Regression and Other Stories

Many textbooks on regression focus on theory and the simplest of examples. Real statistical problems, however, are complex and subtle. This is not a book about the theory of regression. It is a book about how to use regression to solve real problems of comparison, estimation, prediction, and causal inference. It focuses on practical issues such as sample size and missing data and a wide range of goals and techniques. It jumps right in to methods and computer code you can use fresh out of the box.

Key features:

- Real examples, real stories from the authors' real-world experience demonstrate what can be achieved by regression and what the limitations are
- Uses computation with the popular open-source programs R and Stan instead of deriving formulas, with all code available online
- Emphasis on using graphics and presentation to understand and check models that have been fit to data
- Practical advice for understanding assumptions and implementing methods for experiments and observational studies
- Smooth transition to logistic regression and generalized linear models
- Clear presentation of key ideas in data collection, sampling, generalization, and causal inference

The authors are experienced researchers who have published articles in hundreds of different scientific journals in fields including statistics, computer science, policy, public health, political science, economics, sociology, and engineering. They have also published articles in the *Washington Post, New York Times, Slate*, and other public venues. Their previous books include *Bayesian Data Analysis, Teaching Statistics: A Bag of Tricks*, and *Data Analysis and Regression Using Multilevel/Hierarchical Models*.

ANDREW GELMAN is Higgins Professor of Statistics and Professor of Political Science at Columbia University.

JENNIFER HILL is Professor of Applied Statistics at New York University.

AKI VEHTARI is Associate Professor in Computational Probabilistic Modeling at Aalto University.

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 as well as 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*
Stephen L. Morgan, *Johns Hopkins University*
Lawrence L. Wu, *New York University*

Other Titles in the Series

Maximum Likelihood for Social Science, by Michael D. Ward and John S. Ahlquist
Computational Social Science, by R. Michael Alvarez
Spatial Analysis for the Social Sciences, by David Darmofal
Counterfactuals and Causal Inference, Second Edition, by Stephen L. Morgan and Christopher Winship
Time Series Analysis for the Social Sciences, by Janet M. Box-Steffensmeier, John R. Freeman, Matthew P. Hitt, and Jon C. W. Pevehouse
Statistical Modeling and Inference for Social Science, by Sean Gailmard
Formal Models of Domestic Politics, by Scott Gehlbach
Data Analysis Using Regression and Multilevel/Hierarchical Models, by Andrew Gelman and Jennifer Hill
Political Game Theory: An Introduction, by Nolan McCarty and Adam Meirowitz
Essential Mathematics for Political and Social Research, by Jeff Gill
Spatial Models of Parliamentary Voting, by Keith T. Poole
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

Regression and Other Stories

ANDREW GELMAN
Columbia University, New York

JENNIFER HILL
New York University

AKI VEHTARI
Aalto University, Finland

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, 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

79 Anson Road, #06–04/06, Singapore 079906

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/9781107023987
DOI:10.1017/9781139161879

First published 2021
Reprinted 2021

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

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

ISBN 978-1-107-02398-7 Hardback
ISBN 978-1-107-67651-0 Paperback

Contents

Preface

Existing textbooks on regression typically have some mix of cookbook instruction and mathematical derivation. We wrote this book because we saw a new way forward, focusing on understanding regression models, applying them to real problems, and using simulations with fake data to understand how the models are fit. After reading this book and working through the exercises, you should be able to simulate regression models on the computer and build, critically evaluate, and use them for applied problems.

The other special feature of our book, in addition to its wide range of examples and its focus on computer simulation, is its broad coverage, including the basics of statistics and measurement, linear regression, multiple regression, Bayesian inference, logistic regression and generalized linear models, extrapolation from sample to population, and causal inference. Linear regression is the starting point, but it does not make sense to stop there: once you have the basic idea of statistical prediction, it can be best understood by applying it in many different ways and in many different contexts.

After completing Part 1 of this book, you should have access to the tools of mathematics, statistics, and computing that will allow you to work with regression models. These early chapters should serve as a bridge from the methods and ideas you may have learned in an introductory statistics course. Goals for Part 1 include displaying and exploring data, computing and graphing linear relations, understanding basic probability distributions and statistical inferences, and simulation of random processes to represent inferential and forecast uncertainty.

After completing Part 2, you should be able to build, fit, understand, use, and assess the fit of linear regression models. The chapters in this part of the book develop relevant statistical and computational tools in the context of several applied and simulated-data examples. After completing Part 3, you should be able to similarly work with logistic regression and other generalized linear models. Part 4 covers data collection and extrapolation from sample to population, and in Part 5 we cover causal inference, starting with basic methods using regression for controlled experiments and then considering more complicated approaches adjusting for imbalances in observational data or capitalizing on natural experiments. Part 6 introduces more advanced regression models, and the appendixes include some quick tips and an overview on software for model fitting.

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

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 1:* Review key tools and concepts in mathematics, statistics, and computing.
  - *Chapter 1:* Have a sense of the goals and challenges of regression.
  - *Chapter 2:* Explore data and be aware of issues of measurement and adjustment.
  - *Chapter 3:* Graph a straight line and know some basic mathematical tools and probability distributions.
  - *Chapter 4:* Understand statistical estimation and uncertainty assessment, along with the problems of hypothesis testing in applied statistics.
  - *Chapter 5:* Simulate probability models and uncertainty about inferences and predictions.

- *Part 2:* Build linear regression models, use them in real problems, and evaluate their assumptions and fit to data.
  - *Chapter 6:* Distinguish between descriptive and causal interpretations of regression, understanding these in historical context.
  - *Chapter 7:* Understand and work with simple linear regression with one predictor.
  - *Chapter 8:* Gain a conceptual understanding of least squares fitting and be able to perform these fits on the computer.
  - *Chapter 9:* Perform and understand probabilistic prediction and simple Bayesian information aggregation, and be introduced to prior distributions and Bayesian inference.
  - *Chapter 10:* Build, fit, and understand linear models with multiple predictors.
  - *Chapter 11:* Understand the relative importance of different assumptions of regression models and be able to check models and evaluate their fit to data.
  - *Chapter 12:* Apply linear regression more effectively by transforming and combining predictors.
- *Part 3:* Build and work with logistic regression and generalized linear models.
  - *Chapter 13:* Fit, understand, and display logistic regression models for binary data.
  - *Chapter 14:* Build, understand, and evaluate logistic regressions with interactions and other complexities.
  - *Chapter 15:* Fit, understand, and display generalized linear models, including the Poisson and negative binomial regression, ordered logistic regression, and other models.
- *Part 4:* Design studies and use data more effectively in applied settings.
  - *Chapter 16:* Use probability theory and simulation to guide data-collection decisions, without falling into the trap of demanding unrealistic levels of certainty.
  - *Chapter 17:* Use poststratification to generalize from sample to population, and use regression models to impute missing data.
- *Part 5:* Implement and understand basic statistical designs and analyses for causal inference.
  - *Chapter 18:* Understand assumptions underlying causal inference with a focus on randomized experiments.
  - *Chapter 19:* Perform causal inference in simple settings using regressions to estimate treatment effects and interactions.
  - *Chapter 20:* Understand the challenges of causal inference from observational data and statistical tools for adjusting for differences between treatment and control groups.
  - *Chapter 21:* Understand the assumptions underlying more advanced methods that use auxiliary variables or particular data structures to identify causal effects, and be able to fit these models to data.
- *Part 6:* Become aware of more advanced regression models.
  - *Chapter 22:* Get a sense of the directions in which linear and generalized linear models can be extended to attack various classes of applied problems.
- *Appendixes:*
  - *Appendix A:* Get started in the statistical software R, with a focus on data manipulation, statistical graphics, and fitting and using regressions.
  - *Appendix B:* Become aware of some important ideas in regression workflow.

After working through the book, you should be able to fit, graph, understand, and evaluate linear and generalized linear models and use these model fits to make predictions and inferences about quantities of interest, including causal effects of treatments and exposures.

Cambridge University Press
978-1-107-02398-7 — Regression and Other Stories
Andrew Gelman , Jennifer Hill , Aki Vehtari
Frontmatter
More Information

Fun chapter titles

The chapter titles in the book are descriptive. Here are more dramatic titles intended to evoke some of the surprise you should feel when working through this material:

- *Part 1:*
  - *Chapter 1:* Prediction as a unifying theme in statistics and causal inference.
  - *Chapter 2:* Data collection and visualization are important.
  - *Chapter 3:* Here's the math you actually need to know.
  - *Chapter 4:* Time to unlearn what you thought you knew about statistics.
  - *Chapter 5:* You don't understand your model until you can simulate from it.
- *Part 2:*
  - *Chapter 6:* Let's think deeply about regression.
  - *Chapter 7:* You can't just *do* regression, you have to *understand* regression.
  - *Chapter 8:* Least squares and all that.
  - *Chapter 9:* Let's be clear about our uncertainty and about our prior knowledge.
  - *Chapter 10:* You don't just *fit* models, you *build* models.
  - *Chapter 11:* Can you convince *me* to trust *your* model?
  - *Chapter 12:* Only fools work on the raw scale.
- *Part 3:*
  - *Chapter 13:* Modeling probabilities.
  - *Chapter 14:* Logistic regression pro tips.
  - *Chapter 15:* Building models from the inside out.
- *Part 4:*
  - *Chapter 16:* To understand the past, you must first know the future.
  - *Chapter 17:* Enough about your data. Tell me about the population.
- *Part 5:*
  - *Chapter 18:* How can flipping a coin help you estimate causal effects?
  - *Chapter 19:* Using correlation and assumptions to infer causation.
  - *Chapter 20:* Causal inference is just a kind of prediction.
  - *Chapter 21:* More assumptions, more problems.
- *Part 6:*
  - *Chapter 22:* Who's got next?
- *Appendixes:*
  - *Appendix A:* R quick start.
  - *Appendix B:* These are our favorite workflow tips; what are yours?

In this book we present many methods and illustrate their use in many applications; we also try to give a sense of where these methods can fail, and we try to convey the excitement the first time that we learned about these ideas and applied them to our own problems.

Additional material for teaching and learning

Data for the examples and homework assignments; other teaching resources

The website `www.stat.columbia.edu/~gelman/regression` contains pointers to data and code for the examples and homework problems in the book, along with some teaching materials.

Prerequisites

This book does not require advanced mathematics. To understand the linear model in regression, you will need the algebra of the intercept and slope of a straight line, but it will not be necessary to follow the matrix algebra in the derivation of least squares computations. You will use exponents and logarithms at different points, especially in Chapters 12–15 in the context of nonlinear transformations and generalized linear models.

Software

Previous knowledge of programming is not required. You will do a bit of programming in the general-purpose statistical environment R when fitting and using the models in this book, and some of these fits will be performed using the Bayesian inference program Stan, which, like R, is free and open source. Readers new to R or to programming should first work their way through Appendix A.

We fit regressions using the `stan_glm` function in the `rstanarm` package in R, performing Bayesian inference using simulation. This is a slight departure from usual treatments of regression (including our earlier book), which use least squares and maximum likelihood, for example using the `lm` and `glm` functions in R. We discuss differences between these different software options, and between these different modes of inference, in Sections 1.6, 8.4, and 9.5. From the user's perspective, switching to `stan_glm` doesn't matter much except in making it easier to obtain probabilistic predictions and to propagate inferential uncertainty, and in certain problems with collinearity or sparse data (in which case the Bayesian approach in `stan_glm` gives more stable estimates), and when we wish to include prior information in the analysis. For most of the computations done in this book, similar results could be obtained using classical regression software if so desired.

Suggested courses

The material in this book can be broken up in several ways for one-semester courses. Here are some examples:

- *Basic linear regression*: Chapters 1–5 for review, then Chapters 6–9 (linear regression with one predictor) and Chapters 10–12 (multiple regression, diagnostics, and model building).
- *Applied linear regression*: Chapters 1–5 for review, then Chapters 6–12 (linear regression), Chapters 16–17 (design and poststratification), and selected material from Chapters 18–21 (causal inference) and Chapter 22 (advanced regression).
- *Applied regression and causal inference*: Quick review of Chapters 1–5, then Chapters 6–12 (linear regression), Chapter 13 (logistic regression), Chapters 16–17 (design and poststratification), and selected material from Chapters 18–21 (causal inference).
- *Causal inference*: Chapters 1, 7, 10, 11, and 13 for review of linear and logistic regression, then Chapters 18–21 in detail.
- *Generalized linear models*: Some review of Chapters 1–12, then Chapters 13–15 (logistic regression and generalized linear models), followed by selected material from Chapters 16–21 (design, poststratification, and causal inference) and Chapter 22 (advanced regression).

Acknowledgments

We thank the many students and colleagues who have helped us understand and implement these ideas, including everyone thanked on pages xxi–xxii of our earlier book, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. In addition, we thank Pablo Argote, Bill Behrman, Ed Berry, Danilo Bzdok, Andres Castro, Devin Caughey, Zhengchen Cai, Zad Chow, Doug Davidson, Dick De Veaux, Vince Dorie, Mark Fisher, Matěj Grabovsky, Sander Greenland, Daphna Harel, Omri Har-Shemesh, Merlin Heidemanns, Christian Hennig, Nathan Hoffmann, David Kane, Yoav

Kessler, Katharine Khanna, Lydia Krasilnikova, A. Solomon Kurz, Stefano Longo, Gal Matijevic, Michael McLaren, Vicent Modesto, Eam O'Brien, Desislava Petkova, Jenny Pham, Eric Potash, Phil Price, Justin Reppert, Malgorzata Roos, Braden Scherting, Ravi Shroff, Noah Silbert, Michael Sobel, Melinda Song, Scott Spencer, Kenneth Tay, Mireia Triguero, Juliette Unwin, Jasu Vehtari, Jacob Warren, Zane Wolf, Lizzie Wolkovich, Adam Zelizer, Shuli Zhang, and students and teaching assistants from several years of our classes for helpful comments and suggestions, Alan Chen for help with Chapter 20, Andrea Cornejo, Zarni Htet, and Rui Lu for helping to develop the simulation-based exercises for the causal chapters, Ben Silver for help with indexing, Beth Morel and Clare Dennison for copy editing, Luke Keele for the example in Section 21.3, Kaiser Fung for the example in Section 21.5, Mark Broadie for the golf data in Exercise 22.3, Michael Betancourt for the gravity-measuring demonstration in Exercise 22.4, Jerry Reiter for sharing ideas on teaching and presentation of the concepts of regression, Lauren Cowles for many helpful suggestions on the structure of this book, and especially Ben Goodrich and Jonah Gabry for developing the `rstanarm` package which allows regression models to be fit in Stan using familiar R notation.

We also thank the developers of R and Stan, and the U.S. National Science Foundation, Institute for Education Sciences, Office of Naval Research, Defense Advanced Research Projects Agency, Google, Facebook, YouGov, and the Sloan Foundation for financial support.

Above all, we thank our families for their love and support during the writing of this book.