Active Statistics

This book provides statistics instructors and students with complete classroom material for a one- or two-semester course on applied regression and causal inference. It is built around 52 stories, 52 class-participation activities, 52 hands-on computer demonstrations, 52 drills, and 52 discussion problems that allow instructors and students to explore in a fun way the real-world complexity of the subject. The book fosters an engaging "flipped classroom" environment with a focus on visualization and understanding.

The book provides instructors with frameworks for self-study or for structuring the course, along with tips for maintaining student engagement at all levels and practice exam questions to help guide learning.

Designed to accompany the authors' previous textbook *Regression and Other Stories*, its modular nature and wealth of material allow this book to be adapted to different courses and texts or to be used by learners as a hands-on workbook.

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*, the *New York Times*, *Slate*, and other public venues. Their previous books include *Bayesian Data Analysis*, *Teaching Statistics: A Bag of Tricks*, and *Regression and Other Stories*.

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

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

Active Statistics

Stories, Games, Problems, and Hands-On Demonstrations for Applied Regression and Causal Inference

ANDREW GELMAN
Columbia University, New York

AKI VEHTARI
Aalto University, Finland

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/highereducation/ISBN/9781009436212

DOI: 10.1017/9781009436243

First published 2024

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

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-43621-2 Paperback

Contents

How to use this book **vii**

Part 1: Organizing a plan of study **1**

1 Active learning **3**
1.1 Flipped classroom and collaborative learning 3
1.2 What happens during the semester? 4
1.3 Active learning in class 6
1.4 Scheduling 8
1.5 Assessment and feedback 10
1.6 Some general issues in teaching and communication 11

2 Setting up a course of study **13**
2.1 What to learn and how to learn it 13
2.2 Computing 15
2.3 Course material 15
2.4 Real data and simulated data 17
2.5 Two kinds of computer demonstrations 17
2.6 Challenges in learning particular topics 18
2.7 Adapting to your goals and learning style 22
2.8 Using these materials in introductory or more advanced courses 23
2.9 Balance between challenges and solutions 27

Part 2: Stories, activities, problems, and demonstrations **29**

3 Week by week: the first semester **31**
3.1 Introduction to quantitative social science 31
3.2 Prediction as a unifying theme in statistics and causal inference 44
3.3 Data collection and visualization 54
3.4 Review of mathematics and probability 68
3.5 Statistical inference 76
3.6 Simulation 87
3.7 Background on regression modeling 97
3.8 Linear regression with a single predictor 105
3.9 Least squares and fitting regression models 114
3.10 Prediction and Bayesian inference 123
3.11 Linear regression with multiple predictors 133
3.12 Assumptions, diagnostics, and model evaluation 146
3.13 Regression with linear and log transformations 155

4 Week by week: the second semester **163**
4.14 Review of basic statistics and regression modeling 163
4.15 Logistic regression 174

4.16 Working with logistic regression 186
4.17 Other generalized linear models 199
4.18 Design and sample size decisions 214
4.19 Poststratification and missing-data imputation 226
4.20 How can flipping a coin help you estimate causal effects? 238
4.21 Causal inference using regression on the treatment variable 250
4.22 Causal inference as prediction 262
4.23 Imbalance and lack of complete overlap 275
4.24 Additional topics in causal inference 285
4.25 Advanced regression and multilevel models 300
4.26 Review of the course 311

Appendices **319**

A Pre-test questions **321**
A.1 First semester 321
A.2 Second semester 324

B Final exam questions **325**
B.1 Multiple-choice questions for the first semester 325
B.2 Multiple-choice questions for the second semester 340
B.3 Take-home exam 354

C Outlines of classroom activities **357**
C.1 First semester 358
C.2 Second semester 360

How to use this book

We have collected hundreds of stories, class-participation activities, computer demonstrations, and discussion problems for a semester-long or year-long statistics course on applied regression and causal inference, including readings, homework assignments, in-class activities, and exams. The goal is to have a course that is modern both in form (student-centered learning) and content (applied statistics with a computational edge). The material is set up in a modular way so that students and instructors can adapt to their own goals and interests.

The core of this book is Part 2, starting on page 31, with stories, activities, demonstrations, and problems for active learning of statistics. Part 1 of the book discusses how to use this material as part of a course or self-learning program, and appendixes include exam questions and an outline of the active learning tools in the book.

For students. You can use this book as a supplement to *Regression and Other Stories* or as part of a course on applied statistics. We go through every week of a two-semester class on applied regression and causal inference, and for each week we have homework assignments, stories, activities, computer demonstrations, drills, and discussion questions. You can read these on your own as the topics come up in the textbook and go through the computer demonstrations yourself.

The goal of this material for students is to connect statistical ideas and methods, especially involving regression and causal inference, to real-world applications. To this end, the stories, activities, demonstrations, and problems in this book are connected to each week's readings, which correspond to chapters in the textbook. If you are studying on your own or using another book, you should put in the effort to match the items to the topics.

For instructors. This book provides a ready-made two-semester course, and it can also be used as a source of classroom activities and a template for you to compile your own recipe book of stories, class-participation activities, computer demonstrations, and problems to facilitate active learning.

Our goal for instructors is to make it as easy as possible to teach statistical ideas and methods using real-world examples and active learning. An instructor can directly tell these stories in class, do these activities, and work through these live demonstrations; can adapt this material to the appropriate level and pace of the students; or can use this material as inspirations for developing completely new activities.

Adapting to your own needs

If you are using software other than R, adjust the demonstrations accordingly. Again, what is important is not to reproduce all the details but rather to get practice with simulating, analyzing, and plotting models and data.

The materials here are for a course in statistics with a focus on regression and causal inference. If you are a student or teacher of an introductory course or one with a different emphasis, you can adapt our activities and demonstrations accordingly.

This book might well have more material than you think you need. That's fine. We purposely created an overstuffed course with lots of possibilities for student involvement on each topic, to make it easy for you to dip in and use what you can. We encourage you to integrate active learning and real examples into every step of your statistical education.

Statistics is hard. It should not feel tricky.

Many of the stories and class-participation activities in this class have twists, and many of the problems have solutions that are not at first apparent. Indeed, we picked these examples because they are engaging and sometimes surprising. It makes sense to learn through stories—surprises in a narrative represent upending of expectations and are valuable for two reasons: first because they reveal problems with default assumptions, and second because they reveal these implicit assumptions in the first place. Assumptions and models are not bad things in quantitative reasoning; rather, they are a way to move forward in the presence of uncertainty and variation. And it is important to understand the models that we use.[1]

In giving fun stories and activities that feature surprises, we are *not* trying to send the message that statistics is tricky, always with one more pitfall around the corner; rather, we want the models and methods of statistics to feel more natural and intuitive in applied settings.

Online resources

Further material for this book is on the webpage for *Regression and Other Stories*,[2] including data and code for all the examples for both books and slides for material to be displayed during classroom activities.[3]

[1]For discussion of the connection between statistical thinking and surprise in storytelling, see Andrew Gelman and Thomas Basbøll (2014), When do stories work? Evidence and illustration in the social sciences, *Sociological Methods and Research* 43, 547–570.

[2]http://www.stat.columbia.edu/~gelman/regression/.

[3]We thank the U.S. Institute of Education Sciences and Office of Naval Research for partial support of this project and Jonah Gabry, Johannes Hallermeier, Sam Houskeeper, Manu Singh, Diana Lee, Merlin Heidemanns, Jennifer Hill, Elena Llaudet, Raghuveer Parthasarathy, Julie Mueller, Joe Blitzstein, Rich Gonzalez, Rohan Alexander, Pam Davis-Keen, Lauren Cowles, Holly Monteith, two anonymous reviewers, the students in our applied regression classes, and especially Beth Chance for many helpful suggestions and conversations. We also thank everyone who helped us with *Regression and Other Stories*. Above all, we thank our families for their love and support.