

Cambridge University Press

978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill

Frontmatter

[More information](#)

## 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.

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)

Cambridge University Press

978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill

Frontmatter

[More information](#)

## 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.

### *Series Editors:*

R. Michael Alvarez, *California Institute of Technology*

Nathaniel L. Beck, *New York University*

Lawrence L. Wu, *New York University*

### *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

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)

Cambridge University Press

978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill

Frontmatter

[More information](#)

# Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN

*Columbia University*

JENNIFER HILL

*Columbia University*



CAMBRIDGE  
UNIVERSITY PRESS

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)

CAMBRIDGE  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom  
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](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9780521686891](http://www.cambridge.org/9780521686891)

© Andrew Gelman and Jennifer Hill 2007

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

First published 2007  
Reprinted with corrections 2007  
13th printing 2015

Printed in the United States of America by Sheridan Books, Inc.

*A catalog 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). 1. Hill, Jennifer, 1969–  
II. Title. III. Series.  
HA31.3.G45 2006  
519.5'36–dc22 2006040566  
  
ISBN 978-0-521-86706-1 hardback  
ISBN 978-0-521-68689-1 paperback

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.

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)

For Zacky and for Audrey

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)



Contents

List of examples	page xvii
Preface	xix
1 Why?	1
1.1 What is multilevel regression modeling?	1
1.2 Some examples from our own research	3
1.3 Motivations for multilevel modeling	6
1.4 Distinctive features of this book	8
1.5 Computing	9
2 Concepts and methods from basic probability and statistics	13
2.1 Probability distributions	13
2.2 Statistical inference	16
2.3 Classical confidence intervals	18
2.4 Classical hypothesis testing	20
2.5 Problems with statistical significance	22
2.6 55,000 residents desperately need your help!	23
2.7 Bibliographic note	26
2.8 Exercises	26
Part 1A: Single-level regression	29
3 Linear regression: the basics	31
3.1 One predictor	31
3.2 Multiple predictors	32
3.3 Interactions	34
3.4 Statistical inference	37
3.5 Graphical displays of data and fitted model	42
3.6 Assumptions and diagnostics	45
3.7 Prediction and validation	47
3.8 Bibliographic note	49
3.9 Exercises	49
4 Linear regression: before and after fitting the model	53
4.1 Linear transformations	53
4.2 Centering and standardizing, especially for models with interactions	55
4.3 Correlation and “regression to the mean”	57
4.4 Logarithmic transformations	59
4.5 Other transformations	65
4.6 Building regression models for prediction	68
4.7 Fitting a series of regressions	73

x	CONTENTS
4.8	Bibliographic note 74
4.9	Exercises 74
<b>5</b>	<b>Logistic regression 79</b>
5.1	Logistic regression with a single predictor 79
5.2	Interpreting the logistic regression coefficients 81
5.3	Latent-data formulation 85
5.4	Building a logistic regression model: wells in Bangladesh 86
5.5	Logistic regression with interactions 92
5.6	Evaluating, checking, and comparing fitted logistic regressions 97
5.7	Average predictive comparisons on the probability scale 101
5.8	Identifiability and separation 104
5.9	Bibliographic note 105
5.10	Exercises 105
<b>6</b>	<b>Generalized linear models 109</b>
6.1	Introduction 109
6.2	Poisson regression, exposure, and overdispersion 110
6.3	Logistic-binomial model 116
6.4	Probit regression: normally distributed latent data 118
6.5	Ordered and unordered categorical regression 119
6.6	Robust regression using the <i>t</i> model 124
6.7	Building more complex generalized linear models 125
6.8	Constructive choice models 127
6.9	Bibliographic note 131
6.10	Exercises 132
<b>Part 1B:</b>	<b>Working with regression inferences 135</b>
<b>7</b>	<b>Simulation of probability models and statistical inferences 137</b>
7.1	Simulation of probability models 137
7.2	Summarizing linear regressions using simulation: an informal Bayesian approach 140
7.3	Simulation for nonlinear predictions: congressional elections 144
7.4	Predictive simulation for generalized linear models 148
7.5	Bibliographic note 151
7.6	Exercises 152
<b>8</b>	<b>Simulation for checking statistical procedures and model fits 155</b>
8.1	Fake-data simulation 155
8.2	Example: using fake-data simulation to understand residual plots 157
8.3	Simulating from the fitted model and comparing to actual data 158
8.4	Using predictive simulation to check the fit of a time-series model 163
8.5	Bibliographic note 165
8.6	Exercises 165
<b>9</b>	<b>Causal inference using regression on the treatment variable 167</b>
9.1	Causal inference and predictive comparisons 167
9.2	The fundamental problem of causal inference 170
9.3	Randomized experiments 172
9.4	Treatment interactions and poststratification 178

CONTENTS	xi
9.5 Observational studies	181
9.6 Understanding causal inference in observational studies	186
9.7 Do not control for post-treatment variables	188
9.8 Intermediate outcomes and causal paths	190
9.9 Bibliographic note	194
9.10 Exercises	194
<b>10 Causal inference using more advanced models</b>	<b>199</b>
10.1 Imbalance and lack of complete overlap	199
10.2 Subclassification: effects and estimates for different subpopulations	204
10.3 Matching: subsetting the data to get overlapping and balanced treatment and control groups	206
10.4 Lack of overlap when the assignment mechanism is known: regression discontinuity	212
10.5 Estimating causal effects indirectly using instrumental variables	215
10.6 Instrumental variables in a regression framework	220
10.7 Identification strategies that make use of variation within or between groups	226
10.8 Bibliographic note	229
10.9 Exercises	231
<b>Part 2A: Multilevel regression</b>	<b>235</b>
<b>11 Multilevel structures</b>	<b>237</b>
11.1 Varying-intercept and varying-slope models	237
11.2 Clustered data: child support enforcement in cities	237
11.3 Repeated measurements, time-series cross sections, and other non-nested structures	241
11.4 Indicator variables and fixed or random effects	244
11.5 Costs and benefits of multilevel modeling	246
11.6 Bibliographic note	247
11.7 Exercises	248
<b>12 Multilevel linear models: the basics</b>	<b>251</b>
12.1 Notation	251
12.2 Partial pooling with no predictors	252
12.3 Partial pooling with predictors	254
12.4 Quickly fitting multilevel models in R	259
12.5 Five ways to write the same model	262
12.6 Group-level predictors	265
12.7 Model building and statistical significance	270
12.8 Predictions for new observations and new groups	272
12.9 How many groups and how many observations per group are needed to fit a multilevel model?	275
12.10 Bibliographic note	276
12.11 Exercises	277
<b>13 Multilevel linear models: varying slopes, non-nested models, and other complexities</b>	<b>279</b>
13.1 Varying intercepts and slopes	279
13.2 Varying slopes without varying intercepts	283

xii	CONTENTS
13.3	Modeling multiple varying coefficients using the scaled inverse-Wishart distribution 284
13.4	Understanding correlations between group-level intercepts and slopes 287
13.5	Non-nested models 289
13.6	Selecting, transforming, and combining regression inputs 293
13.7	More complex multilevel models 297
13.8	Bibliographic note 297
13.9	Exercises 298
14	<b>Multilevel logistic regression 301</b>
14.1	State-level opinions from national polls 301
14.2	Red states and blue states: what's the matter with Connecticut? 310
14.3	Item-response and ideal-point models 314
14.4	Non-nested overdispersed model for death sentence reversals 320
14.5	Bibliographic note 321
14.6	Exercises 322
15	<b>Multilevel generalized linear models 325</b>
15.1	Overdispersed Poisson regression: police stops and ethnicity 325
15.2	Ordered categorical regression: storable votes 331
15.3	Non-nested negative-binomial model of structure in social networks 332
15.4	Bibliographic note 342
15.5	Exercises 342
Part 2B	<b>Fitting multilevel models 343</b>
16	<b>Multilevel modeling in Bugs and R: the basics 345</b>
16.1	Why you should learn Bugs 345
16.2	Bayesian inference and prior distributions 345
16.3	Fitting and understanding a varying-intercept multilevel model using R and Bugs 348
16.4	Step by step through a Bugs model, as called from R 353
16.5	Adding individual- and group-level predictors 359
16.6	Predictions for new observations and new groups 361
16.7	Fake-data simulation 363
16.8	The principles of modeling in Bugs 366
16.9	Practical issues of implementation 369
16.10	Open-ended modeling in Bugs 370
16.11	Bibliographic note 373
16.12	Exercises 373
17	<b>Fitting multilevel linear and generalized linear models in Bugs and R 375</b>
17.1	Varying-intercept, varying-slope models 375
17.2	Varying intercepts and slopes with group-level predictors 379
17.3	Non-nested models 380
17.4	Multilevel logistic regression 381
17.5	Multilevel Poisson regression 382
17.6	Multilevel ordered categorical regression 383
17.7	Latent-data parameterizations of generalized linear models 384

CONTENTS	xiii
17.8 Bibliographic note	385
17.9 Exercises	385
<b>18 Likelihood and Bayesian inference and computation</b>	<b>387</b>
18.1 Least squares and maximum likelihood estimation	387
18.2 Uncertainty estimates using the likelihood surface	390
18.3 Bayesian inference for classical and multilevel regression	392
18.4 Gibbs sampler for multilevel linear models	397
18.5 Likelihood inference, Bayesian inference, and the Gibbs sampler: the case of censored data	402
18.6 Metropolis algorithm for more general Bayesian computation	408
18.7 Specifying a log posterior density, Gibbs sampler, and Metropolis algorithm in R	409
18.8 Bibliographic note	413
18.9 Exercises	413
<b>19 Debugging and speeding convergence</b>	<b>415</b>
19.1 Debugging and confidence building	415
19.2 General methods for reducing computational requirements	418
19.3 Simple linear transformations	419
19.4 Redundant parameters and intentionally nonidentifiable models	419
19.5 Parameter expansion: multiplicative redundant parameters	424
19.6 Using redundant parameters to create an informative prior distribution for multilevel variance parameters	427
19.7 Bibliographic note	434
19.8 Exercises	434
<b>Part 3: From data collection to model understanding to model checking</b>	<b>435</b>
<b>20 Sample size and power calculations</b>	<b>437</b>
20.1 Choices in the design of data collection	437
20.2 Classical power calculations: general principles, as illustrated by estimates of proportions	439
20.3 Classical power calculations for continuous outcomes	443
20.4 Multilevel power calculation for cluster sampling	447
20.5 Multilevel power calculation using fake-data simulation	449
20.6 Bibliographic note	454
20.7 Exercises	454
<b>21 Understanding and summarizing the fitted models</b>	<b>457</b>
21.1 Uncertainty and variability	457
21.2 Superpopulation and finite-population variances	459
21.3 Contrasts and comparisons of multilevel coefficients	462
21.4 Average predictive comparisons	466
21.5 $R^2$ and explained variance	473
21.6 Summarizing the amount of partial pooling	477
21.7 Adding a predictor can <i>increase</i> the residual variance!	480
21.8 Multiple comparisons and statistical significance	481
21.9 Bibliographic note	484
21.10 Exercises	485

xiv	CONTENTS
<b>22 Analysis of variance</b>	<b>487</b>
22.1 Classical analysis of variance	487
22.2 ANOVA and multilevel linear and generalized linear models	490
22.3 Summarizing multilevel models using ANOVA	492
22.4 Doing ANOVA using multilevel models	494
22.5 Adding predictors: analysis of covariance and contrast analysis	496
22.6 Modeling the variance parameters: a split-plot latin square	498
22.7 Bibliographic note	501
22.8 Exercises	501
<b>23 Causal inference using multilevel models</b>	<b>503</b>
23.1 Multilevel aspects of data collection	503
23.2 Estimating treatment effects in a multilevel observational study	506
23.3 Treatments applied at different levels	507
23.4 Instrumental variables and multilevel modeling	509
23.5 Bibliographic note	512
23.6 Exercises	512
<b>24 Model checking and comparison</b>	<b>513</b>
24.1 Principles of predictive checking	513
24.2 Example: a behavioral learning experiment	515
24.3 Model comparison and deviance	524
24.4 Bibliographic note	526
24.5 Exercises	527
<b>25 Missing-data imputation</b>	<b>529</b>
25.1 Missing-data mechanisms	530
25.2 Missing-data methods that discard data	531
25.3 Simple missing-data approaches that retain all the data	532
25.4 Random imputation of a single variable	533
25.5 Imputation of several missing variables	539
25.6 Model-based imputation	540
25.7 Combining inferences from multiple imputations	542
25.8 Bibliographic note	542
25.9 Exercises	543
<b>Appendixes</b>	<b>545</b>
<b>A Six quick tips to improve your regression modeling</b>	<b>547</b>
A.1 Fit many models	547
A.2 Do a little work to make your computations faster and more reliable	547
A.3 Graphing the relevant and not the irrelevant	548
A.4 Transformations	548
A.5 Consider all coefficients as potentially varying	549
A.6 Estimate causal inferences in a targeted way, not as a byproduct of a large regression	549
<b>B Statistical graphics for research and presentation</b>	<b>551</b>
B.1 Reformulating a graph by focusing on comparisons	552
B.2 Scatterplots	553
B.3 Miscellaneous tips	559

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)

CONTENTS	xv
B.4 Bibliographic note	562
B.5 Exercises	563
<b>C Software</b>	<b>565</b>
C.1 Getting started with R, Bugs, and a text editor	565
C.2 Fitting classical and multilevel regressions in R	565
C.3 Fitting models in Bugs and R	567
C.4 Fitting multilevel models using R, Stata, SAS, and other software	568
C.5 Bibliographic note	573
<b>References</b>	<b>575</b>
<b>Author index</b>	<b>601</b>
<b>Subject index</b>	<b>607</b>

Cambridge University Press  
978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models  
Andrew Gelman and Jennifer Hill  
Frontmatter  
[More information](#)



---

# List of examples

---

Home radon	3, 36, 252, 279, 479
Forecasting elections	3, 144
State-level opinions from national polls	4, 301, 493
Police stops by ethnic group	5, 21, 112, 325
Public opinion on the death penalty	19
Testing for election fraud	23
Sex ratio of births	27, 137
Mothers' education and children's test scores	31, 55
Height and weight	41, 75
Beauty and teaching evaluations	51, 277
Height and earnings	53, 59, 140, 288
Handedness	66
Yields of mesquite bushes	70
Political party identification over time	73
Income and voting	79, 107
Arsenic in drinking water	86, 128, 193
Death-sentencing appeals process	116, 320, 540
Ordered logistic model for storable votes	120, 331
Cockroaches in apartments	126, 161
Behavior of couples at risk for HIV	132, 166
Academy Award voting	133
Incremental cost-effectiveness ratio	152
Unemployment time series	163
The Electric Company TV show	174, 503
Hypothetical study of parenting quality as an intermediate outcome	188
Sesame Street TV show	196
Messy randomized experiment of cow feed	196
Incumbency and congressional elections	197

xviii	LIST OF EXAMPLES
Value of a statistical life	197
Evaluating the Infant Health and Development Program	201, 506
Ideology of congressmembers	213
Hypothetical randomized-encouragement study	216
Child support enforcement	237
Adolescent smoking	241
Rodents in apartments	248
Olympic judging	248
Time series of children’s CD4 counts	249, 277, 449
Flight simulator experiment	289, 464, 488
Latin square agricultural experiment	292, 497
Income and voting by state	310
Item-response models	314
Ideal-point modeling for the Supreme Court	317
Speed dating	322
Social networks	332
Regression with censored data	402
Educational testing experiments	430
Zinc for HIV-positive children	439
Cluster sampling of New York City residents	448
Value added of school teachers	458
Advanced Placement scores and college grades	463
Prison sentences	470
Magnetic fields and brain functioning	481
Analysis of variance for web connect times	492
Split-plot latin square	498
Educational-subsidy program in Mexican villages	508
Checking models of behavioral learning in dogs	515
Missing data in the Social Indicators Survey	529

---

# 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.

Cambridge University Press

978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill

Frontmatter

[More information](#)

xx

## PREFACE

- *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.

*Da a f h e e a e a d h e a i g e a d h e e c e f  
e a c h i g a d e a i g*

The website [www.stat.columbia.edu/~gelman/arm/](http://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](mailto:gelman@stat.columbia.edu).

*O i e f a c e*

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.

*A c k n o w l e d g e m e n t s*

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 Vytlačil for concepts of missing-data modeling and causal inference; Eric

Cambridge University Press

978-0-521-68689-1 - Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill

Frontmatter

[More information](#)

xxii

PREFACE

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 multi-level 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 Arpad for the zinc and HIV study; Eric Verhoogen and Jan von der Goltz for the Progres data; and Iven van Mechelen, Yuri Goegebeur, and Francis Tuerlinx 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.