bugs.
  - *Chapter 16:* Fit varying-intercept regressions and understand the basics of Bugs. Check your programming using fake-data simulation.
  - *Chapter 17:* Use Bugs to fit various models from Part 2A.
  - *Chapter 18:* Understand Bayesian inference as a generalization of least squares and maximum likelihood. Use the Gibbs sampler to fit multilevel models.
  - *Chapter 19:* Use redundant parameterizations to speed the convergence of the Gibbs sampler.

- *Part 3:*
  - *Chapter 20:* Perform sample size and power calculations for classical and hierarchical models: standard-error formulas for basic calculations and fake-data simulation for harder problems.
  - *Chapter 21:* Calculate and understand contrasts, explained variance, partial pooling coefficients, and other summaries of fitted multilevel models.
  - *Chapter 22:* Use the ideas of analysis of variance to summarize fitted multilevel models; use multilevel models to perform analysis of variance.
  - *Chapter 23:* Use multilevel models in causal inference.
  - *Chapter 24:* Check the fit of models using predictive simulation.
  - *Chapter 25:* Use regression to impute missing data in multivariate datasets.

In summary, you should be able to fit, graph, and understand classical and multilevel linear and generalized linear models and to use these model fits to make predictions and inferences about quantities of interest, including causal treatment effects.

Data for the examples and homework assignments and other resources for teaching and learning

The website www.stat.columbia.edu/~gelman/arm/ contains datasets used in the examples and homework problems of the book, as well as sample computer code. The website also includes some tips for teaching regression and multilevel modeling through class participation rather than lecturing. We plan to update these tips based on feedback from instructors and students; please send your comments and suggestions to gelman@stat.columbia.edu.

Outline of a course

When teaching a course based on this book, we recommend starting with a self-contained review of linear regression, logistic regression, and generalized linear models, focusing not on the mathematics but on understanding these methods and implementing them in a reasonable way. This is also a convenient way to introduce the statistical language R, which we use throughout for modeling, computation, and graphics. One thing that will probably be new to the reader is the use of random simulations to summarize inferences and predictions.

We then introduce multilevel models in the simplest case of nested linear models, fitting in the Bayesian modeling language Bugs and examining the results in R. Key concepts covered at this point are partial pooling, variance components, prior distributions, identifiability, and the interpretation of regression coefficients at different levels of the hierarchy. We follow with non-nested models, multilevel logistic regression, and other multilevel generalized linear models.

Next we detail the steps of fitting models in Bugs and give practical tips for reparameterizing a model to make it converge faster and additional tips on debugging. We also present a brief review of Bayesian inference and computation. Once the student is able to fit multilevel models, we move in the final weeks of the class to the final part of the book, which covers more advanced issues in data collection, model understanding, and model checking.

As we show throughout, multilevel modeling fits into a view of statistics that unifies substantive modeling with accurate data fitting, and graphical methods are crucial both for seeing unanticipated features in the data and for understanding the implications of fitted models.

Acknowledgments

We thank the many students and colleagues who have helped us understand and implement these ideas. Most important have been Jouni Kerman, David Park, and Joe Bafumi for years of suggestions throughout this project, and for many insights into how to present this material to students.

In addition, we thank Hal Stern and Gary King for discussions on the structure of this book; Chuanhai Liu, Xiao-Li Meng, Zaiying Huang, John Boscardin, Jouni Kerman, Alan Zaslavsky, David Dunson, Maria Grazia Pittau, Aleks Jakulin, and Yu-Sung Su for discussions about multilevel modeling and statistical computation; Iven Van Mechelen and Hans Berkhof for discussions about model checking; Iain Pardoe for discussions of average predictive effects and other summaries of regression models; Matt Salganik and Wendy McKelvey for suggestions on the presentation of sample size calculations; T. E. Raghunathan, Donald Rubin, Rajeev Dehejia, Michael Sobel, Guido Imbens, Samantha Cook, Ben Hansen, Dylan Small, and Ed Vytlacil for concepts of missing-data modeling and causal inference; Eric

Loken for help in understanding identifiability in item-response models; Niall Bolger, Agustin Calatroni, John Carlin, Rafael Guerrero-Preston, Oliver Kuss, Reid Landes, Eduardo Leoni, and Dan Rabinowitz for code in Stata, SAS, and SPSS; Hans Skaug for code in AD Model Builder; Uwe Ligges, Sibylle Sturtz, Douglas Bates, Peter Dalgaard, Martyn Plummer, and Ravi Varadhan for help with multilevel modeling and general advice on R; and the students in Statistics / Political Science 4330 at Columbia for their invaluable feedback throughout.

Collaborators on specific examples mentioned in this book include Phillip Price on the home radon study; Tom Little, David Park, Joe Bafumi, and Noah Kaplan on the models of opinion polls and political ideal points; Jane Waldfogel, Jeanne Brooks-Gunn, and Wen Han for the mothers and children's intelligence data; Lex van Geen and Alex Pfaff on the arsenic in Bangladesh; Gary King on election forecasting; Jeffrey Fagan and Alex Kiss on the study of police stops; Tian Zheng and Matt Salganik on the social network analysis; John Carlin for the data on mesquite bushes and the adolescent-smoking study; Alessandra Casella and Tom Palfrey for the storable-votes study; Rahul Dodhia for the flight simulator example; Boris Shor, Joe Bafumi, and David Park on the voting and income study; Alan Edelman for the internet connections data; Donald Rubin for the Electric Company and educational-testing examples; Jeanne Brooks-Gunn and Jane Waldfogel for the mother and child IQ scores example and Infant Health and Development Program data; Nabila El-Bassel for the risky behavior data; Lenna Nepomnyaschy for the child support example; Howard Wainer with the Advanced Placement study; Iain Pardoe for the prison-sentencing example; James Liebman, Jeffrey Fagan, Valerie West, and Yves Chretien for the death-penalty study; Marcia Meyers, Julien Teitler, Irv Garfinkel, Marilyn Sinkowicz, and Sandra Garcia with the Social Indicators Study; Wendy McKelvey for the cockroach and rodent examples; Stephen Arpadi for the zinc and HIV study; Eric Verhoogen and Jan von der Goltz for the Progresa data; and Iven van Mechelen, Yuri Goegebeur, and Francis Tuerlincx on the stochastic learning models. These applied projects motivated many of the methodological ideas presented here, for example the display and interpretation of varying-intercept, varying-slope models from the analysis of income and voting (see Section 14.2), the constraints in the model of senators' ideal points (see Section 14.3), and the difficulties with two-level interactions as revealed by the radon study (see Section 21.7). Much of the work in Section 5.7 and Chapter 21 on summarizing regression models was done in collaboration with Iain Pardoe.

Many errors were found and improvements suggested by Brad Carlin, John Carlin, Samantha Cook, Caroline Rosenthal Gelman, Kosuke Imai, Jonathan Katz, Uwe Ligges, Wendy McKelvey, Jong-Hee Park, Martyn Plummer, Phillip Price, Song Qian, Giuseppe Ragusa, Dylan Small, Elizabeth Stuart, Sibylle Sturtz, Alex Tabarrok, and Shravan Vasishth. Brian MacDonald's copyediting has saved us from much embarrassment, and we also thank Yu-Sung Su for typesetting help, Sarah Ryu for assistance with indexing, and Ed Parsons and his colleagues at Cambridge University Press for their help in putting this book together. We especially thank Bob O'Hara and Gregor Gorjanc for incredibly detailed and useful comments on the nearly completed manuscript.

We also thank the developers of free software, especially R (for statistical computation and graphics) and Bugs (for Bayesian modeling), and also Emacs and LaTex (used in the writing of this book). We thank Columbia University for its collaborative environment for research and teaching, and the U.S. National Science Foundation for financial support. Above all, we thank our families for their love and support during the writing of this book.