

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

# **Active statistics: Stories, games, problems, and hands-on demonstrations for applied regression and causal inference**

---

Andrew Gelman

Department of Statistics and Department of Political Science  
Columbia University

Aki Vehtari

Department of Computer Science  
Aalto University

©2024 by Andrew Gelman and Aki Vehtari

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

# Contents

<b>How to use this book</b>	<b>vii</b>
<b>Part 1: Organizing a plan of study</b>	<b>1</b>
<b>1 Active learning</b>	<b>3</b>
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
<b>2 Setting up a course of study</b>	<b>13</b>
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
<b>Part 2: Stories, activities, problems, and demonstrations</b>	<b>29</b>
<b>3 Week by week: the first semester</b>	<b>31</b>
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
<b>4 Week by week: the second semester</b>	<b>163</b>
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
<b>Appendices</b>	<b>319</b>
<b>A Pre-test questions</b>	<b>321</b>
A.1 First semester	321
A.2 Second semester	324
<b>B Final exam questions</b>	<b>325</b>
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
<b>C Outlines of classroom activities</b>	<b>357</b>
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.<sup>1</sup>

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*,<sup>2</sup> including data and code for all the examples for both books and slides for material to be displayed during classroom activities.<sup>3</sup>

---

<sup>1</sup>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.

<sup>2</sup><http://www.stat.columbia.edu/~gelman/regression/>.

<sup>3</sup>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, Greg Mayer, Raghavveer 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.

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

---

## Part 1: Organizing a plan of study

---

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

# Chapter 1

## Active learning

This book supplies material for a course or self-study program in regression and causal inference, aimed at students who have already taken an introductory statistics course or have the need to use statistics and want to jump right into applied methods. But the principles of active learning can be applied at any level, from introductory to advanced courses.

### 1.1 Flipped classroom and collaborative learning

Here is a description of the “flipped classroom” idea:<sup>1</sup>

“Binding credit for student participation to pre-class preparatory work enables class time to be spent in answering informed student questions and participating in in-depth discussions. Ensuring that students are prepared for class maximizes engagement during class time: the session becomes an interactive and dynamic experience where student pedagogical needs and reflections are explored.”

That’s how we think a course should go. Before each class, the students are supposed to read the book and do warmup assignments and homeworks. During class time, students are involved in activities and discussions and work together on problems—kind of like a high school math class, with the instructor as a leader and a coach, not a lecturer.

Much has been written on the benefits of active learning—classroom interactions that involve students doing things, talking with each other, and solving problems together.<sup>2</sup> Here are some key features of cooperative learning:<sup>3</sup>

- “Positive interdependence, that is, all members of a learning team are responsible for the learning of other members.”
- “The teacher designs the learning activities and monitors the groups as they are engaged in team learning. Rather than functioning solely as an expert, dispensing knowledge to students, the teacher in collaborative learning serves as a facilitator.”
- “Explicit attention to social skills. Students are required to cooperate with one another and are often given explicit rules and guidelines for appropriate social skills.”
- “Face-to-face verbal problem solving, which holds advantages for both skilled and less skilled

<sup>1</sup>From Massachusetts Institute of Technology Office of Digital Learning (2022), <https://openlearning.mit.edu/mit-faculty-residential-digital-innovations/student-pre-class-preparation-enhances-class-time/>.

<sup>2</sup>Regarding teaching in general, see, for example, Robert Slavin (1980), Cooperative learning, *Review of Educational Research* 50, 315–342, Donald Bligh (1990), *What's the Point in Discussion*, and Catherine Crouch and Eric Mazur (2001), Peer Instruction: Ten years of experience and results, *American Journal of Physics* 69, 970–977. Regarding statistics education more specifically, see George Cobb (1992), Teaching statistics, in *Heeding the Call for Change: Suggestions for Curricular Action*, edited by Lynn Steen, 3–34, Deborah Nolan and Terence Speed (2000), *Stat Labs: Mathematical Statistics Through Applications*, and Allan Rossman and Beth Chance (2001), Teaching contemporary statistics through active learning, <http://www.rossmanchance.com/pbs/pbs.html>.

<sup>3</sup>From Jim Cooper and Randall Mueck (1990), Student involvement in learning: Cooperative learning and college instruction, *Journal of Excellence in College Teaching* 1, 68–76.

students. Good students benefit from serving as tutors to the other members of the group; less proficient students receive diagnostic and remedial help from their teammates.”

- “Students who are reluctant to participate in large class discussion are often quite comfortable contributing to small group interactions.”

There is evidence that students learn more with active learning but feel like they learn less,<sup>4</sup> and this can be reflected in teaching evaluations.

So it can make sense for an instructor to get students prepared for the flipped classroom. This should start before the semester begins by making the structure clear in the course description, and it should continue on the first day of class with active student participation in the story and activity, as explained in Section 1.2 below. New teachers should also discuss the instructional plan with their supervisors or department chairs.

## 1.2 What happens during the semester?

### First day of class

The instructor should *not* begin with, “Welcome to Statistics 200. I’m Professor Vehtari and in this course we will . . .” It is better to just start the story for the first class (see Section 3.1 or, for first class of the second semester, Section 4.14), with lots of student involvement, stopping at various places to ask questions and have students work in pairs to think about them, and conclude with a path of how the story relates to the week’s reading. Then continue with the class-participation activity for the first class. All this will set a norm of student participation throughout the semester.

After going through the story and activity, the instructor can take a breath and introduce the course, explaining while writing on the board the components of the course (readings, homeworks, feedback sheet, classes, and final exam) and the structure of each class period (story, activity, discuss reading and homework, computer demo, drill, discussion problem). Emphasize that the class period is all about motivation, exploration, problem solving, and answering questions. There will be no “lectures” and no slides, beyond certain images and instructions displayed as an aid to discussion. Go over students’ responsibilities: they are expected before class to do the reading and homework assignment and to contribute to the shared document, and to show up and participate during every class. Discuss the goals of the course and what students will be expected to be able to do once the semester is over. Discuss the roles of mathematics, computing, and applications in the course. Pause to give students a chance to ask questions about the content or structure of the course.

Then is the time to go back and do the computer demo, the drill, and the discussion problem for the first class. At the end of the class, each student can be given a hard copy of the syllabus and reminded of the course website. Throughout the course, the most important things should still be written down for students. Don’t assume that everything said in class is “heard,” and also establish early for students that they will be responsible for being aware of online material.

### Later classes

The other classes during the semester should have a similar structure, except that, in the middle of class, instead of giving an overview of the course, the instructor can talk about some aspect of the week’s reading. This mini-lecture can be prepared ahead or in response to students’ questions on the shared document, or the instructor can simply answer some questions in an unstructured fashion.

Most of each class period can be spent on the prepared material: the story, activity, computer demonstration, drill, and discussion problem. Instructors can use what is in this book or develop their own materials that more directly capture their experiences and perspectives, as well as the interests

<sup>4</sup>Louis Deslauriers, Logan McCarty, Kelly Miller, and Greg Kestin (2019), Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom, *Proceedings of the National Academy of Sciences* 116, 19251–19257.

## 1.2. WHAT HAPPENS DURING THE SEMESTER?

5

and goals of the students. For example, many of our activities have a political science focus but could be changed to work with other subjects. In any case, the instructor should repeatedly explain the connections to the week’s reading and homework. Students will not necessarily see these connections without being told, and it’s important for these not to just be fun stories, activities, and so forth, but also to advance understanding of the core material.

### Projector and slides

We don’t like slides with bullet points, but we use the projector for stories, computer demonstrations, and class discussions. For computer demonstrations we set up an RStudio window and then go through code line by line, typing code into the text editor window and then copy-pasting it into the console to execute the R code.

For each semester of our course we have prepared an accompanying pdf slide deck, which includes all the images we display in our stories, class-participation activities, drills, and discussion problems. By design, the slides are minimal: they present material for the instructor to point to, but that is all.<sup>5</sup> More detailed slides can be helpful if this allows the instructor to better structure the course. We do not, however, recommend that slides be so detailed that they would be used by students as a replacement or shortcut to reading the textbook. The purpose of the slides should be to help the class period go more smoothly, not to present the material. When adapting the course, for example by introducing new examples, instructors can separately prepare what needs to be said in class and keep the slides minimal.

### Blackboard or whiteboard

We find a blackboard or whiteboard to be helpful in discussions. Here are some tips:

- Start at the upper-left corner of the board and use the space efficiently; don’t just start writing from the middle outward. Whatever’s on the board should be organized: students aren’t always paying attention at all times, so when they look up, what’s on the board should make sense.
- When speaking, write key words on the board. For example, when outlining the structure of the course on the first day, as discussed in Section 1.2, write a list on the board: Story, Activity, Discuss reading, etc. When asking students to do a task together, write that task on the board so they don’t have to memorize what they need to do.
- Conversely, when typing code into the computer or writing on the board—for example, sketching a graph or working through formulas—talk it through at the same time. Don’t write silently. Provide that second channel of communication.
- If the setup allows projecting images onto the board, the course can include some fun things such as projecting a scatterplot and then having students go to the board to draw the regression line.
- If the classroom doesn’t have a proper board, or switching between the board and the projector is inconvenient, you can simulate the board by typing with a big font.

### What should students be doing in class?

Students should be working in pairs, talking with each other, and participating in class discussions. The instructor should keep an eye on the students to make sure they are focused. If necessary, ask them to put away their phones or computers (although computers can also be used in many in-class activities for making notes or computations). After asking students to work in pairs, make sure they all pair up. If any students are sitting alone or staring into space, connect them with existing pairs and ask them to move over if needed. When students are working together, walk around the room

<sup>5</sup>These are posted at <http://www.stat.columbia.edu/~gelman/regression/> along with the rest of the material for the course.

taking a look at what they are doing. Make sure they have pens and paper out so they can write out their ideas while talking. Ask them if they understand what they’re being asked to do, and feel free to intervene in their conversations to point them in useful directions.

## 1.3 Active learning in class

### Stories

Stories are fun to tell and can be fun to listen to. We recommend beginning each class with a story. Whether instructors use the stories presented here or choose their own, they should explain the source of every story; consider its difficulty level; include student involvement during the storytelling process; and conclude by connecting to the week’s reading, the themes of the course, and the assumptions implicit in the storyline.

To get a sense of these general principles, stop and read the story for our first class period, the Wikipedia example in Section 3.1.

The *source* of that particular story was a consulting experience of ours, a serendipitous event that illustrates the principle that just about every problem is of statistical interest when looked at carefully. Reality is fractal. There is a tendency in textbooks to smooth over real-life difficulties; here we go into some of the hairy details.<sup>6</sup>

The *difficulty level* of the story is mixed. There are some quantitative steps involved in computing standard errors, and this uses formulas and concepts that aren’t covered until several weeks later in the course, but the main points of the story should be clear even to students who hadn’t seen those topics before.

When *telling the story*, the point is to give a sense of the use of experimentation and statistical analysis to attack an applied problem (in this case, increasing donations to Wikipedia) and then to see how we reacted when things went wrong. *Student involvement* comes when you stop the story at various places and ask questions to the students. After showing students Figures 2 and 3 and explaining that the estimated treatment effect is implausibly large, we ask them to look at the screenshot carefully and discuss in pairs to figure out what went wrong. Then we display Figure 4 and ask how this explains the large treatment effect. After going over this error, we ask the students to discuss in pairs how the problem could be fixed.

We conclude each story by *connecting to the week’s reading and to the themes of the course*. For the story on the first day of class, the relevant topics come in Chapter 1 of *Regression and Other Stories*, and after finishing the story and class discussion, we ask the students to discuss in pairs the specific concerns in this example regarding generalizing from sample to population, from control to treatment, and from observed data to underlying constructs of interest. We then mention the larger themes of the connection between data collection and analysis, the importance of looking at data and checking assumptions and the relevance of mathematical calculations (in this case, the standard error of estimated proportions).

We also recommend the book, *Telling Stories with Data* by Rohan Alexander, which covers a wide range of topics on statistical communication, programming, and modeling to supplement any statistics course or self-learning program.<sup>7</sup>

### Class-participation activities

For each class period we have constructed an activity that involves active student participation. Some of these activities involve collecting and analyzing data from the students in the class; others require students to estimate numbers or assess uncertainty. As with the stories, each of these activities has a

<sup>6</sup>For some examples of confusion and misrepresentation of data even in well-regarded texts, see Andrew Gelman (2011), Going beyond the book: Towards critical reading in statistics teaching, *Teaching Statistics* 34 (3), 82–86.

<sup>7</sup>Rohan Alexander (2022), *Telling Stories with Data*, CRC Press, <https://tellingstorieswithdata.com>.

### 1.3. ACTIVE LEARNING IN CLASS

7

direct point, but it is important also to draw the connection to the week’s readings and homeworks. We often use the trick of throwing questions back at the students, asking *them* why they think we did a particular activity and how it relates to the larger themes of the course. By doing active learning every class period, we try to establish the norm of student participation. On the occasion when we throw a question to students and no one responds, we call on individual students, giving them the option to go along to the line to others if they are not ready to participate.

When teaching the course, we have not required students to hand in any responses to stories and class discussion. The instructor should do make sure that everyone is involved and participating while these activities are happening, talking to individual students as necessary to keep them active.

#### Hands-on computer demonstrations

A short demonstration is a great way to demonstrate the feel of live data analysis while providing an opportunity to field questions about programming, statistical analysis, and graphics using R (or whatever software is being used in the course). There is no need for the examples to be elaborate: even a few lines of code can be enough, and remember that this is just a single component of a 75-minute class. Everything is projected on the screen or board, and the instructor should speak the code aloud while typing it, stopping after each paragraph (as indicated by a blank line in the code) to explore and answer questions. For example, after reading in or simulating a vector called `x`, we might type `print(x)` into the console to take a look at its values. We post the code on the course website so that students can see it ahead of time or review it after class if they would like.

As part of the demonstration it is important that the instructor types the code live, or, if you are doing self-study, that you type the code rather than just copying it from an online source. For class the instructor should get the code working ahead of time so that the main part of the demonstration goes smoothly. Then there should be some time for improvisation, at which point some errors can arise. Some of the mistakes that we and our colleagues have made in class include forgetting how basic functions work (even some of those functions that we wrote!), not finding a data file where we expect, and messing up brackets or variable assignments and spending five minutes searching for the error. When these and other mistakes happen, watching the instructor resolve the problem is one of the most important lessons of the activity. Students need to learn not just how to do things right but also how to identify and fix mistakes.

#### Discussion of questions related to readings and homeworks

There is much more in any textbook than can realistically be covered in class, especially given that we are not following a lecture format. *Regression and Other Stories* includes discussion of general principles and methods, stories, worked examples, and code. During each class period we remind students of the week’s reading and take some time to clarify points that confuse them, while recognizing that we can select only a subset of topics for discussion.

Before each class, students are required to contribute to a shared online file such as a Google doc, either asking a question on the readings or homework assignment due for that class, asking a more general question about the topics of the course, or responding to a question that another student has already asked. Before the class we quickly scan the day’s document and choose a few questions to discuss during class. This is yet another way to promote student involvement as well as an opportunity to resolve confusion. And, when in doubt, use the tactic of asking rather than telling. When a student asks a question, we can reply, What do you think? Eventually we have to answer the question (or admit we don’t know), and we don’t want to drag this out too much, but occasionally bouncing a question back can help clarify what the student’s question really is.

When responding to questions or getting into discussions, we avoid digressions that don’t advance the main thread of the course. It’s fine to skip a question or to put it off until later, and, again, when answering questions and leading discussions it is a good idea to repeatedly touch back to the week’s readings and homeworks. Connections might seem obvious to the instructor but not so clear to

students. If the class has a teaching assistant, these questions can also serve as the basis for further clarification in section meetings.

### Mini-lecture

Responses to student questions will often include a mini-lecture during which the instructor works out a problem on the blackboard. For example, an important but challenging skill in logistic regression is to go back and forth between the coefficients ( $a, b$ ) and the curve,  $\text{logit}^{-1}(a + bx)$ . When we reach logistic regression in the course, we demonstrate this activity in both directions: given the coefficients, drawing the curve, and given the curve, extracting the coefficients; see for example Figure 13.1 in *Regression and Other Stories*. We do this on the board, drawing the curves freehand and performing the computations approximately, emphasizing that these are skills we expect the students to be able to perform themselves for the exams and more generally when interpreting fitted models in the literature.

### Drills

In past years we have often found students to have difficulties with what we had considered to be basic concepts routine computations. In response, we try to spend some time every class on a drill: a set of short easy problems that we write on the board and ask students to work on in their notebooks and then discuss in pairs. This book has a set of drill questions for each class period, with a solution given for the first problem in each group.

Sometimes a drill is just a series of related questions; other times the questions build up in complexity to get at different angles of a problem. In presenting a drill, we solve the first problem on the board, and students can use this as a template to solve the remaining problems on the list. After the drill is done, we can discuss its relevance to the week's readings and the course more generally.

### Discussion problems

Finally, we like to conclude each class with a prepared discussion problem on which students can work together in pairs or small groups. These are similar to “concept tests” in peer instruction.<sup>8</sup> A discussion problem should be challenging, so that students can work for a few minutes together to think about it. Some discussion problems are open ended; others have precise answers. Often the challenge is not in solving the problem but more in setting it up. And, again, end the discussion by tying the problem back to the day’s readings and also situated it within the larger plan of the course.

### Substitutions

This book presents material for 52 class periods, with warmup assignments, homework assignments, a story, a class-participation activity, a computer demonstration, drills, and a discussion problems for each. There is not always a sharp division between all of these. The stories should be told with many pauses for student participation, the activities have aspects of storytelling, the demonstrations often connect to the stories and activities, and the discussion problems often also include stories, as well as being opportunities for students to work together, often with coding. The warmup and homework assignments and drills represent different ways to practice the course material.

## 1.4 Scheduling

Think of the course as presented here as a template: instructors can teach it as is or swap in alternative readings, warmup and homework assignments, stories, activities, computer demonstrations, drills,

<sup>8</sup>See, for example, Eric Mazur and Jessica Watkins (2009), Just-in-time teaching and peer instruction, *Just in Time Teaching: Across the Disciplines, Across the Academy*, edited by Scott Simkins and Mark Maier, 39–62, Sterling, Va.: Stylus.

and discussion problems. The key is for all these pieces to fit together, providing a framework for teaching and learning.

## Workload

The course as designed here is intended to average 10 hours per week, roughly divided as:

- 1 hour reading the textbook
- 0.5 hours on pre-class warmup assignments
- 2.5 hours in the classroom
- 6 hours on homework.

The semester concludes with a 3-hour final exam for which students are expected to study for about a week, that is, another 10 hours.

The point of giving these estimates is not to dictate how other courses should go but rather to provide a baseline: if students are expected to devote more or less than 10 hours each week, the workload should be adjusted accordingly. In addition, in describing in-class plans in this book, we have erred in the direction of providing more details, and realistically it will not be possible to do all the stories, activities, demonstrations, drills, and discussion problems given here for each class. Depending on how the class is going and how much time is needed to go over student questions, it will be necessary to skip some of the listed material. This is fine: the key points are to keep students actively involved and working together during the entire class period and to connect each week’s material with real-world examples.

It’s also important to control the workload of the instructor and any teaching assistants for the course. Indeed, one of our motivations in preparing this book is to make active learning accessible to busy instructors: we have put together these stories, activities, demonstrations, and problems so that instructors can be freed to spend their time helping students learn. In a large class, a big challenge is grading, especially if there are no teaching assistants to help. In that setting, it can make sense to set up a peer grading system with some subset of assignments checked directly by the instructor. The other challenge in a large class is meeting with students; here the scalable solution is scheduled help sessions where the instructor can help students figure things out together, rather than frustrating and time-consuming one-on-one meetings.

## Fitting active learning into a busy class period

We have arranged the activities in this book to fit into a series of 75-minute classes, following roughly this plan:

1. (10 minutes) Story
2. (15 minutes) Class-participation activity
3. (15 minutes) Class discussion of questions related to readings and homeworks
4. (15 minutes) Computer demonstration
5. (10 minutes) Drill
6. (10 minutes) Small-group discussion problems.

These timings are only approximate, as the stories, activities, and demonstrations in this book vary in how much class time they require, and the time allocated for discussion of readings and homeworks is flexible. Sometimes a story can take 20 minutes of class time, especially if it is used as a springboard for class discussion. Realistically, we rarely include all the above segments in every class, as usually one piece or another takes longer than expected. Instructors can adapt based on the progress and struggles the students show in class each week.

Many of the stories, activities, and demonstrations in this book could take a lot more class time if you let them. These are fun examples—we only included things in this course that were interesting to

us—and so, indeed, any single piece contains depths that there would not be time to explore without elbowing out other topics for the week.

With that in mind, you should consider the material here as resources more than as scripts. If a story would take too long to go through in detail, or if an activity has so many steps that it would feel like a distraction, then feel free to use shorter versions that will fit your needs. It's good to switch up activities every 10 to 20 minutes rather than getting stuck too long doing the same thing, even active learning. Similarly with the computer demonstrations: you can start one, entering and altering code until you feel that you've learned something, and then stop, with no requirement to go to the end. In this book we would rather include more detail than less, with the understanding that instructors and students will take what they need, and then the rest is always available for later study.

### Pre-class warmup assignments

We have prepared a few quick questions due before each class to keep students on track regarding the week's reading. Instructors should feel free to adapt these to the needs of their classes. The warmup assignments are intended to be straightforward for students who have read the relevant textbook material (in other words, when the students look at the questions they are guided to read the relevant material).

### Homework assignments

Everything that students are expected to learn should be in homework assignments and drills. There is no point in doing a derivation on the board, for example, if students are not required to reproduce it in some form by themselves. But with this realization comes a challenge, as it is difficult to include homework problems on all the important topics covered in the course while still respecting the limited time that students have to spend on any course. We suspect that the best solution here is to limit what is covered in class—better to cover one key concept and set of skills and make sure that students can do them well, than to hit several topics shallowly. For now, though, we'll just say that the instructor should affirmatively decide on what skills the students are expected to learn, give practice in these skills, and not to expect learning on topics that are mentioned in class or the textbook but not included in homework assignments and exams. It can be helpful to provide guidance on readings to help students decide where to focus their effort. Also it's important for the instructor to monitor progress during the semester to see which assignments are giving students trouble, to give a chance to reinforce these lessons.

We assign homework assignments with a range of difficulty levels. It's valuable to nail down key skills that can be used as building blocks, to have more involved problems that challenge students and promote independent thinking, and to give students practice collaborating on some of the computing challenges involved in working with real applications.

## 1.5 Assessment and feedback

### Tracking students during the semester

Warmup assignments and homeworks are due for every class, and students should be actively involved during the class period, so there are many opportunities to observe and evaluate them and to talk with them when they run into trouble. If the class is large and the instructor does not have the time or resources to grade the warmup and homework assignments, they can be graded using peer grading. It is important that students try these problems so they can follow the discussions in class and learn from their mistakes.

### Final exams and practice final exams

We have prepared several multiple-choice exam questions for each chapter covered in the course. Having created this test bank, we sample one question at random from each chapter and put these together to form a practice exam to give to the class a few weeks ahead of time. We then create the exam itself by taking a new sample of two questions per chapter. The exam questions we prepared for the two semesters corresponding to Chapters 1–12 and 13–22 of *Regression and Other Stories* are in Appendices B.1 and B.2 of this book. Instructors can use these questions directly, create new questions by making small changes to the ones here, or write entirely new questions. The exam question bank can be open to the students, which guides the students to learn what is also tested. When creating new exam questions, it is good to think are what the learning objectives and how student learning is guided by these new questions.

### Grading

Grading should be transparent and efficient. We use a weighted average of scores on pre-class assignments (graded based on completion rather than correct answers), homeworks, class participation, and the final exam.

At each stage, we recommend that grading be clear and simple. We grade each homework problem as 1 (correct), 0.5 (attempted but wrong), or 0 (not attempted), with no partial credit beyond that. Class participation for each period is a 1 (contributed meaningfully to the shared document) or 0 (did not contribute). The final exam is multiple choice, so each problem gets a grade of 0 or 1, with no partial credit. From a statistical point of view, this works because the grade is averaged over multiple homework problems for each class period, many class periods, and many final exam questions, so it is not necessary to make fine-grained decisions for each problem. Simple grading also makes it more feasible to have students correct and resubmit their homework assignments.

## 1.6 Some general issues in teaching and communication

We recommend *How to Talk So Kids Will Listen and Listen So Kids Will Talk* by Adele Faber and Elaine Mazlish.<sup>9</sup> That book, which originally appeared in 1980, is not explicitly about teaching, but we find its themes and techniques to be directly relevant to working with students and communication more generally. A lot of the ideas involve being direct about your goals and obstacles.

When communicating with students about the course itself, we have to be careful: we want students' feedback, but we don't want to turn them into theater critics either. We try to elicit their reactions and suggestions while making it clear that this is all within the bounds of the course as we have designed it.

There's also general advice for communication that works in class too, such as using people's names, answering questions with questions such as "What do you think?" that motivate students to stay focused in the conversation, and never describing a task as easy. Instead of saying, "We'll start with a simple example," or "Here's an easy problem for you," say, "Here's a problem that's challenging but doable, if you go about it the right way."

Pair discussion gives students a space to consider ideas they might not be ready to share with the whole class, and working in pairs gives them a chance to explain things to each other. Teaching can be an effective way to learn.

<sup>9</sup>Adele Faber and Elaine Mazlish (1980), *How to Talk So Kids Will Listen and Listen So Kids Will Talk*, New York: Simon and Schuster.

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

## Chapter 2

# Setting up a course of study

There are two key questions when designing a course: What to teach? and How to teach it? Or, from a student's perspective, What to learn? and How to learn it?

## 2.1 What to learn and how to learn it

### The importance of structure

Stories, group activities, and computer demonstrations are great tools for learning applied statistics—not on their own, but as part of a structured course. At the time scale of the semester, the instructor should be aware of what are the learning objectives, and these should be conveyed to the students. A starting point is to put together the final exam before the course begins so that it's clear what specific skills the students will need to learn. Then as the class goes on, the instructor should continue to connect each aspect of the class to the progress of the semester—no loose ends, and with each lesson tied to a skill that students can practice.<sup>1</sup>

If you are a student following this book on your own, we encourage you to pay attention to the placement of the stories, activities, and demonstrations within the schedule of the course so that you can see statistical concepts in action while getting practice in specific techniques.

### What to learn

There are a million possible things to learn in statistics and adjacent fields such as mathematics, programming, and numerical analysis. Statistics is in the overlap of data description and mathematical and probabilistic modeling, and from that perspective it makes sense for a first course in applied statistics to be centered on linear regression, which is both a mathematical model and a tool for data description and causal inference.

Here we present a two-semester course on applied regression and causal inference that we have taught to political science students at Columbia University. The first semester of the course goes through Chapters 1–12 of *Regression and Other Stories*, covering the following topics:

- Applied regression: measurement, data visualization, modeling and inference, transformations, and linear regression.
- Simulation, model fitting, and programming in R.

The second semester covers Chapters 13–22 of *Regression and Other Stories*, including the following topics:

- Applied regression: logistic regression, generalized linear models, poststratification, and design of studies.

<sup>1</sup>See Andrew Gelman (2019), Here's why you need to bring a rubber band to every class you teach, every time, <https://statmodeling.stat.columbia.edu/2019/09/08/rubber-band/>.

- Causal inference from experiments and observational studies using regression, matching, instrumental variables, discontinuity analysis, and other identification strategies.
- Simulation, model fitting, and programming in R.

Throughout we touch on the key statistical problems of adjusting for differences between sample and population, adjusting for differences between treatment and control groups, extrapolating from past to future, and using observed data to learn about latent constructs of interest.

We follow the usual practice of centering the course around a textbook, in this case our own *Regression and Other Stories*. The homework assignments come from the book, and the classroom activities relate to each week's readings. If a different book is being used, the other aspects of the course should be changed accordingly. As a precursor to warm up before getting into the details of regression modeling we recommend *Data Analysis for Social Science* by Elena Llaudet and Kosuke Imai, a compact introduction to statistical analysis, computing, and causal inference.<sup>2</sup>

When the year is over, students are expected to be able to graph scatterplots and regression lines, simulate data (that is, program the computer to generate synthetic or “fake” data), interpret linear and logistic regressions, and use them to make causal inferences when appropriate. When designing a course, an instructor may have different goals; whatever they are, it is important to make them clear to the class.

### How to learn it

Student-centered learning is an approach to instruction in which students play an active role. In a fully student-centered course, the student would choose what topics to learn and how best to learn them. Our proposed course could perhaps best be described as teacher-centered in its structure but student-centered in the classroom. The instructor decides ahead of time the material to be learned, along with homework assignments and exams, which are the same for all students. But the class periods are structured to facilitate active student involvement.

Here are the components of the course, beyond what is in the textbook:

- Final exam and homework assignments. These determine what the students will be expected to learn.
- Pre-test to be given at the beginning of the course, so that instructors and students can find out students' level of preparation and get a sense of what they have learned during the semester.
- In-class instruction. We assume two 75-minute classes per week, with each class proceeding as follows:
  - A statistics story related to the class topic.
  - A class-participation activity.
  - Discussion of student questions related to the most recent readings and homeworks.
  - A demonstration on the computer, in which the instructor goes through each step and fields questions from the students.
  - A set of drills: quick problems that reinforce key skills.
  - A discussion problem: a harder problem that students work on in pairs.

We divide the students into groups of two or three at different times during the class to facilitate discussion and to ensure that they stay focused during the entire class period. When students are working in small groups, the instructor should walk around the classroom to make sure the students are making progress, to get a sense of what they are doing, and to help as needed.

Our in-class schedule is busy, and we recognize that it might not always be possible to get through all these activities listed above in 75 minutes. If the discussion takes longer than anticipated, it's

<sup>2</sup>Elena Llaudet and Kosuke Imai (2023), *Data Analysis for Social Science*, Princeton University Press.

fine to skip some of the planned activities. The point is to keep students engaged and thinking about statistics during the entire class.

In the present book we go through the 26 weeks of the two-semester course, listing all the above components for the two classes each week.

## 2.2 Computing

We don't expect you, as a student being introduced to applied statistics, to learn the mathematical derivations of linear regression or related points of statistical inference. But you should learn how to perform these statistical procedures on the computer, and you also will need to learn how to use the computer as an experimental tool, simulating fake data as well as plotting and fitting models.

Programming is a key part of our course for three reasons:

- To truly engage with data you need to be able to work flexibly, that is by programming rather than just clicking through options in a menu. This goes for data manipulation, visualization, modeling, and post-processing of inferences.
- Computer experimentation allows you to work directly with probability models, for example by simulating data and then checking that the underlying parameters can be recovered in a model fit.
- Visualizing data and fitting models are core skills that you should learn in any applied statistics course.

So the instructor needs to choose some statistical programming language to use in the course. In *Regression and Other Stories*, we use R and RStudio, along with the `rstanarm` package.<sup>3</sup> Depending on the students in the class and what they plan to do next, it could make sense instead to use Python, Stata, or some other software. In any case, it is important to allocate enough time at the beginning and throughout the semester for learning and practicing computing skills.

Even after choosing a computer language for the course, the instructor will need to make some decisions on how to teach it. We use what is called “base R,” which is derived from the statistical programming language S which we learned many years ago. Base R has a benefit that the code looks similar to other programming languages students might have learned. Nowadays, many of the best practitioners of R use the `ggplot2` and `tidyverse` R packages for graphics and data analysis. These have great benefits when working with data frames but use a different programming paradigm.

We suggest instructors use software with which they are comfortable and which they think will serve students’ future needs.

Key computing tasks include downloading or entering data, manipulating data, graphing data, simulating data from probability models, fitting models to data, and plotting fitted models. For this course, some of the datasets are collected from students in class and typed directly into files or read from online forms, and we also access many datasets that we have prepared and put online. We purposely have kept these data online and not included them as prepared datasets within an R package, because we want students to gain experience of downloading data that have not been prepackaged. Similarly, much of our code for data manipulation, plotting, simulation, and fitting is unpolished: we are modeling how a student or applied researcher might work through an analysis in real time, rather than presenting publishable code in final form.

## 2.3 Course material

We follow the sequence laid out in the preface to *Regression and Other Stories*. That book is structured through models and examples, with the intention that after each chapter the student should

<sup>3</sup>See these websites: <https://www.r-project.org/>, <https://posit.co/products/open-source/rstudio/>, and <https://mc-stan.org/rstanarm/>.

have certain skills in fitting, understanding, and displaying models. The list of topics is long, and for your course or self-study program you might just choose a relevant subset.

- *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 liner 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.

- *Appendices:*
  - *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, students 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.

In a two-semester course we cover Chapters 1–12 in the first semester and Chapters 13–22 in the second semester. In a single-semester course for better-prepared students, we cover most of the book, skipping a few topics as necessary for the schedule.

## 2.4 Real data and simulated data

The stories, activities, and demonstrations in this book take on two different forms: real data and simulated data.

- Real-data examples are valuable most directly in demonstrating how statistical methods are relevant to live problems. In addition, as the saying goes, God is in every leaf of every tree: look closely enough at any real example and you will come across something unexpected, an anomaly that, if you allow it, can recenter how you think about statistics.

We demonstrate using the examples from the first two weeks of the course. The Wikipedia story (page 33) demonstrates how clean experiments can fail, and also how careful study of available data can raise suspicion. The story of the *Literary Digest* poll (page 36) first shows how things can go wrong with a non-randomly-sampled survey and then expands into a story of how problems with data can partially be fixed using modeling and adjustment. The United Nations peacekeeping story (page 45) introduces the use of pre-treatment measurements to adjust for selection bias in causal inference. The story about girls' participation in sports (page 47) shows how we can look carefully into a published claim and figure out where it is coming from.

Each of these stories involves methods that go beyond what is covered in the course. The stories are aspirational and represent what statistics can do; they also reveal pitfalls from naive uses of statistical methods.

- Simulated data can come from the computer or simple data-collection activities involving students in the class. In either case, the most direct lesson is how to write the code to create the simulated data or go through the steps to collect data from students. More generally, simulating data is the statistical equivalent of laboratory experimentation in science: you can control the conditions and see what comes out, then vary the conditions and see how things change. All of this is crucial to understanding. Statistics is not just a set of tools for analyzing data; it also includes models for generating data. Often real data don't look like we expect, and we can train our expectations by understanding what is implied by our probability models.

It is valuable to demonstrate both sorts of examples—real data and simulated—as they represent different aspects of statistical workflow.

## 2.5 Two kinds of computer demonstrations

This book includes some demonstrations where the code is simple and the focus is on reading the output, and some where the lesson is in the coding itself. Here is an example of the first kind:

```
library("rstanarm")
# Read in the data
hibbs <- read.table(paste0("https://raw.githubusercontent.com/avehtari/",
```

```
"ROS-Examples/master/ElectionsEconomy/data/hibbs.dat"), header=TRUE)
# Plot the data
plot(hibbs$growth, hibbs$vote,
      xlab="Average recent growth in personal income",
      ylab="Incumbent party's vote share")
# Estimate regression and display fitted model
M1 <- stan_glm(vote ~ growth, data=hibbs, refresh=0)
print(M1)
plot(M1)
```

The code here is straightforward, and students can focus on interpreting the results.

In contrast, here is an example of a computer demonstration that focuses on the code:

```
library("rstanarm")
n <- 50
a <- 2
b <- 3
sigma <- 4
n_loop <- 100
inside_68_interval <- rep(NA, n_loop)
for (loop in 1:n_loop) {
  print(loop)
  x <- runif(n, 0, 10)
  y <- rnorm(n, a + b*x, sigma)
  fake <- data.frame(x, y)
  fit <- stan_glm(y ~ x, data=fake, refresh=0)
  b_hat <- coef(fit)[["x"]]
  b_se <- se(fit)[["x"]]
  inside_68_interval[loop] <- abs(b_hat - b) < b_se
}
print(mean(inside_68_interval))
```

This example has statistical content—it's a check that approximately 68% of the estimates fall within  $\pm 1$  standard error of the true value—but the main challenge here for students will be the programming.

We recommend that, when going through in this book, you adapt the computer demonstrations to your needs, focusing more on interpreting output or on coding depending on your goals and where you are in any particular week.

When doing either sort of demonstration, we often take the opportunity to perturb the code in different ways and see what happens. This is also a great way to involve students. We can ask questions of the form, What would happen if you try this?, or What could you do to cause this code to break?, and then discuss various possibilities before trying them out live.

The computer demonstrations in this book are *not* intended to show best practices for R, or for programming more generally. The code we present here is messy, representing the way that we program in real time as we are trying to figure something out, either analyzing existing data or simulating fake data as a way to better understand our models. So take the demonstrations in that spirit, and consider our code as examples of what can be done, a set of starting points rather than an ideal.

## 2.6 Challenges in learning particular topics

There are some important topics in statistics that are particularly difficult to learn. We offer no silver bullets here, just some scattered thoughts.

**Programming, statistical software, and working with data.** To do all but the most basic data analysis requires programming. This is fine, as programming, like mathematics, is a useful skill in itself. But many students do not know how to program, or have difficulty generalizing their understanding of

## 2.6. CHALLENGES IN LEARNING PARTICULAR TOPICS

19

programming to R or whatever other language is used in the course. So time must be scheduled at the beginning of the course to getting up to speed on data manipulation and graphics on the computer. This can include help sessions every week and an online bulletin board for students to help each other.

You can get started by working through Appendix A of *Regression and Other Stories*; in addition we reinforce the importance of coding with a computer demonstration during every class period, with each demonstration being self-contained enough that the instructor can explain it while typing in the code live in class.

You might need extra help on coding during the first few weeks, so the instructor should arrange help sessions and encourage students to work together to get each other unstuck.

**Logarithms.** Exponential and power-law relationships (for example, a population that increases 5% per year, or an elasticity function by which a 1% increase in price leads to a 0.6% decrease in sales) are fundamental to serious quantitative reasoning. Unfortunately, students are typically uncomfortable with logarithms and exponential functions, which are the mathematical building blocks of these curves.

When teaching statistics and quantitative reasoning, the instructor needs to choose: cover logarithms or don't. Either way, commit to the choice. If the course includes logarithms, this could take a week or more of the semester. If the course does not cover logarithms, then it's good to be clear to students what they are missing, so they can go back later and learn the topic. We include some examples and activities on logarithms, exponentials, and power laws; see Sections 3.4, 3.13, 4.17, and 4.21.

**Probability.** The traditional introductory undergraduate or high school statistics class includes a few weeks of probability: the laws of probability, probability trees, conditional probability, random variables, expectations, and variances. This all has two purposes: First, probability is an important application of mathematics in its own right, relevant for decision making, forecasting, and other problems involving uncertainty. So there is value in teaching probability, even setting aside its use in the theory of statistical inference. Second, probability theory and probability models underlie central ideas in applied statistics, including standard errors, uncertainty intervals, hypothesis testing, prediction, and Bayesian information aggregation.

A challenge in an applied statistics course is how much probability to include. Mathematical derivations of the law of large numbers and central limit theorem: that's way too much. It is possible, though, to teach probability trees and conditional probability and to demonstrate these ideas on the computer. It's fine to devote a week or two to this or just to skip it entirely. It's just a decision for the instructor to make, what to include and what to leave out of the course. Even if there is time to teach the basics of probability, there will not be time in an applied statistics course to fully connect probability theory to statistical inference. So you have to accept ahead of time that estimation and uncertainty will be handled non-rigorously.

**Sampling distribution of means, differences, and regression coefficients.** Even simpler derivations, for example the proof that  $E(x+y) = E(x) + E(y)$  or that  $\text{var}(x+y) = \text{var}(x) + \text{var}(y)$  for independent variables  $x$  and  $y$ , are really too much to fit in a one-semester statistics course. It also doesn't work just to state these relationships and then try to use them in further reasoning. One thing we don't like that we've seen in some introductory courses is a sort of pseudo-derivation, where the expressions for  $E(x+y)$  and  $\text{var}(x+y)$  as given and then used to derive the sampling distribution of the sample mean, using reasoning such as,  $E(\bar{y}) = E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n E(y_i) = E(y)$  and  $\text{var}(\bar{y}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(y_i) = \text{var}(y)/n$ . These algebraic steps may seem simple and intuitive to an instructor with a solid background in probability theory but will go over the heads of students coming in without that mathematics.

That said, the formulas for the means and variances of sums, differences, and weighted averages are important in applications, so we need to teach them, even if we can't really derive them. The idea that the standard deviation of the average often declines like  $1/\sqrt{n}$ : that's important, but it just needs to be taught directly as a fact that students need to learn. It's not just an empirical pattern—it can be proved mathematically—but the proof is beyond the scope of the course, and we give drills,

homework exercises, and exam problems for which students need to apply this idea. Similarly for expressions such as the standard deviation of a weighted average and the behavior of uncertainty intervals as sample size increases.

**The multiple regression estimate and its derivation.** In class we do not present the expression for the least-squares estimate; indeed we do not even discuss the idea of inverting a matrix or solving a linear system. Much depends on the level of the course and the background of students; at some universities, applied statistics comes after calculus and linear algebra, while at other universities, such mathematical knowledge might be in the course’s official prerequisite but is not actually expected of the students. In the applied regression and causal inference material discussed in this book, an understanding of matrix algebra is not expected, but in a course where students have such knowledge, these connections can be made during class. Even in the most applied version of the course, we do discuss the concept of least squares (see Sections 8.1–8.3 of *Regression and Other Stories*) but we neither derive nor explain the estimate itself. As with other topics at this mathematical level, for this course we would rather forthrightly *not* cover it than present a sort of fake proof that would just take time away from material that students will actually be prepared to learn and use. If they are interested in the topic they can study it in a theoretical statistics course.

**Bayesian inference.** We had a difficult time with Bayesian inference when writing *Regression and Other Stories*. On one hand, a Bayesian perspective is pretty much necessary for the computational approach that we want to use for inference and prediction: we summarize uncertainty in any fitted model using simulations of possible values of the unknown parameters, and we propagate that uncertainty when making predictions. On the other hand, the mathematics of Bayesian inference for regression models requires calculus, and we do not expect students to be performing integration or multivariate analysis in this course. Even students who have taken calculus do not always find it easy to apply these ideas to distributions without learning probability theory, which is beyond the scope of our applied statistics class.

As with other aspects of mathematical statistics, we walk a fine line here, using Bayesian methods and trying to develop some intuition while making it clear that we do not attempt to derive the results. It’s similar to how you can learn to drive a car and have some sense how it works without a full understanding of the physics or the ability to build a car yourself.

**Informative prior distributions.** We have a tough choice when introducing and using priors in regression. As discussed above, we definitely want to use a Bayesian approach—at least implicitly, in the sense of summarizing inferences and predictions using simulations from a probability distribution—even if we have no plans to derive the relevant formulas or fully explain the Bayesian approach mathematically in terms of a joint distribution of parameters and data.

But even within this implicit Bayesian framework, we have a choice of what prior distribution to use on the regression coefficients. Three options naturally present themselves:

1. *Uniform prior distribution.* This has the virtue of yielding the same point estimate as least-squares regression but has the practical problem of yielding noisy estimates when data are sparse relative to the number of predictors.
2. *Default weakly-informative prior distribution.* This is what we recommend as a starting point in applied work, but it requires more explanation when introducing the topic.
3. *Informative prior distribution tailored to the problem at hand.* This is ideal but it goes beyond the scope of the course. Also, even when informative priors are recommended, it is good to have a basic default regression as a starting point.

When writing *Regression and Other Stories* we chose the second option above, using a default weakly informative prior and repeatedly explaining how to set the software to use a flat prior if that is desired; see Sections 1.5, 8.4, and 9.5 of that book.

In retrospect, perhaps it would have been better to have started with simple least-squares regression and then later introducing weakly-informative and strongly-informative priors in the

context of Bayesian inference. Such an approach would have the drawback of requiring a shift of software (from `lm` to `stan_glm`) and conceptual framework (from least squares and maximum likelihood to Bayes) midstream, but with the advantage of more clearly connecting to standard non-Bayesian practice.

**Standard errors, uncertainty intervals, and the  $t$  distribution.** A more subtle difference between classical and Bayesian regression arises with small-sample uncertainty. With least-squares regression comes the standard error, and then the 95% confidence interval is approximately the estimate  $\pm 2$  standard errors, but with a larger factor when sample size is small, using the  $t$  distribution. The standard error is the estimated sampling variation of the estimate, but because the standard error is itself estimated with uncertainty, you need the  $t$  distribution. With Bayesian inference, all the uncertainty gets swallowed up in the simulation, and you can simply report the posterior mean  $\pm 2$  times the posterior standard deviation (or the approximate equivalent based on median and median absolute deviation, as explained in Section 5.3 of *Regression and Other Stories*), without having to worry about the  $t$  distribution. This is a place where Bayesian inference makes things simpler, but, again, it might be necessary to explain what's going on here to fully link to the classical approach.

**Generalized linear models.** Our course focuses on linear regression, with a couple of weeks on logistic regression as well. These models are part of a larger mathematical class of *generalized linear models*. We use that term as the title of Chapter 15 of *Regression and Other Stories* but not exactly following the formal definition in statistics.<sup>4</sup> In particular, we include ordered logistic regression, which is a useful generalization of linear modeling and feels like what should be called a “generalized linear model” but does not formally fall into that category. This is no big deal; we just need to be careful when teaching negative binomial regression, ordered logistic regression, and the other models in Chapter 15 that we are using an informal definition of generalized linear model.

**Poisson regression.** Poisson regression is a generalized linear model that is used for count data, where the outcome can take on values 0, 1, 2, etc., for example predicting the number of car crashes in different intersections in a city, given information such as local traffic levels and speed limits. As discussed in Section 15.2 of *Regression and Other Stories*, the Poisson model has serious limitations and we almost always recommend using negative binomial regression instead. We mention this here only because it is standard for texts on generalized linear models to present Poisson regression as the go-to method for count data, and we don't want you to fall into that trap.

**Causal inference.** We cover causal inference—estimating treatment effects—informally throughout the course and *Regression and Other Stories* and then formally in Part 5 of the book and the final weeks of the course. The informal coverage begins on the first page of Chapter 1, where we characterize one of the three challenges of statistics as, “generalizing from treatment to control group, a problem that is associated with causal inference, which is implicitly or explicitly part of the interpretation of most regressions we have seen.” We continue throughout with examples of regressions for causal inference and also examples for which direct causal interpretations are inappropriate. This causal thread runs throughout, and then students should be ready near the end of the course for a more formal treatment of the topic.

**Data collection and design.** Introductory statistics courses often emphasize the importance of careful data collection, most notably random sampling in surveys and random assignment of treatments in experiments. In the course laid out here, we also do this, with two additions. First, we discuss the challenges of measurement; see Chapter 2 of *Regression and Other Stories* and Section 3.3 of this book. Second, we discuss adjustments for imbalanced comparisons and non-representative samples; see Sections 17.1 and 19.4 of *Regression and Other Stories* and various examples in this book. Indeed, the two stories included here for the first week of classes (Section 3.1) concern biases in real-world experiments and surveys, and how to adjust for these biases. We hold random sampling and

<sup>4</sup>See John Nelder and Robert Wedderburn (1972), Generalized linear models, *Journal of the Royal Statistical Society A* 135, 370–384.

experimentation as ideal without implying that it is hopeless to analyze data collected nonrandomly or with incomplete randomization.

**Hypothesis testing and  $p$ -values.** There is a cluster of ideas in classical statistical theory and practice that we do not like, including null hypothesis significance testing,  $p$ -values, and type 1 and type 2 errors. We define all these in Chapter 4 of *Regression and Other Stories*—and we explain there why we do not find these concepts useful. We do not cover them in class, nor do we include them in homeworks or exams. We do include the related concepts of standard errors, uncertainty, and interval estimation; we just do not frame these in terms of hypothesis testing, significance levels, or  $p$ -values.

If these methods are to be included in the course—and we recognize that there are good reasons to do so, including keeping up with other statistics classes and much of the applied literature, along with being able to communicate with researchers who have received classical statistical training—then we recommend assigning drills, homework exercises, and exam questions on these methods.

## 2.7 Adapting to your goals and learning style

**Background and speed of the course.** We taught this course in 2021–2022 to about 40 political science students at Columbia University. When teaching to statistics majors or more quantitatively-focused social science students, we teach this material at twice the speed, thus compressing a year-long course with 26 weeks of classroom time to a single 13-week course. This book presents materials for one week at a time; to double the speed, just change each week to a single class (assuming two classes per week) and choose one of the two stories, one of the two class-participation activities, and so on, for each class. If your background is different, you have to think carefully about your aims. The flipped classroom described here works best with students who have similar backgrounds and interests. There are many alternative ways of how to encourage active learning, for example group work, peer grading, and asynchronous online chats.

**Applied focus.** Our book has an applied focus and substitutes computing for mathematics. If your intended learning trajectory is more theoretical, then it would make sense to use stories and computer demonstrations that are more theory-focused, for example a counterexample to linear regression where there is no unique least-squares solution<sup>5</sup> and a simulation showing the central limit theorem as decreasing discrepancies from the normal distribution as sample size increases. We focus on examples in political science and related fields. If you are adapting the course for psychology, biology, or some other subject, it would make sense to look for stories and discussion problems from these fields. The point is to keep a coherent focus so as to achieve the larger goal of developing skills and understanding.

In addition, *Regression and Other Stories* has many examples not included in this course plan that can be used in class, and any end-of-chapter exercises that are not assigned as homework can be adapted to become discussion problems or drills. Again, the particular items in the plan can be considered as a starting point. It is easier to construct a course by adapting an existing model than to create an entirely new course from scratch. What we are offering here is intended to serve as that starting point, allowing self-learners or instructors to spend less time on course planning and more time on direct interactions with students.

**Student interests.** Many of the homework assignments and in-class activities require students to pick topics of interest to them. This can work great in a statistics class in an applied field such as biology or political science, but students in a general introductory course might feel they have no specific topics of interest. When this happens, the instructor should work with them, asking what they are interested in outside of statistics. Topics could include business, sports, personal relationships, entertainment and the arts, health, . . . , just about anything. An important theme of applied statistics is the generality of the subject, and it's worth spending some time with students, individually or in

<sup>5</sup>This can be done by simulating data points  $(x, y)$  where all the  $x$ -values are equal, so that there is no leverage to estimate the slope.

## 2.8. USING THESE MATERIALS IN INTRODUCTORY OR MORE ADVANCED COURSES

23

pairs, to connect statistical ideas to some of their interests in life. You can also look at web pages such as the Data and Story Library<sup>6</sup> with data sets that are likely to have some general interest.

**Class size.** We have used the active-learning approach in classes of up to 50 students. With increasing number of students, time needed for the class activities increases, and a smaller portion of the students can be active in the discussion. For larger classes it might work to monitor pairs or small groups of students using teaching assistants and also to have weekly section meetings. With increasing number of students, resources needed to check the pre-class assignments and homeworks increases. Instead of using a teaching assistant to check these, it can be better to use peer grading to manage the flow of homeworks, and use a bigger portion of the teaching assistant resources for having more teachers per student during the class activities. During class, students should be able to work in pairs even in a large lecture hall, as long as there is room for instructors and assistants to walk around the room and keep them on task.<sup>7</sup>

We do not recommend lecturing during the class period. Material in a large passive lecture can just as well be learned from a textbook, as long as such a book is available. When adapting a textbook for a particular course, an instructor can also assign additional reading or short videos.

**Remote teaching and self-learning.** If students are not sitting in the same room together, it can be a challenge to do some of the class-participation activities described in this book. But we still think it is valuable to work in pairs, so ideally remote learners should be connected so they can work together, discuss, and ask and answer each others' questions.

This sort of collaborative learning can be done without requiring special features of a remote learning platform. Each pair or small group of students can set up their own shared online document where they can work together on problems and communicate privately while the class is going on.

### What we have here and what's still missing

We have a solid textbook, a workable modular structure for the course, and a full set of activities for student-centered learning in the classroom.

We still lack a clear set of priorities—a list of what are the most important skills to be learned, and in what order—and a path toward using these skills to understand the world. The textbook has many real-world examples, and our in-class stories provide motivation by relating technical material to real-world questions, but we remain concerned that our classes are too focused on the details of math and coding.

Consider the analogy of learning statistics to learning a new language. We tell our students that the class is particularly difficult because they're learning two foreign languages—statistics and R—both of which can be challenging without a strong background in math and programming. The challenge is to learn how to engage in and understand conversations, not merely memorize stock phrases and rules of grammar.

The bad news is that you might have only one or two semesters of formal statistics instruction. The good news is that there are many opportunities for statistics immersion, just by reading empirical papers in social science.

## 2.8 Using these materials in introductory or more advanced courses

The stories, activities, and examples in this book can be adapted to fit the interests of students and the goals of the class. We demonstrate this here using an activity we have developed based on the

<sup>6</sup><https://dasl.datadescription.com/>.

<sup>7</sup>See Rhonda Magel (1996), Increasing student participation in large introductory statistics classes, *American Statistician* 50, 51–56, and Rhonda Magel (1998), Using cooperative learning in a large introductory statistics class, *Journal of Statistics Education* 6 (3).

“two truths and a lie” game.<sup>8</sup> Before reading the discussion here, please read the description of the class-participation activity on page 179.

The activity should be fun for any group of students, but the relevant statistical lesson will depend on the level of the course being taught: it connects to several important topics, including measurement, uncertainty, prediction, calibration, and logistic regression. Because of its social aspect, it makes sense to do this during the first or second week of the semester, but it is also always important to explicitly connect classroom activities to the material being covered that week, as well as to the course as whole. We give some details about how this could go at a few different levels.

**Introductory statistics.** For an introductory course, the focus can be on probability and uncertainty. Before the activity begins, the instructor should ask students to speculate on how accurate their guesses will be. On average, will they be able to guess the lie every time? 90% of the time? 50%? More than 33%, we hope, right? This can be an opportunity to introduce the concept of a null hypothesis: from pure random guessing, the number of correct guesses would have the binomial distribution with probability  $\frac{1}{3}$ . Depending on when during the semester this activity is done, the instructor can follow up with an estimate of average probability of guessing correctly, along with a standard error, confidence interval, and hypothesis test comparing to the reference level of  $\frac{1}{3}$ . At the same time it is important to keep the larger perspective of the sampling distribution, so when presenting these results we should engage the students in discussion of how these numbers would all change if the data were different: with the given  $n$ , how many successful guesses would be needed for the null hypothesis to be rejected at the 5% level or the 95% interval for the success rate to exclude  $\frac{1}{3}$ , and so forth.

This can be connected to other problems of probability estimation such as weather forecasting or election forecasting. The instructor can display the fitted curve without going into detail on logistic regression, just giving this as an example of an advanced statistical method. For an example during the first weeks of an introductory class, the lesson here is not any particular technique, but rather the way that statistical analysis can be used to learn information from subjective certainty statements. Statistical modeling bridges between qualitative and quantitative worlds.

**Bayesian statistics.** For a course on Bayesian statistics, the activity can be used to demonstrate the principle of calibration. In this activity, the certainty judgments represent prior information but not prior *distributions*, and the step of fitting a model to predict accuracy of guesses  $y$  given certainty judgments  $x$  can be seen as a data-based construction of a prior distribution. For example, suppose the model is  $\Pr(y = 1) = \text{logit}^{-1}(-0.6 + 0.3x)$ ; this corresponds to a probability of correct guess ranging from 0.35 when  $x = 0$  to 0.92 when  $x = 10$ , and for a new guess with certainty judgment  $x$ , the value  $\text{logit}^{-1}(-0.6 + 0.3x)$  can be taken as the prior probability that this guess is correct. The point here is that priors for real problems can be calibrated based on the accuracy of past guesses; this is, for example, how point spreads for sporting events can be translated into betting odds.<sup>9</sup>

These points can be placed in the context of a class discussion via a series of prompts. As always, it is best to start the discussion *before* the data have been revealed. To start, students can consider in pairs their prior probability that a particular guess is correct: at what odds would they be willing to bet that they actually caught the lie? The next question is how this prior probability varies with  $x$ . From this they can see the certainty judgments as a device for constructing an empirical prior distribution. We can then ask how large a dataset might be needed for this prior to be useful in practice. It is easiest to get a sense of this using simulation, starting with some assumption about the function  $E(y|x)$ , trying out a sample size, simulating data, and seeing what the plot of  $E(y|x)$  vs.  $x$  looks like. The connection of all this to the class-participation activity is that, by giving the certainty statements and guesses themselves, students should get a picture of the challenges of constructing empirically-based priors.

<sup>8</sup>This discussion is taken from Andrew Gelman (2023), “Two truths and a lie” as a class-participation activity, *American Statistician* 77, 97–101.

<sup>9</sup>See Hal Stern (1997), How accurately can sports outcomes be predicted?, *Chance* 10 (4), 19–23.

## 2.8. USING THESE MATERIALS IN INTRODUCTORY OR MORE ADVANCED COURSES

25

**Generalized linear models or machine learning.** For a class on generalized linear models or machine learning, the “two truths and a lie” activity can be used as an introduction to logistic regression, showing the details of fitting and graphing a model, interpreting coefficient estimates and standard errors, and using predictions to make probabilistic forecasts for new cases. Here the activity ties directly into the material taught in the class, and after the model has been fit, graphed, and explained, there is a sequence of logical followups. Students can discuss the range of predicted probabilities: Will they always fall between 0 and 1? (Yes.) Will they always fall between  $\frac{1}{3}$  and 1? (Not necessarily.) How many measurements would be necessary for the slope of the curve to be estimated with some desired level of accuracy? You can approximately figure this out using the rule that the standard error scales like  $1/\sqrt{n}$  and can also check by simulating fake data.

Depending on the course material, this activity can be followed up in different ways. For example, a simulation can be performed to assess the statistical power of the study given sample size  $n$ , the distribution of observed certainty scores  $x$  in the observed data, and assumed values of the intercept  $a$  and slope  $b$  of the logistic regression; in R:

```
n_loop <- 1000
slope_est <- rep(NA, n_loop)
slope_se <- rep(NA, n_loop)
for (i in 1:n_loop){
  x_sim <- sample(x, n, replace=TRUE)
  y_sim <- rbinom(n, 1, invlogit(a + b*x_sim))
  sim <- data.frame(x_sim, y_sim)
  fit <- glm(y_sim ~ x_sim, family=binomial(link="logit"), data=sim)
  slope_est[i] <- summary(fit)$coefficients["x_sim", "Estimate"]
  slope_se[i] <- summary(fit)$coefficients["x_sim", "Std. Error"]
}
power <- mean(abs(slope_est)/slope_se > 2)
```

Here we have computed the power following the conventional rule that the estimated slope is statistically significant if it is more than two standard errors from zero. There is no need to perform this particular calculation; we are just illustrating how the data collected in this activity can be used as a starting point for relevant lessons.

**Psychometrics and multilevel modeling.** Another direction is to turn this into a lesson on reliability and validity of measurement. What is meant by that certainty score? How useful would you expect the certainty score to be in making a probabilistic forecast? This sort of calibration problem arises in many areas of science and policy. For example, consider a hiring setting where interviewers give numerical ratings for the candidates, and then later when there are data on job performance, the ratings can be retrospectively calibrated. One direction is to set up some comparison points, for example by asking respondents to give certainty scores for other outcomes such as weather or sporting events. There is a large literature in psychology on challenges with assessing the accuracy of subjective guesses.<sup>10</sup>

For a class on psychometrics or multilevel modeling, this discussion of measurement can serve as an entry point to the design and analysis of repeated measures data. What if the confidence of the guess is highly predictive of accuracy at the individual level, but with an effect that disappears when aggregated across guessers? Students can discuss in pairs how such a pattern can arise, if less accurate guessers tend also to be overconfident. To learn this pattern we would need to gather multiple measurements on each guesser, for example by having students make guesses and certainty statements individually rather than via consultation, and then the resulting data could be fit using a multilevel model with intercept and perhaps slope that vary by guesser. There is also clustering in the data, with four measurements per group. We have no strong reason to expect group effects (i.e., some

<sup>10</sup>See, for example, Richard Nisbett and Timothy Wilson (1977), Telling more than we can know: Verbal reports on mental processes, *Psychological Review* 84, 231–259, and Annelies Vredeveldt and James Sauer (2015), Effects of eye-closure on confidence-accuracy relations in eyewitness testimony, *Journal of Applied Research in Memory and Cognition* 4, 51–58.

groups guessing more accurately than others, more than would be expected from chance variation alone), but it would be easy to look for this by fitting a multilevel model.

**Debriefing.** A key part of any class-participation activity is what happens after the activity ends. The instructor should not just stop and leave students to ponder. It is fun when an activity has a twist, but it is not a magic trick; the point is not to amaze students but to bring them closer to the material being taught. We want the activity not to mystify but to de-mystify. So it is important to follow up the activity with explicit discussion, both of its connection to the material being taught in the class and its relevance to real-world applications of statistics in areas such as education, business, politics, or health, depending on the interests of the students.

We can also consider what lessons students might take away from this activity. “Statistics is fun”: that’s a good memory. “I got fooled by Jason’s lie: he’s not really adopted”: that’s fine too, as it serves the goal of students getting to know each other. “You can use logistic regression to convert a certainty score into a predicted probability”: that’s good because it’s a vivification of a general mathematical lesson. “The estimated slope was smaller than the standard error so we couldn’t distinguish it from zero”: that’s not a bad lesson either. The instructor should think about what memories to aim to create, and keep the discussion focused. For example, the details of the truths and lies are fun, and there could be a temptation to share some of the most successful lies with the class—but for a class on statistics or research methods, those sorts of details could be counterproductive, reinforcing memories that would distract from the statistical lessons. We want the activity to be vivid and memorable but for the right reasons.

Relatedly, having students make predictions in advance helps them confront misconceptions and be convinced to change, but it’s also possible for students to remember their initial naive guesses more than the right answer, so it’s important to have these clarifying discussions to reinforce clear thinking about the point of the activity and its larger message.

**Fitting the activity into the course as a whole.** In our experience we have seen three kinds of positive outcomes associated with this sort of activity, especially when performed near the beginning of the semester. The first is that students get used to the idea that attendance is active, not passive, and we hope the alertness required to perform these activities translates into better participation throughout the class period. The second is that people typically find data more interesting and relatable when they can see themselves in the scatterplot. The third valuable outcome is that the “two truths and a lie” activity is a social icebreaker. It is our hope that in laying out this activity—not just the general concepts but also the details of implementation, including instructions, Google form, sample data and analysis, and post-analysis discussion points—we have lowered the barrier of difficulty so that instructors in a wide range of statistics courses can try it out in their own classes, at minimal cost in classroom time and with the potential to get students more involved in their learning process.

That said, we do not have direct empirical evidence of the effectiveness of this activity on student learning. As is typically the case in education, it is easier to develop a new idea than it is to quantitatively evaluate its effects in the classroom.<sup>11</sup> We can still learn from experience, but such learning tends to be qualitative, from observing student reactions and discussions. We do not expect miracles from these activities. We are satisfied if they motivate students to participate in class and think about each week’s material.

The biggest risk or opportunity cost in introducing a new class-participation activity is that time spent in the activity could be spent working on lecturing or problem solving. For this reason, the activity should be closely tied to the course material and performed efficiently, with instructions and Google form prepared ahead of time and with code all set up to analyze the data when they come in. Instructors can also use the material in this book as a template for designing and implementing their own class-participation activities.

<sup>11</sup>For some discussions of the difficulties of evaluating teaching innovations in statistics, see Beth Chance, Joan Garfield, Elsa Medina, and Dani Ben-Zvi (2008), Assessment in statistics education, in *Developing Students’ Statistical Reasoning: Connecting Research and Teaching Practice*, edited by Joan Garfield and Dani Ben-Zvi, and Andrew Gelman and Eric Loken (2012), Statisticians: When we teach, we don’t practice what we preach, *Chance* 25 (1), 47–48.

## 2.9 Balance between challenges and solutions

In the first week of class we tell the story of the 1936 *Literary Digest* pre-election poll from 1936, a famous example of a large prestigious survey that was done without using random sampling and which yielded really bad inference; see the story on page 36 of this book. This episode is traditionally presented in statistics classes as a cautionary tale: the primary message here is to use random sampling and the secondary message is to not trust surveys that don't use random sampling. These are not terrible messages—but statisticians Sharon Lohr and J. Michael Brick took the next step and found that, using information that was available at the time, it is possible to adjust the survey to get an improved, although not perfect, estimate.

This example illustrates the balance we seek to strike between challenges and solutions. Statistics is hard. It should not be tricky. Our stories and activities often include twists that reveal hidden assumptions. In the *Literary Digest* polling story, there was a mistaken mathematical assumption that a large sample will automatically be representative of the population, and a mistaken sociological assumption that a well-publicized project would be done well. Recognizing these assumptions gets you thinking, if a large sample size is not enough, what is actually necessary for a survey to be accurate?

A quick answer is that random sampling will do the job. But we do not want to stop there, for two reasons. First, real-world surveys of people are not random samples, and we do not want to send the unrealistic message that samples must be random to be useful. Yes, the *Literary Digest* survey had big problems, but modern political surveys are very accurate, despite response rates below 10% and respondents who do not look like a random sample of the population.<sup>12</sup>

The second reason for not ending the story in failure is that statistical methods in general, and regression modeling in particular, are valuable in large part by bridging between data and underlying models of the world. The mathematics of random sampling, with its unbiased estimates and standard errors, is one such bridge. Regression adjustment and poststratification for nonrepresentative samples is another bridge. Survey adjustment, is an important practical tool in statistics as well as being an application of linear regression. Ending the story with adjustment is a way to pull back the curtain and demonstrate that statistics is difficult but not tricky: the adjustment steps are transparent and intuitive. That said, we do not want the solution to seem *too* clean, as it is still imperfect and depends on strong, false, assumptions. This is part of the balance required in any science teaching, as we go back and forth between using models and questioning them.

<sup>12</sup>For more background, see Andrew Gelman (2021), Failure and success in political polling and election forecasting, *Statistics and Public Policy* 8, 67–72.

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

## Part 2: Stories, activities, problems, and demonstrations

---

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

## Chapter 3

# Week by week: the first semester

Our course has two 75-minute classes per week; here we share two sets of readings and homework, two stories, two class-participation activities, two computer demonstrations, and so forth, for each of the 13 weeks of the first semester. Chapter 4 has the corresponding material for the second semester.

### 3.1 Introduction to quantitative social science

#### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Wikipedia experiment	Design a study to measure some quantity of interest	Collect and analyze simulated data	Design a study to estimate a causal effect	Find the hidden assumption
Literary Digest poll of 1936	Design an experiment to distinguish between two hypotheses	Predict elections from economy	Generalize	Find the hidden assumption

#### Reading

Chapter 1 of *Regression and Other Stories*: Overview

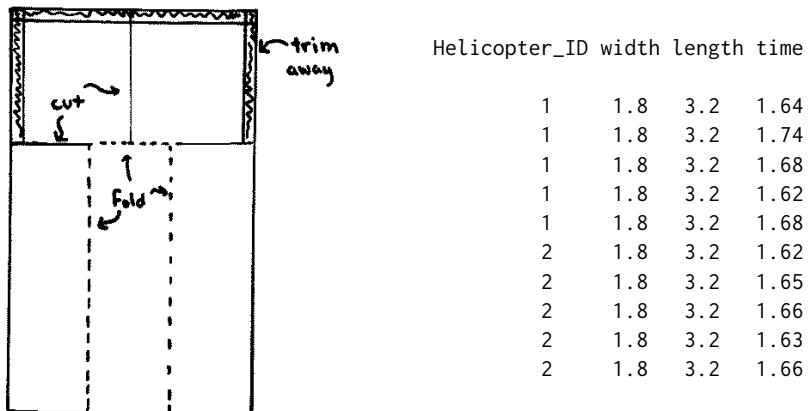
#### Pre-class warmup assignments

1. *No assignment before the first class.*
2. Overview of statistics
  - (a) What are the three challenges of statistical inference?
  - (b) What are the key skills you should learn from the textbook?
  - (c) What is the URL of the website with data and examples from the textbook?

#### Homework assignments

1. *No homework assignment due for the first class.*
2. From design to decision (Exercise 1.1 of *Regression and Other Stories*)

Figure 1 displays the prototype for a paper “helicopter.” The goal of this assignment is to design a helicopter that takes as long as possible to reach the floor when dropped from a fixed height, for example 8 feet. The helicopters are restricted to have the general form shown in the sketch. No



**Figure 1** (a) Diagram for making a “helicopter” from half a sheet of paper and a paper clip. The long segments on the left and right are folded toward the middle, and the resulting long 3-ply strip is held together by a paper clip. One of the two segments at the top is folded forward and the other backward. The helicopter spins in the air when dropped. (b) Example data file showing flight times, in seconds, for 5 flights each of two identical helicopters with wing width 1.8 inches and wing length 3.2 inches dropped from a height of approximately 8 feet.

additional folds, creases, or perforations are allowed. The wing length and the wing width of the helicopter are the only two design parameters, that is, the only two aspects of the helicopter that can be changed. The body width and length must remain the same for all helicopters. A metal paper clip is attached to the bottom of the helicopter.

Here are some comments from previous students who were given this assignment:

Rich creased the wings too much and the helicopters dropped like a rock, turned upside down, turned sideways, etc.

Helis seem to react very positively to added length. Too much width seems to make the helis unstable. They flip-flop during flight.

Andy proposes to use an index card to make a template for folding the base into thirds.

After practicing, we decided to switch jobs. It worked better with Yee timing and John dropping. 3 – 2 – 1 – GO.

Each group of students will need 25 half-sheets of paper and 2 paper clips. The body width will be one-third of the width of the sheets, so the wing width can be anywhere from  $\frac{1}{6}$  to  $\frac{1}{2}$  of the body width; see Figure 1a. The body length will be specified by the instructor. For example, if the sheets are U.S.-sized ( $8.5 \times 5.5$  inches) and the body length is set to 3 inches, then the wing width could be anywhere from 0.91 to 2.75 inches and the wing length could be anywhere from 0 to 5.5 inches. In this assignment you can experiment using your 25 half-sheets and 2 paper clips. You can make each half-sheet into only one helicopter. But you are allowed to design sequentially, setting the wing width and body length for each helicopter given the data you have already recorded. Take a few measurements using each helicopter, each time dropping it from the required height and timing how long it takes to land.

- Record the wing width and body length for each of your 25 helicopters along with your time measurements, all in a file in which each observation is in its own row, following the pattern of `helicopters.txt` in the folder `Helicopters`, also shown in Figure 1b.<sup>1</sup>
- Graph your data in a way that seems reasonable to you.

<sup>1</sup>Data for examples and assignments are at <http://www.stat.columbia.edu/~gelman/regression/>.

- (c) Given your results, propose a design (wing width and length) that you think will maximize the helicopter’s expected time aloft. It is not necessary for you to fit a formal regression model here, but you should think about the general concerns of regression.

## Stories

### 1. Wikipedia experiment

We were contacted by two people working for the Wikimedia Foundation, the organization that funds Wikipedia.<sup>2</sup> They had supervised a large number of online experiments in an attempt to increase the rate of donations to the project. Figure 2 shows an example of a manipulation of the request screen: the “control” or existing condition used a box with rounded corners, and the “treatment” or alternative used square corners. The experiment was performed to see if this treatment had any effect on donations.

This is an example of what is called “randomized experimentation” in statistics or “A/B testing” in industry—here, option A was the control condition and option B was the new version with rounded corners, with each potential donor randomly assigned to one of the two options. The idea is that if the treatment has a large enough effect, it can be detected experimentally, with the random assignment ensuring approximate comparability of the groups receiving the two conditions. A large set of such tests can potentially reveal which of various proposed small improvements would increase the donation rate.

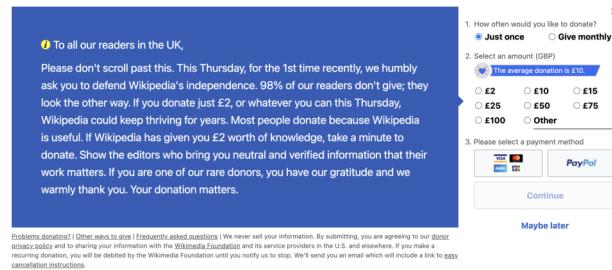
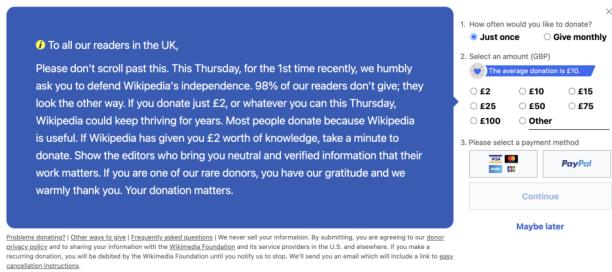
The Wikimedia team had contacted us because some of the effects they’d estimated were implausibly large. As an example, they shared the screenshot shown in Figure 3 giving the result from the experiment comparing square to round corners.

Reading Figure 3 from the bottom up: In the experiment, donations were 15% lower in the square-corners condition than in the control condition (an estimated effect of  $-15\%$ ), and the data are compatible with effects in the range  $[-19\%, -11\%]$ —that is the 95% confidence interval shown in the lower right of the display. Moving up, the success rate was 0.51% for the controls and 0.43% for the square corners, with the relative effect size being estimated at  $(0.0043 - 0.0051)/0.0051 = -0.15$  or  $-15\%$ . And the top of the display shows the raw data: 4861 successes out of 954 630 trials for the control group and 4695 successes out of 1 082 180 trials for the treatment group, an impressive sample size given that the experiment only ran for 12 hours. Wikipedia has a lot of users!

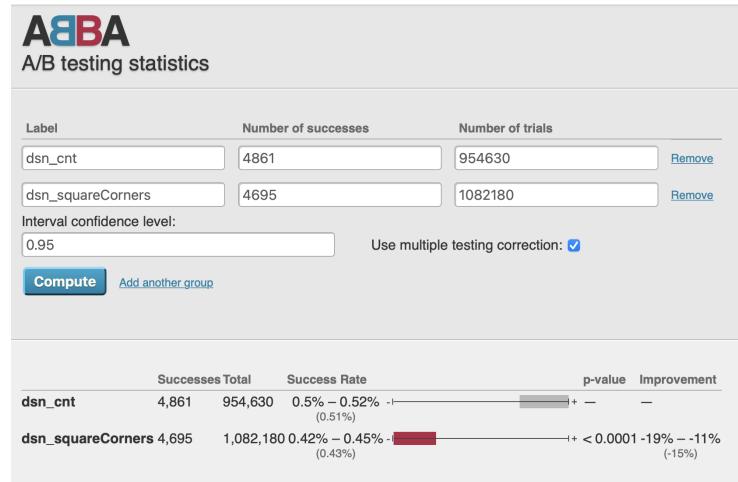
Is it possible the large estimated effect of 15% was just due to chance? No, as we can see because the 95% confidence interval is well away from zero. The sample size of two million trials is huge, and we would not see a difference of 15% just by chance. We can also see this using a quick calculation: if the true success rate is 0.0051, then the standard deviation of the estimated success rate from a million trials is  $\sqrt{0.0051 * 0.9949 / 1\,000\,000} = 0.00007$ , and so the standard deviation of the difference between two success rates is  $\sqrt{0.00007^2 + 0.00007^2} = 0.0001$ . This is an absolute difference in proportions. To convert this to a percentage scale, divide by the baseline of 0.0051 to get  $0.0001 / 0.0051 = 0.02$ , or 2%. A 95% interval is an estimate  $\pm 2$  standard errors, hence in this case  $[-15\% \pm 2 * 2\%] = [-19\%, -11\%]$ , which is indeed what we saw. This is a standard calculation which shows that with this sample size we would not expect to see such a large effect by chance alone.

When telling the story, we perform the above calculations on the blackboard without slowing down and trying to explain each step. The point here is not to teach inference for proportions—this will be covered a few weeks later in the course—but rather to model the use of statistical analysis in understanding a real-world problem.

<sup>2</sup>[https://en.wikipedia.org/wiki/Wikimedia\\_Foundation](https://en.wikipedia.org/wiki/Wikimedia_Foundation).



**Figure 2** Wikipedia messaging: control (round corners) and treatment (square corners) were compared to see which led to higher donation rates. This was one of many experiments (“A/B tests”) done by Wikipedia, notable because of the surprisingly large observed difference; see Figure 3.



**Figure 3** Output from the Wikipedia experiment. The display shows the estimated effect of the new treatment (square corners on the donation box) compared to the control condition (the status quo of round corners). The treatment is estimated to decrease donation rates by an implausibly large amount of 15%. Followup is in Figure 4.

Continuing with the narrative: Having established that the observed difference in success rates between the treatment and control group could not plausibly have arisen by chance, we suggested that perhaps the effect in the range  $[-19\%, -11\%]$  was real but not stable. Perhaps square corners could have an effect of  $-15\%$  on weekdays, when many people are accessing Wikipedia from work or school, and a positive effect of  $+10\%$ , say, on weekends. Maybe square corners are perceived as friendlier in some countries than in others. Or the effect could vary with other contextual factors, such as what was in the news? An effect in a single 12-hour experiment might not be representative of general conditions.

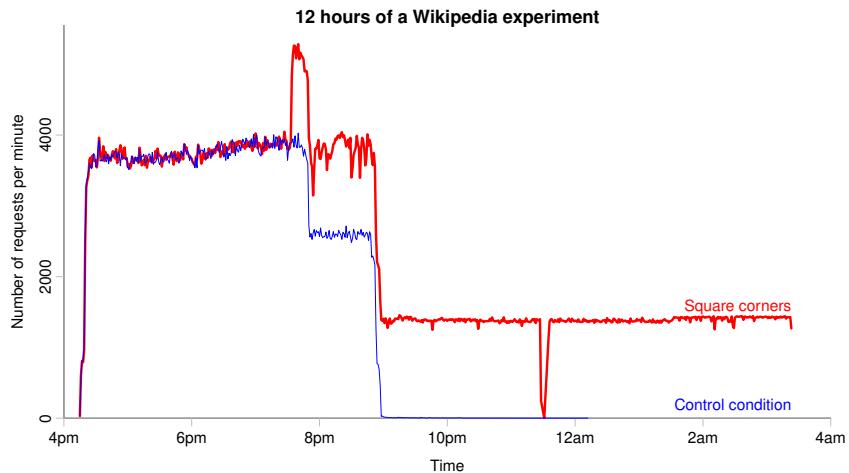


Figure 4 Further information from the Wikipedia experiment, following up on Figure 3. After about 7 pm, there were big differences in the number of people given the treatment and control, suggesting there was some flaw in the randomization.

Our correspondents were doubtful. They said that the proportion of solicitations that yield a donation varies a lot—for example, it is higher in days and evenings than in the middle of the night—but that in their history of A/B tests, they haven’t found the effects of any particular treatment to vary so much. They were skeptical about my idea that the effect that could be  $-15\%$  at some times and  $+10\%$  in other settings.

So we decided to look more carefully at Figure 3. And we noticed something. In a simple randomized experiment, you expect to see approximately 50% of cases in the control group and 50% in the treatment group. Not exactly an even split—if you flip a coin  $N$  times, it’s unlikely you’ll see exactly  $N/2$  heads and  $N/2$  tails—but close. In this experiment, however, there are 950 000 control trials and 1 080 000 with the treatment. That sort of split can’t just happen by chance.

To do a quick calculation: the proportion of trials in the treatment group is  $1 082 180 / (954 630 + 1 082 180) = 0.53$ , or 53%. If we flip 2 million coins, the proportion of heads is a random variable with mean 0.5 and standard deviation  $\sqrt{0.5 * 0.5 / 2 000 000} = 0.0004$ , or 0.04%. Using the  $\pm 2$  standard deviation rule, we’d expect the percentage of heads to be somewhere in the range  $[50\% \pm 2 * 0.04\%] = [49.92\%, 50.08\%]$ . No way we’d see anything like 53%. So something seemed wrong. Our guess was that not all the data from the experiment had been recorded, perhaps some of the data from the control group got lost somewhere in the pipeline?

The Wikipedia team got back to us and shared the data plotted in Figure 4, which shows the number of trials during each minute for each of the two conditions during the period of the experiment. Something happened between 7 and 8 pm; we’re not sure what, but all of a sudden the treatment group was consistently getting more action than the control group.

And this in turn explains the different success rates! Remember how they said that conversion rates were higher during the daytime and early evening than during the middle of the night? Due to the botched randomization, the control condition was mostly applied during the daytime and the early evening, whereas the square-corners intervention had relatively more nighttime exposure. Hence the observed drop.

This story relates to the week’s reading on the goals of statistics: generalizing from sample to population, from control to treatment, and from observed data to underlying constructs of interest. Here, the people in the study are the sample and Wikipedia users in general are the population, the

control is round corners and the treatment is square corners, and the observed data are whether these particular people donated. The underlying goal is . . . perhaps to get more people to give, or to get existing donors to give more? It's good to think about these questions even if we don't have clear answers.

The Wikipedia story also illustrates some recurring themes of the course regarding data collection and data exploration, the importance of checking the assumptions underlying our inferences, and the role of probability models and mathematical calculations (in this case, the standard error of estimated proportions).

## 2. Literary Digest poll of 1936

Statisticians Sharon Lohr and J. Michael Brick write:<sup>3</sup>

"The *Literary Digest* poll of 1936 is a byword for bad survey research. Textbooks have long used it as a prime example of how sampling goes bad . . .

The story of the 1936 poll is well known. Ten million ballots were sent out . . . The mailing list was 'drawn from every telephone book in the United States, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail-order and occupational data.' Of these, almost 2.4 million ballots were returned. The final results from the poll predicted that Republican Alfred Landon would win with 54 percent of the popular vote and 370 electoral votes. In the election, Franklin Roosevelt won more than 60 percent of the popular vote and 523 electoral votes, carrying every state except Maine and Vermont.

Gallup (1938) blamed the poll's inaccuracy on the sampling frame, which was constructed largely from lists of telephone and automobile owners and thus overrepresented the well-to-do. . . . Other postmortems . . . pointed to nonresponse as a primary factor in the poll's inaccuracy, relying on a survey taken by Gallup in 1937 that asked respondents whether they had participated in the 1936 LD poll. These analyses concluded that telephone and automobile owners both supported Roosevelt (though not to the same extent as persons without telephones and automobiles), and that the low response rate combined with the flawed sample produced the inaccurate forecast . . .

But the 24 percent response rate of the 1936 LD poll is much higher than the response rate in many of today's polls. One difference is that today's polls weight the data to attempt to compensate for nonresponse bias. Typically, the weights of the respondents are adjusted so that weighted estimates match the voting population demographics; some polls also weight by political party.

Demographic weighting could not have been used for the 1936 LD poll because those data were not collected from the respondents. But the ballot [see Figure 5] did ask respondents for whom they voted in the 1932 election. This information could have been used to weight the data using techniques that were known at that time and involved simple computations. . . ."

And here's what Lohr and Brick found:

"If information collected by the poll about votes cast in 1932 had been used to weight the results, the poll would have predicted a majority of electoral votes for Roosevelt in 1936, and thus would have correctly predicted the winner of the election. We explore alternative weighting methods for the 1936 poll and the models that support them. While weighting would have resulted in Roosevelt being projected as the winner, the bias in the estimates is still very large."

<sup>3</sup>Sharon Lohr and J. Michael Brick (2017), Roosevelt predicted to win: Revisiting the 1936 Literary Digest poll, *Statistics, Politics and Policy* 8, 65–84. For further discussion, see Andrew Gelman (2021), The Xbox before its time: Using the famous 1936 Literary Digest survey as a positive example of statistical adjustment rather than a negative example of non-probability sampling, <https://statmodeling.stat.columbia.edu/2021/07/16/literary-digest/>.

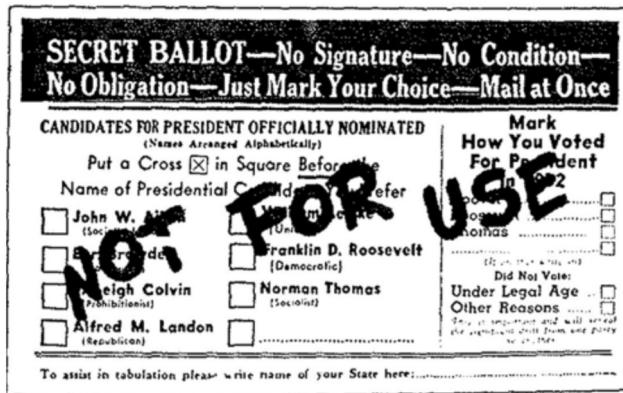


Figure 5 Ballot from the 1936 Literary Digest poll. A better estimate could have been obtained by using the “Mark How You Voted For President in 1932” question to adjust the survey results.

This provides an important lesson for statistics learning. The usual way the *Literary Digest* poll is used in statistics textbooks is as a warning: Use random sampling, or the boogeyman will get you! That’s pretty silly, actually, given that all polls use non-random sampling if you consider nonresponse as part of the sampling procedure, which in effect it is. We much prefer the positive message: All samples of humans are flawed, but we can do our best to adjust for those flaws, while also recognizing the imperfections of our adjustments.

This example relates to two ideas in the week’s reading: first, the problem of generalizing from sample to population; second, the use of regression modeling to adjust for differences between sample and population.

### Class-participation activities

#### 1. Design a social science study to measure some quantity of interest

For the very first class-participation activity, we have students work together to design a hypothetical study involving social measurement. They start by discussing possible research topics in social science: each pair of students should come up with a proposed topic, then a few students share their ideas, and then the entire class decides on a topic. We then take them through similar steps to choose a quantity to be measured. Possible examples include political polarization, gentrification, life satisfaction, the spread of misinformation, friendship networks, or all sorts of other things.

Once the quantity has been picked, the students can work in pairs to consider how to measure it. To do so they will need to define the underlying quantity of interest, consider what data they could gather to construct a measure, decide how they would collect the data, make a plan for estimating the measure given the data, and then come up with specifics such as sample size and details of the measurement. Finally, once we have all that, we can discuss how to measure variation over time or across states and countries.

While the class is having this conversation, we write the items in Figure 6 on the board, one at a time, to structure the discussion.

As always, this activity should be concluded by tying it to the reading. Chapter 1 of *Regression and Other Stories* discusses the three tasks of statistics: generalizing from sample to population, from control to treatment group, and from observed data to underlying construct of interest. This activity focuses on the third of these generalizations, a topic that is often neglected in statistics classes. Pointing toward topics to be covered later in the course, the instructor can also discuss the use of simulation to explore possible designs of data collection.

1. Topic
2. Quantity to measure
3. Definition
4. Data
5. Data collection
6. Estimation
7. Specifics
8. Variation

**Figure 6** Steps for the class-participation activity designing a study to measure a quantity of interest in social science. As each step is introduced and discussed, the instructor should write it on the board, and students can then follow along as they work on the activity.

1. Topic
2. Two hypotheses
3. Ideal data that would establish hypothesis 1 or 2
4. Scenario of ambiguous data
5. Scenario of data consistent with neither hypothesis
6. Data collection and measurement
7. Inference
8. Specifics

**Figure 7** Steps for the class-participation activity designing an experiment to distinguish between two hypotheses in social science. As each step is introduced and discussed, the instructor should write it on the board, and students can then follow along as they work on the activity.

## 2. Design an experiment to distinguish between two social science hypotheses of interest

The first week of our course introduces statistics in the context of quantitative social science, and in the student-participation activity for the second class of that week we lead a collaborative design of a hypothetical experiment. We start by having students work in pairs to discuss possible topic areas and hypotheses. For example, in the problem of encouraging low-income villagers to use bed nets to reduce malaria risk, one hypothesis is that it is better to give the nets for free to increase availability, while a competing hypothesis is that selling them for a small amount of money will encourage middlemen to promote their use. For another example, there is a hypothesis in education that a technique called mindset interventions will increase student performance, and a competing hypothesis that this intervention distracts from learning. Another example in education is the controversy over the importance of peer effects: what educational benefits, if any, do children get from being in the same class as high-performing pupils? If students are stuck coming up with a topic and hypotheses, the instructor can give one or two of these examples, but in our experience it is not difficult for students to come up with interesting ideas, so all that is needed is to keep the discussion focused on a specific research question.

Once the two hypotheses have been chosen, the instructor can lead a discussion of how to use data to distinguish between them. The first step is to consider ideal data that would establish one hypothesis or the other, then equivocal data which would be consistent with both hypotheses, and, from the other direction, hypothetical data that would be consistent with neither hypothesis. The next step is to work with the class to come up with a plan for data collection and measurement, inference (how to use the data to distinguish between the hypothesis), and specifics of sampling,

1. Reading
2. Homework
3. Feedback sheet
4. Classes
5. Final exam

**Figure 8** Components of the course. The instructor should project these onto the screen or write them on the board on the first day of class.

treatment assignment, and measurement. While this discussion is going on, the instructor can write the items in Figure 7 on the board, one at a time, to structure the conversation.

This activity connects to the reading in Chapter 1 of *Regression and Other Stories* on the use of statistics to learn about scientific theories. In this case the topic is experimentation, which involves generalizing from control to treatment group, so it is a good time for the instructor to explicitly introduce some ideas of causal inference and potential outcomes (what would happen if X were done instead of Y, and how that could be measured), topics that recur throughout the course and to which we return more formally near the end of the second semester.

### Discussion of reading and homework

1. For the first day of class, the instructor should discuss the course plan (see Section 1.2):
  - (a) Goals of the course
  - (b) Components of the course; discuss Figure 8
  - (c) Structure of each class period; discuss Figure 9
  - (d) Students' responsibilities
  - (e) Roles of mathematics, computing, and applications
  - (f) Also discuss questions from final exam
2. For later classes, this is a time to discuss issues that students raised in the shared document for that class; see Section 1.3 of this book. For example, here are some issues that came up for the second day of class when we were teaching:
  - (a) Helicopter homework assignment
  - (b) Bayesian aspect of the course
  - (c) What sample size is necessary? Different sources of error
  - (d) United Nations vs. gun control studies: importance of asking a focused question
  - (e) Questions about R
  - (f) Selection bias in what gets reported: Why does it say in the textbook that the estimate in the Jamaica study wouldn't have been published if it had not been at least 2 standard errors from zero?
  - (g) Jamaica experiment: If that study really is too noisy to be useful, then what should be done? Increase the sample size? Just give up?
  - (h) Xbox study: At what point does a convenience sample become too far removed from the desired sample to no longer be considered effective?
  - (i) How do social scientists minimize selection bias in a study where you can't just experimentally assign treatments?
  - (j) Do you always need both a treatment and a control group?

1. Story
2. Activity
3. Discussion of reading and homework
4. Computer demo
5. Drill
6. Discussion problem

**Figure 9** Structure of each class period. The instructor should project these onto the screen or write them on the board on the first day of class.

## Computer demonstrations

### 1. Collect and analyze simulated data

The computer demonstrations at the beginning of the semester are intended to introduce R and the idea of working in a script-based, rather than point-and-click, computing environment, and also to demonstrate how we use simulation to understand statistical data collection and analysis.

For the first computer demonstration of the course, we constructed a fairly elaborate simulation that the instructor, or student working in self study, can perform in real time. At this point as a student at the beginning of the course we do not expect you to follow the details of the code or the underlying statistical reasoning; rather, the point of the demonstration is to show directly how computing can be used to simulate statistical data collection, visualization, and analysis. In this particular demonstration, we simulate before-after data, then fit and display a linear regression, then simulate an experiment with constant treatment effect, fit that linear regression, and display it.

This demonstration connects to the week's reading, which presents an overview of applied statistics. This example shows a randomized experiment such as discussed in Section 1.3 of *Regression and Other Stories*. As with all the computer demonstrations, we recommend typing the code directly into the text editor (not cutting and pasting) and saying the code aloud while entering it, as this more closely models the process of problem solving on the computer. When copying the code into the console to be executed by R (or whatever software is being used), the instructor has another chance to explain what the code is doing. And, as always, if the demonstration below seems too long, just go through part of it.

```
# Load package
library("rstanarm")

# Generate 50 data points
N <- 50
midterm <- runif(N, 0, 100)
a <- 20
b <- 0.5
error <- rnorm(N, 0, 5)
final <- a + b*midterm + error
fake <- data.frame(midterm, final)

# Plot data
plot(midterm, final)

# Fit regression
fit <- stan_glm(final ~ midterm, data=fake, refresh=0)
print(fit)
```

### 3.1. INTRODUCTION TO QUANTITATIVE SOCIAL SCIENCE

41

```
# Add fitted regression line to data
a_hat <- coef(fit)[1]
b_hat <- coef(fit)[2]
curve(a_hat + b_hat*x, add=TRUE)

# Add a treatment effect
treatment <- sample(c(0,1), N, replace=TRUE)
theta <- 10
final <- a + b*midterm + theta*treatment + error
fake <- data.frame(midterm, treatment, final)

# Plot data
plot(midterm, final, type="n")
points(midterm[treatment==0], final[treatment==0], col="blue")
points(midterm[treatment==1], final[treatment==1], col="red")

# Fit regression
fit_2 <- stan_glm(final ~ treatment + midterm, data=fake, refresh=0)
print(fit_2)

# Add fitted regression line to data
a_hat <- coef(fit_2)[1]
theta_hat <- coef(fit_2)[2]
b_hat <- coef(fit_2)[3]
curve(a_hat + b_hat*x, add=TRUE, col="blue")
curve(a_hat + theta_hat + b_hat*x, add=TRUE, col="red")
```

## 2. Regression predicting election outcome from the economy

For this demonstration, we go through the steps of graphing data and understanding a fitted regression.<sup>4</sup> There are no surprises here; it's just a straightforward example of a basic and important data analysis procedure.

```
# Load package
library("rstanarm")

# Read data from here: https://github.com/avehtari/ROS-Examples
hibbs <- read.table(paste0("https://raw.githubusercontent.com/avehtari/", 
    "ROS-Examples/master/ElectionsEconomy/data/hibbs.dat"), header=TRUE)

# Plot the data
plot(hibbs$growth, hibbs$vote,
      xlab="Average recent growth in personal income",
      ylab="Incumbent party's vote share")

# Estimate regression and display fitted model
M1 <- stan_glm(vote ~ growth, data=hibbs, refresh=0)
print(M1)

# Add fitted line to graph
abline(coef(M1), col="gray")

# Predict for a new election where growth = 2.0
prediction <- coef(M1)["(Intercept)"] + coef(M1)[["growth"]] * 2
```

<sup>4</sup>Example from Section 1.2 of *Regression and Other Stories*.

1. Treatments
2. Population
3. Sample
4. Treatment assignment
5. Pre-test measurement
6. Outcome measurement

Figure 10 *Steps for the drill on designing a study. These should be projected onto the screen in preparation for the drill so that students can follow these steps while answering the drill questions.*

## Drills

### 1. Design a study

Below is a series of research questions regarding estimation of causal effects. For each, the exercise is to suggest a hypothetical experiment and then fill in the details by going through the following steps listed in Figure 10: decide on the treatments, population, sample, treatment assignment, pre-test measurement, and outcome measurement. There will be no single correct answer.

In doing this drill, you should not overthink each problem. The goal is to be able to go quickly through the six steps of Figure 10.

(a) Estimate the effect of a tennis class.

*Solution:* For example, (1) Treatment is a particular tennis class, control is no class; (2) Population is all adults who are inexperienced tennis players and are interested in the sport; (3) Sample is the group of people who volunteer to participate in the experiment; (4) Treatments are assigned by flipping a coin; (5) Pre-test measurement is a tennis skill test; (5) Post-test measurement is the same skills test.

(b) Estimate the effect of doubling the number of polling places.

(c) Estimate the effect of campaign contributions.

(d) Estimate the effect of the flipped classroom.

(e) Estimate the effect of remote teaching.

Each of these examples is, in statistics jargon, a *controlled experiment*. Causal effects can also be estimated from *observational studies* (where treatments or exposures are not assigned by the experimenter), and, as discussed in Chapter 1 of *Regression and Other Stories*, there are also non-causal studies where the goal is measurement, prediction, or discovery.

### 2. Generalize

Below is a series of hypothetical social science studies. For each, the exercise is to state how it is generalizing from sample to population, from treatment to control group, and from measurement to underlying construct. See Figure 11.

(a) A survey of political participation.

*Solution:* (1) Sample is people in the study, population is all potential voters in the country; (2) No treatment or control groups for this problem; (3) Measurement is how participation is recorded (for example a survey response or checking the voter registration records), underlying construct is actual participation (for example, did you actually vote).

(b) An experiment measuring the effect on political attitudes of watching a TV news show.

(c) A study of time trends of voter turnout in several countries.

(d) A focus group on attitudes toward military intervention in Haiti.

1. From sample to population
2. From treatment to control group
3. From measurement to underlying construct

Figure 11 *Steps for the drill on generalizing. These should be projected onto the screen in preparation for the drill so that students can follow these steps while answering the drill questions.*

- (e) An experiment measuring patterns of social interaction among a group of students playing a strategy game.

### Discussion problems

1. Find the hidden assumption and error in a naive empirical claim

Assume that all the facts and assertions in the paragraph below are correct. Why do the conclusions not follow? (This does not mean that the conclusions are actually false.) What are alternative explanations for the facts? What data could be gathered to confirm or disconfirm these explanations? This is not a trick question.<sup>5</sup>

“Many theorists claim that domestic instability tends to lead to foreign aggression. Others have made the claim that domestic instability makes it less likely that a country will engage in an aggressive foreign policy. The posited linkages are obvious. Suppose you develop a good measure of both variables, and for each year you compute the total amount of domestic instability in all countries in the international system and correlate this with the total amount of external aggression by all countries. You find no correlation at all and conclude that, contrary to both theories, there is no connection between domestic instability and war.”

2. Find the hidden assumption and error in a naive empirical claim

Assume that all the facts and assertions in the paragraph below are correct. Why do the conclusions not follow? (This does not mean that the conclusions are actually false.) What are alternative explanations for the facts? What data could be gathered to confirm or disconfirm these explanations? This is not a trick question.<sup>5</sup>

“There is a positive correlation between the per capita GDP of a country and the degree to which it is democratic. Therefore as poor countries get richer, they will also become more democratic.”

<sup>5</sup>From Robert Jervis; see Andrew Gelman (2021), Statistical fallacies as they arise in political science (from Bob Jervis), <https://statmodeling.stat.columbia.edu/2021/03/03/fallacies-jervis/>.

## 3.2 Prediction as a unifying theme in statistics and causal inference

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
United Nations peacekeeping	Bag of candies and sampling bias	Graph of data and fitted line	Describe a regression slope	Height and earnings
Girls and sports	Gather and plot data from students	Tinker with an example	Simple coding	Graph hypothetical data

### Reading

Appendix A: Computing in R, Sections A.1–A.3

### Pre-class warmup assignments

#### 1. Simple coding: sequences and sampling

For each question below, give R code.

- Generate and print the sequence  $(0, 0.2, 0.4, \dots, 20)$ .
- Generate and print one random number, uniformly distributed between 0 and 10.
- Generate and print ten random numbers, uniformly distributed between 0 and 10.

#### 2. Simple coding: sampling, looping, and vectors

For each question below, give R code.

- Write a loop to generate and print 10 uniformly-distributed numbers, five times.
- Create a vector  $a$  that contains the names of four celebrities.
- Pick one of the names from vector  $a$  with equal probability.
- Repeat with unequal probabilities by setting up a vector  $p$  that specifies the probabilities, and including it in your sample function.

### Homework assignments

#### 1. (a) Sketching a regression model and data (Exercise 1.2 of *Regression and Other Stories*)

Figure 1.1b of *Regression and Other Stories* shows data corresponding to the fitted line  $y = 46.3 + 3.0x$  with residual standard deviation 3.9, and values of  $x$  ranging roughly from 0 to 4%.

- Sketch hypothetical data with the same range of  $x$  but corresponding to the line  $y = 30+10x$  with residual standard deviation 3.9.
- Sketch hypothetical data with the same range of  $x$  but corresponding to the line  $y = 30+10x$  with residual standard deviation 10.

#### (b) Goals of regression (Exercise 1.3 of *Regression and Other Stories*)

Download some data on a topic of interest to you. Without graphing the data or performing any statistical analysis, discuss how you might use these data to do the following things:

- Fit a regression to estimate a relationship of interest.
- Use regression to adjust for differences between treatment and control groups.
- Use a regression to make predictions.

### 3.2. PREDICTION AS A UNIFYING THEME IN STATISTICS AND CAUSAL INFERENCE

45

#### (c) Problems of statistics (Exercise 1.4 of *Regression and Other Stories*)

Give examples of applied statistics problems of interest to you with challenges in:

- i. Generalizing from sample to population.
- ii. Generalizing from treatment to control group.
- iii. Generalizing from observed measurements to the underlying constructs of interest.

Explain your answers.

#### 2. (a) Goals of regression (Exercise 1.5 of *Regression and Other Stories*)

Give examples of applied statistics problems of interest to you in which the goals are:

- i. Forecasting/classification.
- ii. Exploring associations.
- iii. Extrapolation.
- iv. Causal inference.

Explain your answers.

#### (b) Causal inference (Exercise 1.6 of *Regression and Other Stories*)

Find a real-world example of interest with a treatment group, control group, a pre-treatment predictor, and a post-treatment predictor. Make a graph like Figure 1.8 using the data from this example.

#### (c) In pairs: Working through your own example (Exercise 1.10 of *Regression and Other Stories*)

Download or collect some data on a topic of interest to you. You can use this example to work though the concepts and methods covered in the book, so the example should be worth your time and should have some complexity. This assignment continues throughout the book as the final exercise of each chapter. For this first exercise, discuss your applied goals in studying this example and how the data can address these goals.

## Stories

### 1. United Nations peacekeeping

Several years ago, political scientist Page Fortna conducted a study on the effectiveness of international peacekeeping. She analyzed data from countries that had been involved in civil wars between 1989 and 1999, comparing countries with and without United Nations peacekeeping. The outcome measure was whether there was a return to civil war in the country and, if so, the length of time until that happened. Data collection ended in 2004, so any countries that had not returned to civil war by the end of that year were characterized as being still at peace. The subset of the data summarized here contains 96 ceasefires, corresponding to 64 different wars.<sup>6</sup>

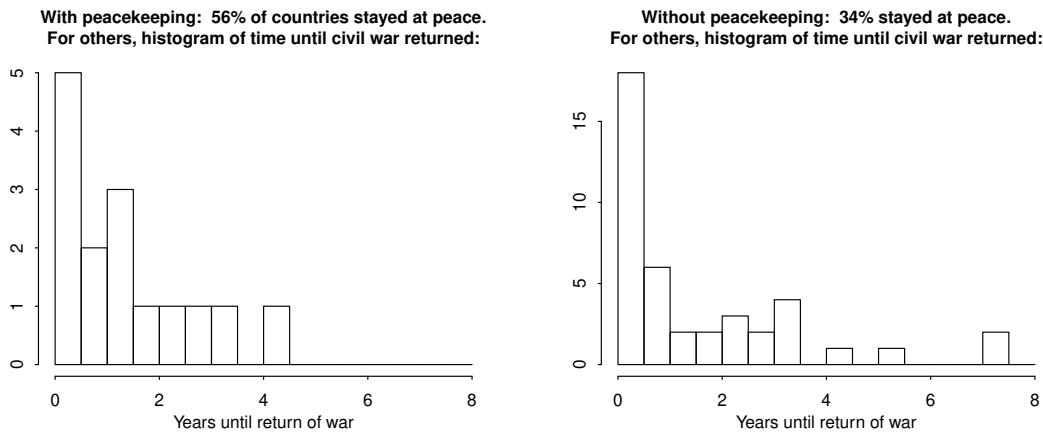
A quick comparison found better outcomes after peacekeeping: 56% stayed at peace, compared to 34% of countries without peacekeeping. When civil war did return, it typically came soon: the average lag between ceasefire and revival of the fighting was 17 months in the presence of peacekeeping and 18 months without. Figure 12 shows the results.

There is, however, a concern about *selection bias*: perhaps peacekeepers chose the easy cases. Maybe the really bad civil wars were so dangerous that peacekeepers didn't go to those places, which would explain the difference in outcomes.

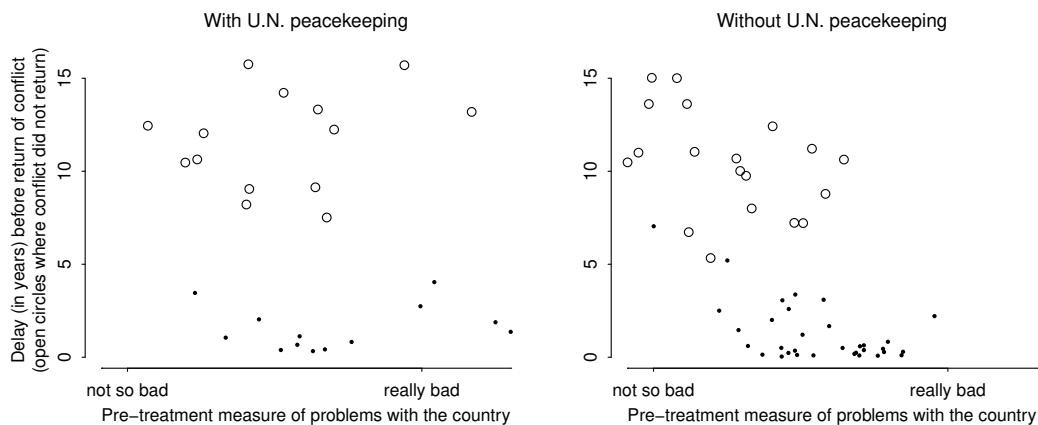
To put this in more general terms: in this study, the “treatment”—peacekeeping—was not randomly assigned. In statistics jargon, Fortna had an *observational study* rather than an *experiment*, and in an observational study we must do our best to adjust for pre-treatment differences between the treatment and control groups.

<sup>6</sup>Page Fortna (2008), *Does Peacekeeping Work? Shaping Belligerents' Choices after Civil War*, Princeton University Press. The discussion here is taken from Section 1.4 of *Regression and Other Stories*.

### 3. WEEK BY WEEK: THE FIRST SEMESTER



**Figure 12** Outcomes after civil war in countries with and without United Nations peacekeeping. The countries with peacekeeping were more likely to stay at peace and took on average about the same amount of time to return to war when that happened. However, there is a concern that countries with and without peacekeeping may differ in their pre-existing conditions; see Figure 13.



**Figure 13** Outcomes after civil war in countries with and without United Nations peacekeeping, plotted vs. a measure of how bad the situation was in the country. After adjusting for this pre-treatment variable, peacekeeping remains associated with longer periods without war.

Fortna adjusted for how badly off the country was before the peacekeeping-or-no-peacekeeping decision was made, using some objective measures of conditions within the country. The analysis was further complicated because in some countries we know the time until return to civil war, whereas in other countries all we can say is that civil war had not yet returned during the period of data collection. In statistics, this sort of incomplete data process is called “censoring,” which does not mean that someone has refused to provide the data but rather that, due to the process of data collection, certain ranges of data cannot be observed: in this case, the length of time until resumption of civil war is inherently unknowable for the countries that remained at peace through the date at which data collection had concluded. Fortna addressed this using a “survival model,” a complexity that we will ignore here. For our purposes here we summarize the combination of pre-treatment predictors as a scalar “badness score,” which ranges from 1.9 for the Yemen civil war in 1994 and 2.0 for India’s Sikh rebellion in 1993, to the cases with the highest badness scores, 6.9 for Angola in 1991 and 6.5 for Liberia in 1990.

Figure 13 shows outcomes for treated and control countries as a function of badness score, with some missing cases where not all the variables were available to make that assessment. According

to these data, peacekeeping was actually performed in tougher conditions, on average. As a result, adjusting for badness in the analysis (while recognizing that this adjustment is only as good as the data and model used to perform it) *increases* the estimated beneficial effects of peacekeeping, at least during the period of this study.

This example is relevant to the week's reading and to the course as a whole as it introduces ideas of comparison and causal inference with observational data that are important in many areas of social science.

## 2. Girls and sports: error of individual-level claim from state-level correlations

A claim was made in a research paper published in 2014 that "sports participation [in high school] causes women to be less likely to be religious . . . more likely to have children . . . more likely to be single mothers."<sup>7</sup> And the advertised effects were huge: "a ten percentage-point increase in state-level female sports participation generates a five to six percentage-point rise in the rate of female secularism, a five percentage-point increase in the proportion of women who are mothers, and a six percentage-point rise in the proportion of mothers who, at the time that they are interviewed, are single mothers." This is, as the saying goes, "big if true"—elasticities of 50% for things that have nothing (apparently) to do with the treatment—and are even larger when you consider that the outcomes are binary and, for example, sports participation can't make you secular if you were already going to be secular anyway, it can't cause you to have a child if you were already going to have a child anyway, and so forth.

But, as the authors explain in their blog posts and the scholarly article, they were not measuring the effects of sports participation directly. Rather, they were using aggregate statistics, "comparing women in states with greater levels of 1971 male [high school] sports participation . . . to women in states with lower levels of 1971 male sports participation." The outcomes are state-level average responses to General Social Survey questions for "respondents who completed tenth grade and who either attended high school before Title IX was passed in 1972 or after it came into full effect in 1978." Title IX is a law requiring equal treatment of males and females in schools.

The people who did this study were doing their best to target their analysis on the group of women who'd be affected by the treatment. Then they ran individual-level regressions on binary variables (just as a minor point, we think it would have been better to keep the original ordered responses so as not to throw away information), but the action is all coming from the state-level predictor, the measure of male athletic participation in 1971, by state.

The trouble is that (a) the treatment is at the group, not the individual, level, and (b) it's not a clean "natural experiment." Think of it this way. Suppose some states were randomly selected to get the Title IX treatment and some were not. This would be the ideal scenario—but, even there, you're measuring the effect of an aggregate policy, not the effect on individual participation. But it's much worse than that. Actually, the treatment was applied to all the states, so all that could be studied was an *interaction*—that is, a treatment effect that was different in some states than others. Finally, the interaction being studied is not random; there are systematic differences between states with higher and lower boys' high school sports participation in 1971. The highest rates are reported in North Dakota, Nebraska, Minnesota, Iowa, Kansas, Montana, Arkansas, South Dakota, Vermont, Idaho, Oregon, and Wyoming.

How do we think about this? We can start from the two ends: the published claims and the observed data pattern.

Start with the implausibility of the reported estimates. As noted above, if a ten percentage-point increase in state-level female sports participation is associated with a five percentage-point increase

<sup>7</sup>Phoebe Clarke and Ian Ayres (2014), The Chastain effect: Using Title IX to measure the causal effect of participating in high school sports on adult women's social lives, *Journal of Socio-Economics* 48, 62–71. The discussion here is taken from Andrew Gelman (2014), How much can we learn about individual-level causal claims from state-level correlations?, <https://statmodeling.stat.columbia.edu/2014/05/16/individual-level-causation/>.

in the proportion of women who are mothers, there's no way that most of this can be coming from a direct effect. The implication would be that there's this huge group of girls who (i) will have children if they do sports, and (ii) will not have children if they do not do sports. These estimated elasticities have to be driven by big differences between states that possibly have nothing to do with high school sports. One step would be to look at lots of different state-level predictors (not just boys' 1971 high school sports participation) and lots of different state-level outcomes. We put this on our blog, and a commenter remarked that "high state level sports participation by boys is associated with a high proportion of small (presumably rural) high schools in those states." Many of the states listed above have relatively low levels of religious participation.

We'd also suggest, for each outcome, to make a scatterplot of the state-level aggregate outcomes vs. boys' sports participation in 1971. If the researchers on this project want to make the causal leap, go for it—but make clear that it's a leap. In the meantime, the scatterplot, with the 50 states labeled by their convenient two-letter abbreviations, could give a lot of insight.

As is generally the case with these mismatches between correlation and causation, we don't want to say that the research hypotheses are false; rather, we'd say that the claims are not demonstrated from the data and that we think that any true effects are smaller and less consistent than implied by the researchers. It may still be true that, at the individual level, "sports participation causes women to be less religious, more likely to have children, and, if they do have children, more likely to be single mothers," even if the actual effects are an order of magnitude lower than claimed. Any effects could also go in the opposite direction. State-level correlations don't tell us much about this. Recall that if we were studying state-level correlations of income and voting, we'd come to the false conclusion that poor people are more likely to vote Republican. In the present example, the Title IX story helps, but only a little.

This example relates to the week's reading for two reasons. First, it's an example of the generalization from treatment to control group and an example of an observational study. Second, it demonstrates the difficulty of interpreting coefficients in a predictive model. You have to learn how to say things accurately: instead of "sports participation [in high school] causes women to be less likely to be religious" or even "girls who participate in sports are, on average, less likely to be religious," it's something like, "in states with greater levels of male sports participation in 1971, girls in 1978 were less likely to be religious, compared to girls in 1978 in states that were comparable in 1971 but with lower levels of male sports participation." It's a mouthful to say all this—but that's kind of the point: that's what can be learned from these data! The example also is relevant to the course as a whole in demonstrating quantitative reasoning about causal claims.

### Class-participation activities

#### 1. Bag of candies and sampling bias

We illustrate the challenges of sampling and inference using an activity in which students attempt to grab candies at random from a bag. The challenge is that the candies are of different sizes, and it is difficult to draw a representative sample.<sup>8</sup>

##### *Preparation*

The instructor should buy 100 candies of different sizes and shapes and put them in an opaque bag. Get something like 20 large full-sized candy bars, 20 or 30 smaller items like mini Snickers bars and mini Peppermint Patties, and then 50 or 60 really little things like tiny Tootsie Rolls, lollipops, and individually-wrapped Life Savers. Make sure it's exactly 100.

The activity also requires a digital kitchen scale that reads out in grams. This all takes some

<sup>8</sup>See Andrew Gelman (2008), The candy weighing demonstration, or the unwisdom of crowds, [https://statmodeling.stat.columbia.edu/2008/05/08/doing\\_the\\_candy/](https://statmodeling.stat.columbia.edu/2008/05/08/doing_the_candy/), and Section 9.1 of Andrew Gelman and Deborah Nolan (2017), *Teaching Statistics: A Bag of Tricks*, second edition, Oxford University Press.

1. Pull 5 candies out of the bag
2. Weigh the candies
3. Write down the weight
4. Put the candies back in the bag!!
5. Pass the scale and bag to your neighbors
6. Silently multiply the weight of the 5 candies by 20

Figure 14 Instructions for the candy-weighing activity. This should be projected on the screen or written on the board, and students can then follow along as they work on the activity.

preparation but we think it's worth it in creating a memorable learning experience that includes many statistical lessons.

The instructor should bring a note inside a sealed envelope (details below). Before class starts, the instructor should unobtrusively put the envelope somewhere, for example between two books on a shelf or behind a window shade.

#### *Setup*

The instructor should hold up the bag of candy, then project or write onto the board the instructions in Figure 14, saying these instructions aloud while writing or projecting them.

The students should work in pairs. Their goal is to estimate the total weight of all the candies in the bag. They can choose their 5 candies using any method—systematic sampling, random sampling, whatever. Whichever pair guesses closest to the true weight, they get the whole bag!

After demonstrating how to zero the scale, the instructor should give the scale and the bag of candies to a pair of students in the front row, and let them go.

#### *Action*

The activity will proceed silently while the rest of the class proceeds. So the instructor should continue on with the class, keeping track that the scale and bag are moving slowly through the room. After about 30 or 40 minutes, it should reach the back and the students will be done.

At this point, the pairs, one at a time, can call out their estimates, which the instructor should write on the board. They will be numbers like 3080, 2400, and 4340. Once all the numbers are written, a crude histogram (for example, bins for 1000–2000 grams, 2000–3000, 3000–4000, and 4000–5000) can be drawn. This represents the sampling distribution of the estimates.

Two students from the class (but not from the same pair) should then come to the board and look at all the estimates and give their best guess, having seen this information. The instructor should ask the class if they agree with these two students and then give the bag to the two students in the front of the room and have them weigh it.

#### *Punch line*

The weight of all 100 candies will be something like 1658. Every time we have done this activity, this true weight has been lower than all or almost all of the sampling-based estimates produced by students. The instructor should write this true weight as a vertical bar on the histogram on the blackboard. If this activity is used later in the course, it could be used to illustrate the concepts of bias and standard error of an estimate, but at this point in the semester we are focusing on data collection and the general challenges of statistics.

The instructor can then call out to the students who are sitting near the hidden envelope: “Um, uh, what’s that over there . . . Is it an envelope??? Really? What’s inside? Could you open it up?” A student opens it and reads out what’s written on the sheet inside: “Your guesses are too high!”

It is possible for students to have anticipated this bias and corrected for it in their sampling, by very carefully including many smaller candies in their sample, but in a large class there should only be a few students who will have thought things through so carefully, and so most of the guesses will be much higher than the true weight.

#### *Aftermath*

Now's the time to talk about sampling. Students should work in pairs to try to figure out what happened. The explanation is that large candies are easy to see and to grab, while small candies fall through the gaps between the large ones and end up at the bottom of the bag. Consider the analogy to conduct a survey by going to a shopping mall or by sending out an email survey and seeing who responds. Students can work in pairs to figure out how to take a random sample of 5 out of 100 candies. It won't be obvious to many of them that the way to do a random sample is to number each of the candies from 1 to 100 and pick numbers at random. Also, as noted above, this is an example that can be used later in the semester to illustrate bias and standard error.

This example connects to the week's readings because it demonstrates the challenge of gathering data that allows generalization from sample to population.

#### 2. Gather, plot, and discuss data from students

This activity begins with the instructor discussing with students what they would be interested in learning from each other. These can be questions about politics, school, relationships, family, ... things you'd like to share, or more neutral topics if the group is more shy. You will be making scatterplots, so for this activity, it is best if the measurements are continuous or approximately continuous. For example, a question on political identification (left, center, or right) will only have three values which will be hard to show in a scatterplot, so it would be better in that case to take a continuous measure on a 0–100 scale, in which case it would be necessary to calibrate the scale in some way. Once the class has decided on a few questions, they should enter their responses on a Google form; Figure 15 provides an example.

After the data are entered, the instructor can download the Google data as a .csv file and give it a header with a name for each question; for example, with the questions shown in Figure 15, these names might be `Biden_feeling`, `salary`, and `music_talent`. Then the instructor should read the data into R and make a scatterplot of a randomly chosen pair of variables (with blank axes). Here is some code to plot two randomly-selected columns of the data:

```
students <- read.csv("students.csv")
K <- ncol(students)
k <- sample(K, 2)
x <- students[,k[1]]
y <- students[,k[2]]
plot(students[,k[1]], students[,k[2]], xaxt="n", yaxt="n", xlab="", ylab="")
```

With the plot displayed on the screen, students should try to figure out which two variables have been plotted. Once students have made their guesses, the instructor can display a plot including numbers on the axes:

```
plot(x, y, xlab="", ylab="")
```

And then include the variable names:

```
plot(x, y)
```

This activity connects to the week's reading as an example of generalization from observed measurements to underlying constructs of interest, and this activity concerns some of the real-world challenges of measurements, such as accuracy, confidentiality, and, most fundamentally, defining

What is your feeling about Joe Biden on a 0 to 100 scale, where 0 very cold, 50 is neutral, and 100 is very warm?

Your answer \_\_\_\_\_

What annual salary do you guess you will have in your next job?

Your answer \_\_\_\_\_

Guess your level of musical talent relative to the other students in this class on a 0 to 100 scale, where 0 is that you are the worst, 50 is that you are in the middle, and 100 is that you are the best.

Your answer \_\_\_\_\_

**Submit**      **Clear form**

Figure 15 Example of a short survey form for the class using questions based on student input and discussion.

the ultimate goal: Why are these questions being asked in first place, What is the underlying quantity being measured?

In addition, it relates to the upcoming week's reading on data exploration and graphics, and the class can discuss the challenge of displaying these data, for example in dealing with discreteness.

### Computer demonstrations

#### 1. Graph of data and fitted regression lines

Here you make a graph similar to Figure 1.8 of *Regression and Other Stories*, showing a regression with imbalance between treatment and control group. This simple example demonstrates simulation, model fitting, comparison and adjustment, and graphics.

```
library("rstanarm")

# Simulate data from two groups
N <- 100
z <- sample(c(0,1), N, replace=TRUE)
x <- ifelse(z==0, rnorm(N, 2, 2), rnorm(N, 6, 2))
y <- 20 + 5*x + 15*z + rnorm(N, 0, 10)
data <- data.frame(x, z, y)

# Estimate regression
fit <- stan_glm(y ~ x + z, data=data, refresh=0)
print(fit)

# Display data and fitted model
plot(x, y, xlab="Pre-treatment predictor", ylab="Outcome measurement", bty="l",
      main="Continuous pre-treatment predictor and binary treatment", type="n")
points(x[z==0], y[z==0], pch=20, col="blue")
```

```
points(x[z==1], y[z==1], pch=20, col="red")
abline(coef(fit)[1], coef(fit)[2], col="blue")
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], col="red")

# Try simple comparison
diff <- mean(y[z==1]) - mean(y[z==0])
print(diff)
```

In this simulation, the true value of the intercept and the coefficients for  $x$  and  $z$  are 20, 5, and 15, respectively, and the estimates from the fitted regression will be close to that. The treatment effect, or coefficient of  $z$ , is 15.

But the raw difference between the mean of  $y$  when  $z = 1$  (the red dots) and the mean of  $y$  when  $z = 0$  (the blue dots) is large. That's from the imbalance.

The demonstration fits into the week's reading as an example of regression to adjust for differences between treatment and control groups, which is one of the main themes of the course. Later we will conduct more elaborate simulations in which the regression model is itself wrong.

## 2. Tinker with an example

You can continue the previous demonstration by playing around with the numbers to change the result. A key part of the above code is this line, which creates imbalance between treatment and control groups:

```
x <- ifelse(z==0, rnorm(N, 2, 2), rnorm(N, 6, 2))
```

To see the imbalance, just type:

```
print(mean(x[z==0]))
print(mean(x[z==1]))
```

If you change the simulation and increase the imbalance between the two groups, this will propagate into a greater error in the simple comparison, compared to the regression. If you lower the imbalance to zero, you'll get a better estimate of the treatment effect.

This demonstration connects to the week's reading by showing the concepts of balance in causal inference, and it is relevant to the course as a whole by introducing the idea of learning by experimenting with code.

## Drills

### 1. Describe a regression slope in words

A continuing theme in the course is precision in language. All the mathematical rigor in the world won't help a researcher who is sloppy in the interpretations of results.

For each example below, explain the meaning of the underlined number, first wrongly and then correctly.

- (a)  $\text{final} = 30 + \underline{0.8} * \text{midterm} + 10 * \text{math\_major} + \text{error}$

*Solution:* *Wrong summary:* Increasing the midterm score by 10 points increases the final score by 8 points. *Correct summary:* Under the fitted model, the average difference in final exam scores, comparing two people with the same math major status but differing by 10 points in the midterm, is 8. Or, comparing two students who are identical in their math major status, on average you'd predict the student who scored 10 points higher on the midterm to score 8 points higher on the final.

The correct summary differs from the wrong summary in three ways, first by using the language of comparison rather than intervention, second by explicitly stating that the number is an average, and third by stating that the number is an attribute of the fitted model.

### 3.2. PREDICTION AS A UNIFYING THEME IN STATISTICS AND CAUSAL INFERENCE

53

- (b)  $\text{final} = 30 + 0.8 * \text{midterm} + 10 * \text{math\_major} + \text{error}$
- (c)  $\text{vote\_share} = 0.13 + 0.11 * \text{incumbency} + 0.65 * \text{previous\_vote\_share} + \text{error}$

Assume here that this is a model of elections, where `incumbency` is coded as 1 for incumbents running for reelection and 0 for open seats.

- (d)  $\text{vote\_share} = 0.13 + 0.11 * \text{incumbency} + 0.65 * \text{previous\_vote\_share} + \text{error}$

#### 2. Simple coding: computing and graphing functions

For each question below, give R code.

- (a) Compute the values of  $y = 2 + 3x$  for the values  $x = 1, 2, 3, 4, 5$ .

*Solution:*

```
x <- 1:5  
y <- 2 + 3*x
```

- (b) Compute the values of  $y = 2 + 3x$  for the values  $x = -2, -1.5, -1, \dots, 1.5, 2$ .
- (c) Compute the values of  $y = |2 + 3x|$  for the values  $x = -2, -1.5, -1, \dots, 1.5, 2$ .
- (d) Graph the function,  $y(x) = 2 + 3x$ , for  $x$  in the range  $(-2, 2)$ .
- (e) Graph the function,  $y(x) = 2 + 3 \exp(-4x + 5)$ , for  $x$  in the range  $(-2, 2)$ .
- (f) Graph the function,  $y(x) = 1/x$ , for  $x$  in the range  $(-2, 2)$ .

### Discussion problems

#### 1. Height and earnings

Consider the following model that has been fit to data from a survey taken in 1990:

$$\text{earnings} = -26000 + 600 * \text{height} + 10600 * \text{male} + \text{error}.$$

Sketch this fitted line and interpret each coefficient in words.<sup>9</sup>

#### 2. Graph hypothetical data

Consider an experiment performed on 50 people, in which each person gets a pre-test, then half the people get the treatment and half get the control, and then later everyone gets a post-test. You will make series of graphs, plotting post-test vs. pre-test, using different symbols for treatment and control. Examples of such plots are in Figure 1.4 of *Regression and Other Stories*. The challenge of this discussion problem is to make such plots under different scenarios.

In each scenario, assume that the pre-test is a strong but not perfect predictor of the post-test. The scenarios will differ regarding the treatment effect and the balance between treatment and control group.

- (a) Sketch a graph of hypothetical data arising from an experiment in which the treatment assignment is *balanced with respect to the pre-test* and the treatment has a *positive effect*.
- (b) Sketch a graph of hypothetical data arising from an experiment in which *people with lower pre-test scores are more likely to get the treatment*, and the treatment has a *positive effect*.
- (c) Sketch a graph of hypothetical data arising from an experiment in which people with lower pre-test scores are more likely to get the treatment, and the treatment has *no effect*.

Once you have these scatterplots, you can discuss the role of regression in causal inference. In the second and third scenarios, the simple difference comparing post-test averages of treatments and controls will be misleading, as it is necessary to adjust for pre-treatment differences between treatment and control groups.

<sup>9</sup>From Section 6.3 of *Regression and Other Stories*.

### 3.3 Data collection and visualization

#### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Political leanings of sports fans	Measure handedness	Download and work with data	All graphs are comparisons	Tell stories with graphs
Use comparisons to redraw a graph	Scatterplot charades	Make plots clearer	Graph criticism	Plots of baby names

#### Reading

Chapter 2 of *Regression and Other Stories*: Data and measurement

#### Pre-class warmup assignments

1. Plots
  - (a) In R, create data  $x$  taking on the values 0, 0.5, 1, ..., 10, then create  $y = \sqrt{x}$  and make a scatterplot of  $y$  vs.  $x$ .
  - (b) In R, create a plot of the line  $y = \sqrt{x}$ , for  $x$  ranging from 0 to 10.
2. Overlaying colors
  - (a) In R, create datasets  $(x, y)$  corresponding to two groups of people. First, create the data for group 1, with  $x$  taking on the values 0, 0.5, 1, ..., 10, and  $y = \sqrt{x}$ . Then create the data for group 2, with  $x$  taking on the values 0, 2, 4, ..., 20, and  $y = 2.5 + \sqrt{x}$ . Make a single scatterplot showing the points  $(x, y)$  with data from group 1 in blue and data from group 2 in red.
  - (b) In R, create a single plot with two lines: the line  $y = \sqrt{x}$  in blue, and the line  $y = 2.5 + \sqrt{x}$  in red, for  $x$  ranging from 0 to 20.

#### Homework assignments

1. (a) Statistics as generalization (Exercise 1.7 of *Regression and Other Stories*)  
Find a published paper on a topic of interest where you feel there has been insufficient attention to:
  - i. Generalizing from sample to population.
  - ii. Generalizing from treatment to control group.
  - iii. Generalizing from observed measurements to the underlying constructs of interest.Explain your answers.  
(b) Statistics as generalization (Exercise 1.8 of *Regression and Other Stories*)  
Find a published paper on a topic of interest where you feel the following issues *have* been addressed well:
  - i. Generalizing from sample to population.
  - ii. Generalizing from treatment to control group.
  - iii. Generalizing from observed measurements to the underlying constructs of interest.Explain your answers.  
(c) A problem with linear models (Exercise 1.9 of *Regression and Other Stories*)

### 3.3. DATA COLLECTION AND VISUALIZATION

55

Consider the helicopter design experiment in Exercise 1.1 of *Regression and Other Stories*. Suppose you were to construct 25 helicopters, measure their falling times, fit a linear model predicting that outcome given wing width and body length,

$$\text{time} = \beta_0 + \beta_1 * \text{width} + \beta_2 * \text{length} + \text{error},$$

and then use the fitted model,  $\text{time} = \beta_0 + \beta_1 * \text{width} + \beta_2 * \text{length}$ , to estimate the values of wing width and body length that will maximize expected time aloft.

- i. Why will this approach fail?
  - ii. Suggest a better model to fit that would not have this problem.
2. (a) Composite measures (Exercise 2.1 of *Regression and Other Stories*)
- Following the example of the Human Development Index in Section 2.1 of *Regression and Other Stories*, find a composite measure on a topic of interest to you. Track down the individual components of the measure and use scatterplots to understand how the measure works, as was done for that example in the book.

## Stories

### 1. Political leanings of sports fans

In January, 2009, political commentator Brad Miner wrote:<sup>10</sup>

“With the Super Bowl coming up this weekend, I want to write about sports, which I consider a key to building a larger conservative coalition in America. . . .

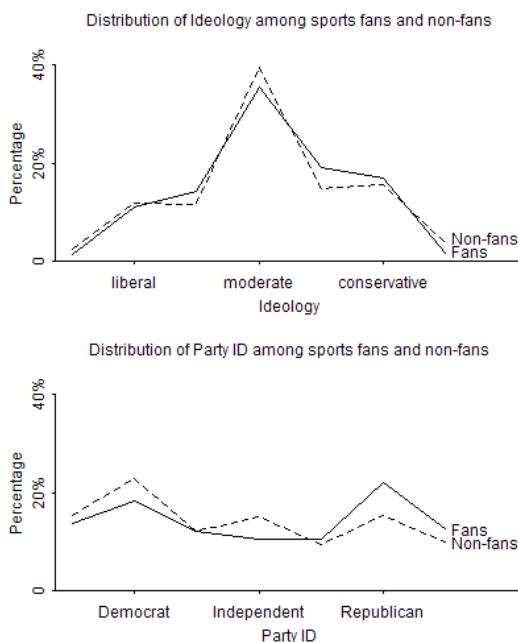
If you did a survey of the political philosophies of 75,000 randomly selected Americans you’d expect the usual—if somewhat mystifying—results: ‘Only about one-in-five Americans currently call themselves liberal (21%), while 38% say they are conservative and 36% describe themselves as moderate.’ So said the folks at Pew Research, and this was after the November election. Do that same poll among the fans at Raymond James Stadium in Tampa on Sunday and the results would likely be more like 15% liberal, 30% moderate, and 50% conservative. And a bunch of those liberals would probably be gun owners. Obviously those numbers are just speculation on my part, but I guarantee that Steelers fans are more conservative than all Pennsylvanians and ditto Cardinals devotees and the rest of Arizona. Which is not to say that these folks cast their ballots in November more for McCain than Obama. That’s the problem.”

What do the data say? We looked up the “attended sporting event in the past year” item on the General Social Survey. Unfortunately, the question was only asked once, in the 1993–1996 survey, at which time 56% of respondents said they attended an “amateur or professional sports event” during the past twelve months. Figure 16 shows how they differed from the 44% who didn’t.

At least in the mid-1990s, sports attenders were a bit more Republican than other Americans (the categories in Figure 16 are Strong Democrat, Democrat, lean Democrat, Independent, lean Republican, Republican, strong Republican), but not much different in their liberal-conservative ideology.

So these data do not appear to support Miner’s claim. Miner expected sports fans to label themselves as more conservative but maybe not to be more likely to vote Republican; actually, sports fans were more likely to call themselves Republican but no more likely to describe themselves as conservative.

<sup>10</sup>See Andrew Gelman (2009), Sports fans as potential Republicans?, [https://statmodeling.stat.columbia.edu/2009/01/27/sports\\_fans\\_as/](https://statmodeling.stat.columbia.edu/2009/01/27/sports_fans_as/), and Andrew Gelman (2010), Political leanings of sports fans, [https://statmodeling.stat.columbia.edu/2010/04/07/political\\_lean/](https://statmodeling.stat.columbia.edu/2010/04/07/political_lean/).



**Figure 16** From the General Social Survey in 1993–1996, distribution of political ideology and party identification scores from respondents who answered Yes or No to the question of whether they attended a sporting event during the past year. There was not much difference between the two groups in political ideology, but sports fans were more Republican, on average, than non-fans.

That said, the sporting event attended could be anything between the Super Bowl and a kids' soccer game. Maybe more dramatic results would be obtained by considering a more restricted group of sports fans. There are lots of surveys of TV watching, so there should be data that would give more detail on ideology, voting, and spectator sports watching. And, more generally, we never want to rely too strongly on just one survey. Still, you have to start somewhere.

A year later, political reporter Reid Wilson shared a graph showing reported political attitude and voter turnout among watchers of different TV sports. The National Football League was in the middle but with a bit more Republican than Democratic viewers. You can find the graph online.<sup>11</sup>

The graph at that link is excellent. In particular, its use of red and blue coloring (indicating points to the left or right of the zero line) and light/dark shading (indicating points above or below the center line on the vertical axis) are good ideas, we think, despite that they convey no additional information, in that they draw attention to key aspects of the data. The sizes of the circles in that graph appear to be proportional to the number of fans of the different sports. The only thing that bothers us is that we don't see any definition of the "Republican Index," "Democrat Index," or "Voter Turnout Index." We understand the relative positions of the different circles on the graphs but not the absolute numbers. For example, the National Football League has a GOP Index of 111, a Dem Index of 103, and a Voter Turnout Index of 111. We don't know what these numbers represent, exactly.

This story is relevant to the week's reading because it shows how open questions can be addressed through data, indeed often through data that have already been collected. The example also demonstrates what we can learn from simple data exploration and graphics, and the difficulty understanding the details in the second figure reveals the challenges of interpreting processed data.

<sup>11</sup>Reid Wilson (2010), Sports viewers largely Republican, *National Journal*, [https://web.archive.org/web/20100402234855/http://hotlineoncall.nationaljournal.com/archives/2010/03/sports\\_viewers.php/](https://web.archive.org/web/20100402234855/http://hotlineoncall.nationaljournal.com/archives/2010/03/sports_viewers.php/).

### 3.3. DATA COLLECTION AND VISUALIZATION

57

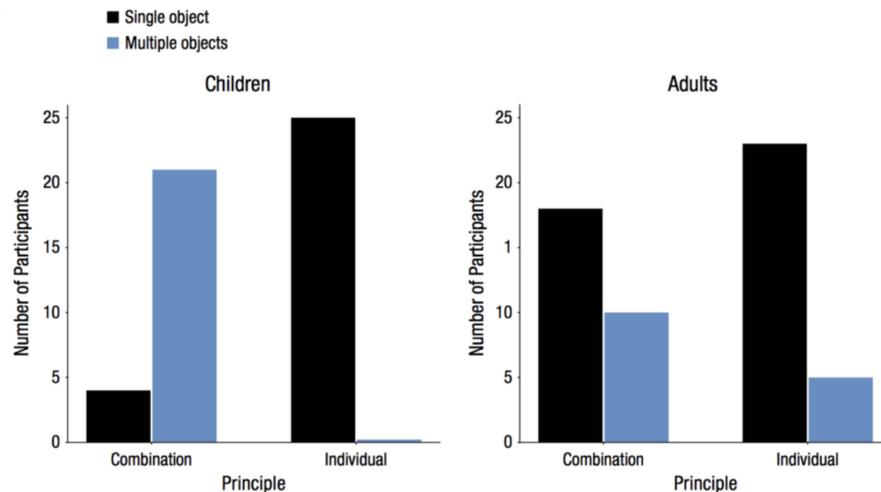


Figure 17 Graph from a published paper summarizing the results of a psychology experiment. We think this graph can be improved; see Figure 19.

Going further, we might ask what new data could be collected to assess the potential value of future political pitches aimed at sports fans.

#### 2. Use the “graphs as comparisons” principle to redraw a graph

Figure 17 shows a pair of bar plots that take up half a page in a published article in psychology.<sup>12</sup> We project this onto the screen and ask the class how these plots could be improved.

We could start with the details. To read the graph, you need to go back and forth between the legend and the bars to keep the color scheme fresh in your mind. The negative space in the middle of each plot looks a bit like a white bar, which is not directly confusing but makes the display harder to follow. And we had to do a double-take to interpret the infinitesimal blue bar of zero height on the left plot. Also, it’s not clear how to interpret the *y*-axes: 25 participants out of how many? And what’s up with the *y*-axis on the second graph? The 15 got written as a 1, which makes us suspect that the graph was re-drawn from an original, which then leads to concern that other mistakes may have been introduced in the re-drawing process.

But that’s not the direction we want to go. There are problems with the visual display, but going and fixing them one by one would not resolve the larger problem. To get there, we need to think about goals.

To structure the discussion, we write on the board the text on Figure 18—ordinarily we would project pre-written material onto the screen, but in this case the screen is occupied by Figure 17. A graph is a set of comparisons, and the two goals of a graph are:

- To understand and communicate the size and directions of comparisons that you were already interested in, before you made the graph.
- To facilitate discovery of new patterns in data that go beyond what you expected to see.

Both these goals are important. We want to understand and communicate what we think we know, and we also want to put ourselves in a position where we can learn more.

The question, when making any graph, is: what comparisons does it make it easy to see? After all, if you just wanted the numbers you could put them in a table.

Now let’s apply these principles to the bar plots in Figure 17. In that display it is easy to compare

<sup>12</sup>From Alison Gopnik, Thomas Griffiths, and Christopher Lucas (2015), When younger learners can be better (or at least

Graphs as comparisons. Two goals:

1. Understand and communicate the size and directions of comparisons that were already of interest
2. Discovery of new patterns

**Figure 18** *Two goals of a graph. The instructor should write these on the board in preparation for the discussion of redrawing the graph in Figure 17.*

the heights of two bars right next to each other—for example, you can see that the black bars all higher than the blue bars, except for the pair on the left . . . hmmmm, which are blue and which are black, again? Finally, the graph lacks an overall title that would point us to the most important comparisons it is presenting.

We're not trying to rag on the authors here. This sort of Excel-style bar graph is standard in so many presentations. We just think they could do better.

So, how to do better? Let's start with the goals.

- (a) What are the key comparisons that the authors want to emphasize? From the caption, it seems that the most important comparisons are between children and adults. We want a graph that shows the following patterns:
  - i. In the Combination scenario, children tended to choose multiple objects (the correct response, it seems) and adults tended to choose single objects (the wrong response).
  - ii. In the Individual scenario, both children and adults tended to choose a single object (the correct response). Actually, we would re-order these and first look at the Individual scenario, which seems to be some sort of baseline, and then go to Combination which is displaying something new.
- (b) What might we want to learn from a graph of these data, beyond the comparisons listed just above? This one's not clear, so we'll guess. Who were those kids and adults who got the wrong answer in the Combination scenario? Did they have other problems? What about the adults who got the wrong answer in the Individual scenario, which was presumably easier? Did they also get the answer wrong in the other case? There must be some other things to learn from these data too—it's hard to get people to participate in a psychology experiment, and once you have them there, it makes sense to give them as many tasks as can be done. But from this figure alone, we're not sure what these other questions would be.

We next consider how to make the revised graph. Given that we don't have the raw data, and we're just trying to redo the figure above, we'll focus on the first task: displaying the key comparisons clearly.

Hey—we just realized something! The two outcomes in this study are “Single object” and “Multiple object”—that's all there is! And, looking carefully, we see that the numbers in each graph add up to a constant: it's 25 children and, ummm, let's read this carefully . . . 28 adults!

This simplifies our task considerably, as now we have only four numbers to display instead of eight.

We can easily display four numbers with a line plot. The outcome is percentage who give the “Single object” response, and the two predictors are Child / Adult and Individual Principle / Combination Principle.

---

more open-minded) than older ones, *Current Directions in Psychological Science* 24, 87–92, <https://cocosci.princeton.edu/tom/papers/LabPublications/GopnicketalYoungLearners.pdf>. The discussion here is taken from Andrew Gelman (2017), Graphs as comparisons: A case study, <https://statmodeling.stat.columbia.edu/2017/07/16/graphs-comparisons-case-study/>.

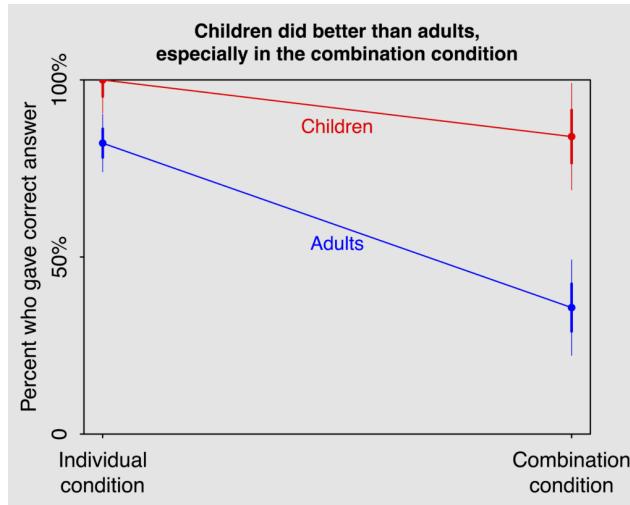


Figure 19 Redrawn graph conveying the same information as in Figure 17, but using less space, and with the key comparisons easier to see.

One of these predictors will go on the  $x$ -axis, one will correspond to the two lines, and the outcome will go on the  $y$ -axis.

In this case, which of our two predictors goes on the  $x$ -axis?

Sometimes the choice is easy: if one predictor is binary (or discrete with only a few categories) and the other is continuous, it's simplest to put the continuous predictor as  $x$ , and use the discrete predictor to label the lines. In this case, though, both predictors are binary, so we need to think more carefully to decide what to do.

We would like to use logical or time order, as that's easy to follow. There are two options:

- Time order in age, thus Children, then Adults; or
- Logical order in the experiment, thus Individual Principle, then Combination Principle, as Individual is in a sense the control case and Combination is the new condition.

We tried it both ways and we think the second option was clearer. So we'll show this graph and the corresponding R code. The first option could work too.

Figure 19 shows the result, which we think is better than the bar graphs from the original article, for two reasons. First, we can see everything in one place: as the title says, "Children did better than adults, specially in the combination condition." Second, we can directly make both sorts of comparisons: we can compare children to adults, and we can also make the secondary comparison of seeing that both groups performed worse under the combination condition than the individual condition.

We also threw in  $\pm 1$  and 2 standard error bars, using the formula based on  $(y + 2)/(n + 4)$  for the uncertainty of a proportion; see page 52 of *Regression and Other Stories*. The one thing this graph does not show is whether the adults who got it wrong on the individual condition were more likely to get it wrong in the combination condition, but that information wasn't in the original graph either.

On the whole, we are satisfied that the replacement graph contains all the information in less space and is much clearer than the original. Again, this is not a slam on the authors of the paper. They were not working within a tradition in which graphical display is important.

This example connects to the week's reading as an example of thinking about graphs as comparisons. It relates to the course as a whole in that it is about communication and learning from data.

### Class-participation activities

#### 1. Measure handedness

One way to explore the surprises of real-world statistics is to look carefully at data collected from students. Before displaying the collected data, it can be instructive to ask students to share their guesses of what the data look like. An example that has worked with us is a simple survey of handedness.<sup>13</sup>

Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would never use the other hand for that activity. If in any case you are really indifferent, put + in both columns.

Task	Left	Right
Writing		
Drawing		
Throwing		
Scissors		
Toothbrush		
Spoon		
Total		

$$\text{Right} - \text{Left} : \quad \text{Right} + \text{Left} : \quad \frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$$

Create Left and Right scores by counting the total number of + signs in each column. Your handedness score is  $(\text{Right} - \text{Left})/(\text{Right} + \text{Left})$ : thus, a pure right-hander will have a score of  $(12 - 0)/(12 + 0) = 1$ , and a pure left-hander will have a score of  $(0 - 12)/(0 + 12) = -1$ .

Figure 20 Handedness inventory. Each student should fill out this form and report the total score. Students can then divide into pairs and sketch their guesses of the histogram of these scores for the students in the class, as illustrated in Figure 21.

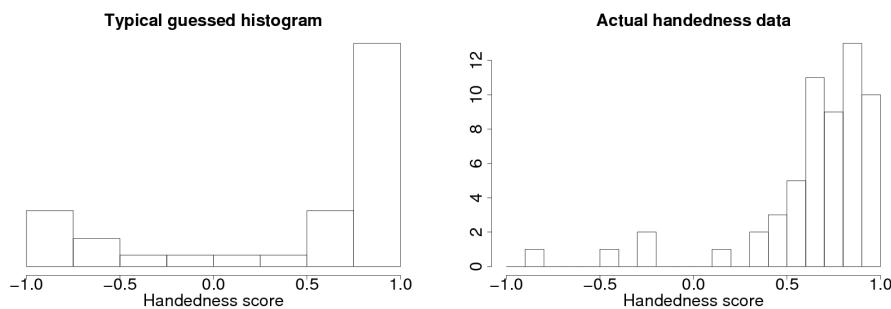
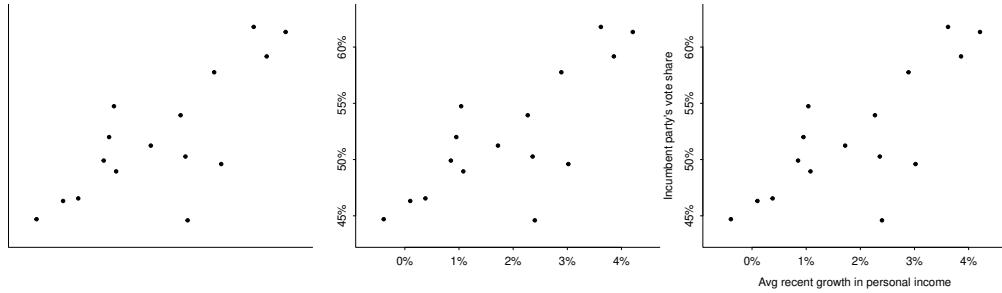


Figure 21 After completing their handedness forms (see Figure 20), students should work in pairs to guess and draw the distribution of responses for students in their class. (a) A guess from a group of students of the histogram of handedness scores in their class; (b) actual data. As in this example, students' guessed distributions typically show a sharp division into left- and right-handers; depending on the questions used in the survey, actual data can show a smoother distribution including mixed-handers.

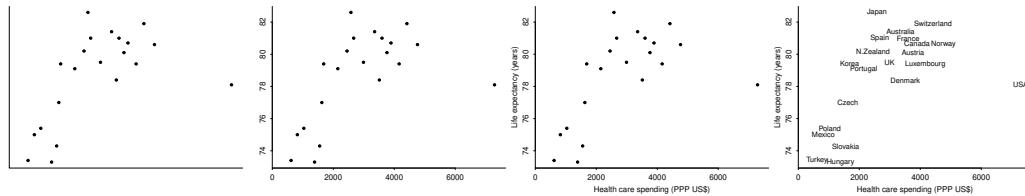
<sup>13</sup>From Section 3.3.2 of *Teaching Statistics: A Bag of Tricks*.

### 3.3. DATA COLLECTION AND VISUALIZATION

61



**Figure 22** Example of scatterplot charades, using Figure 1.1b of Regression and Other Stories. Show one graph at a time, first displaying just the scatterplot, then axes, then axis labels. In this case, the data show results from recent U.S. presidential elections. Further scatterplot charades examples appear in Figures 23–25.



**Figure 23** Another example of scatterplot charades, using Figure 2.4 of Regression and Other Stories. Show one graph at a time, first displaying just the scatterplot, then axes, then axis labels, then labels on the points. Further scatterplot charades examples appear in Figures 24–25.

The instructor should print copies of Figure 20 for all students and have them compute their handedness scores, which range from  $-1$  to  $1$ ; see Figure 20. The next step is to collect these forms and ask one student to enter the scores into the computer (just the final  $(\text{Right} - \text{Left}) / (\text{Right} + \text{Left})$  score, no need to type in all the other information on the forms). While this is happening, the students should pair up, and each pair should take two minutes to draw their guess of the distribution of the handedness scores of the students in the class.

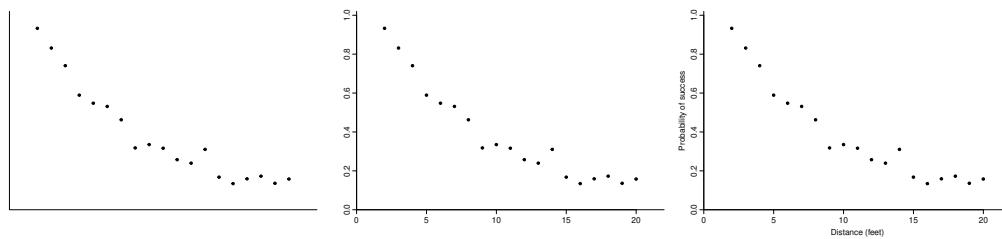
Then one pair should present their histogram on the board—it typically looks like Figure 21a—and invite comments from the class. Since they have all just worked on the problem, many students should be eager to participate. The instructor should adjust the drawn histogram following the suggestions of the class, and then display the actual data that by now the student has typed in and read into R. An example from a recent class appears in Figure 21b.

This example fits into the week’s reading on the challenges of measurement. Is the Handedness Inventory a good way to measure handedness? Are there other skills (for example, texting) that should be added to the survey?

#### 2. Scatterplot charades

Students do this activity in pairs. Each student should come to class with a scatterplot on some interesting topic printed on paper or visible on their computer or phone, and then reveal the plot to the other student in the pair, a bit at a time, starting with the dots only and then successively uncovering units, axes and titles. At each stage, the other student should try to guess what is being plotted, with the final graph being the reveal.

Then the two students should switch roles. Once they’ve gone through both examples, the students in each pair should then discuss why it was easy or difficult to guess each plot’s content and message, and then they can discuss how it might be possible to improve each plot to make its visual patterns easier to grasp.



**Figure 24** Another example of scatterplot charades, using Figure 22.1 of Regression and Other Stories. Show one graph at a time, first displaying just the scatterplot, then axes, then axis labels. This graph shows data on the success rate of putts in professional golf. A final scatterplot charades example appears in Figure 25.

To give a sense of how this works, Figures 22–25 demonstrate with some examples from *Regression and Other Stories*. At the end of the class period *before* we plan to do this activity, we go through one or two of these examples with students ahead of time, to give a sense of what sort of graphs they could bring to class.

### Computer demonstrations

#### 1. Download and work with data

It is useful to be able to grab data online and work with these data. We demonstrate the steps of data cleaning, manipulation, and plotting, using an example from one of the datasets from the website of *Regression and Other Stories*.

```
# Read data from here:  https://github.com/avehtari/ROS-Examples
library("foreign")
library("dplyr")
pew_all <- read.dta(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Pew/data/",
  "pew_research_center_june_elect_wknd_data.dta"))

# Look at data
pew_all
names(pew_all)
pew_all$age
table(pew_all$age)
pew_all$party
as.numeric(pew_all$party)
table(pew_all$party)
table(as.numeric(pew_all$party))
table(pew_all$sex)
table(pew_all$marital)

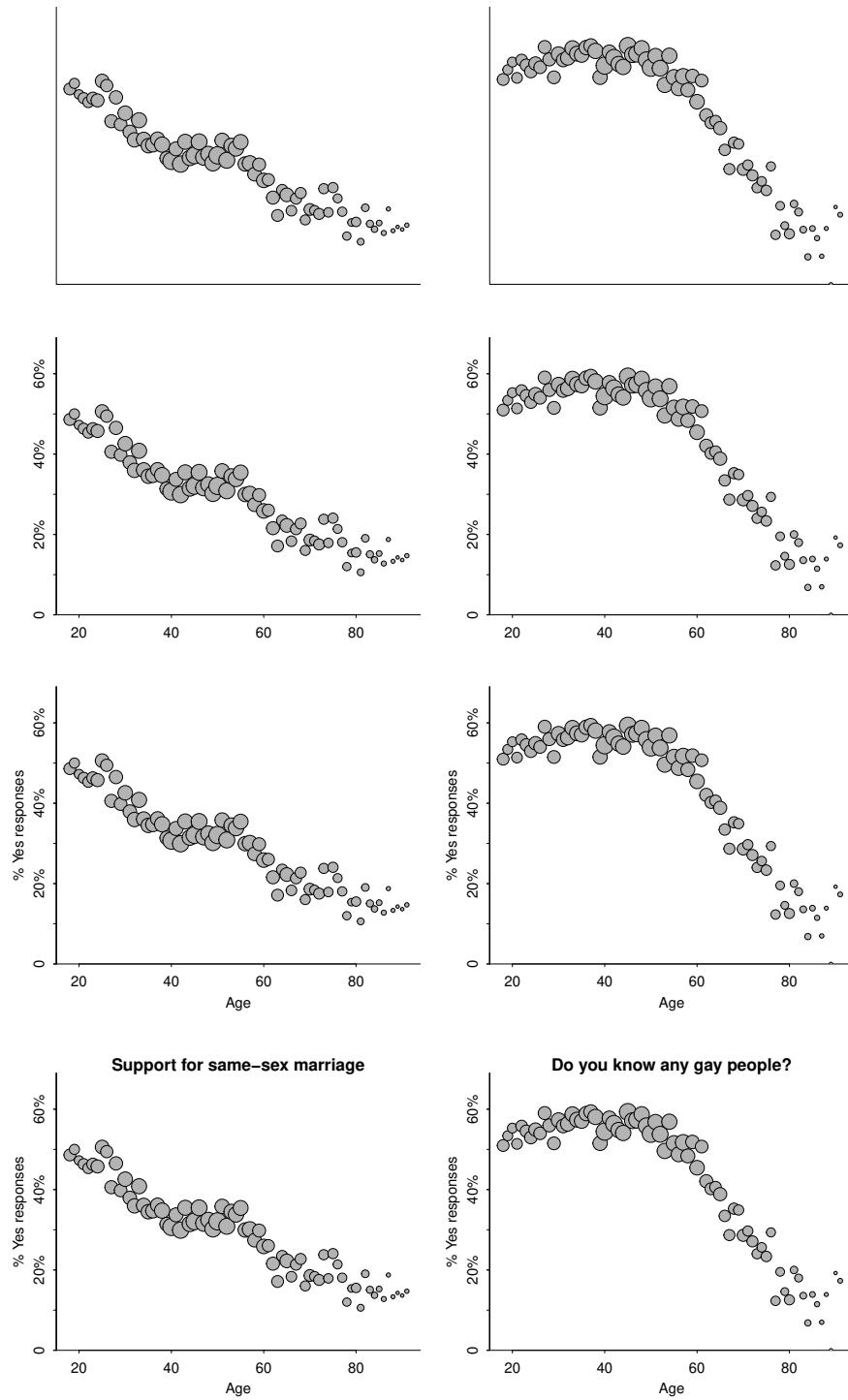
# Recode party into 3 categories:
# -1 for Democrat, 1 for Republican, 0 otherwise
pew_all$party_simple <- recode(pew_all$party, "missing/not asked" = 0,
  "republican" = 1, "democrat" = -1, "independent" = 0, "no preference" = 0,
  "other" = 0, "dk" = 0)

# Extract a few questions, remove NA's, and remove data with weird ages
pew <- pew_all %>% select(c("age", "sex", "marital", "party_simple")) %>%
  na.omit() %>% filter(age < 97)

plot(pew$age, pew$party_simple)
```

### 3.3. DATA COLLECTION AND VISUALIZATION

63



**Figure 25** Another example of scatterplot charades, using Figure 22.5 of Regression and Other Stories. Show one row of plots at a time, first displaying just the scatterplot, then with axes, then with axis labels. The final row shows the fully labeled graphs based on a U.S. public opinion survey from 2004.

```
# Plot with jitter
n <- nrow(pew)
plot(pew$age + runif(n, -0.1, 0.1), as.numeric(pew$party_simple) +
    runif(n, -0.1, 0.1), xlab="Age", ylab="Party")
```

This final graph is ugly! Some data are not meant to be scatterplotted.

The main point of this demonstration is to give a hands-on feeling for downloading and working with data, but this example also shows that sometimes the natural-seeming graph won't be useful. More work could be done to make a better graph of these data, but we won't do that here.

## 2. Make plots clearer

In this demonstration we simulate some data, plot them using the default settings in R, and then go through some steps to make the graph more readable. This is not the only way to make clean graphs in R—many users recommend the `ggplot2` package. The point here is just that it is possible to put in some effort and make cleaner, more informative plots.

```
# Generate 50 data points
N <- 50
midterm <- runif(N, 0, 100)
a <- 20
b <- 0.5
error <- rnorm(N, 0, 5)
treatment <- sample(c(0,1), N, replace=TRUE)
theta <- 10
final <- a + b*midterm + theta*treatment + error
fake <- data.frame(midterm, treatment, final)

# Plot data
plot(fake$midterm, fake$final)

# Improved plot
plot(fake$midterm, fake$final, main="Comparing teaching plans",
     xlab="Midterm", ylab="Final", col=ifelse(fake$treatment==1,
     "red", "blue"), pch=20, xlim=c(0, 100), ylim=c(0, 100))
```

There are other ways to effectively plot data, just as there are many ways to express a thought in human language. The point is that a programming language such as R allows us to improve a graph, step by step, as necessary.

## Drills

### 1. All graphs are comparisons

For each of a series of graphs, identify at least one implicit or explicit comparison that is facilitated by the graph. You can pick graphs from the internet or just use the ones below.<sup>14</sup>

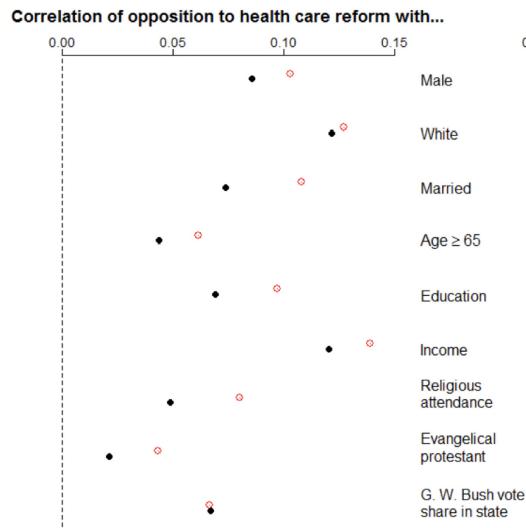
(a) Dot plot:<sup>15</sup>

<sup>14</sup>For a few more examples, see Andrew Gelman (2021), Making better data charts: From communication goals to graphics design, <https://statmodeling.stat.columbia.edu/2021/09/23/design-charts/>.

<sup>15</sup>From Andrew Gelman, Daniel Lee, and Yair Ghitza (2010), Public opinion on health care reform, *The Forum* 8 (1), 8.

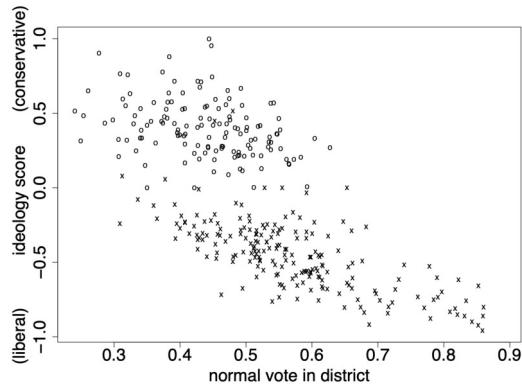
### 3.3. DATA COLLECTION AND VISUALIZATION

65

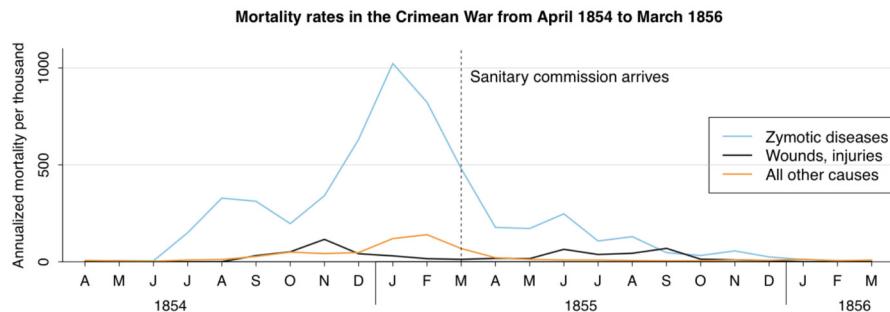


*Solution:* This graph displays several comparisons. First, for each predictor, the solid dot and the open circle are compared to each other. Second, the dots for different predictors are being compared. Third, all the dots are compared to zero.

(b) Scatterplot:<sup>16</sup>



(c) Multiple time series:<sup>17</sup>



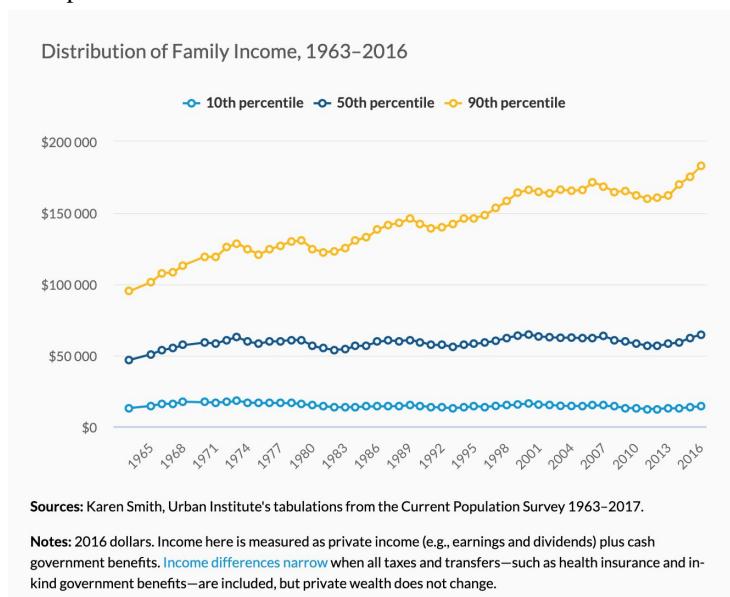
<sup>16</sup>From Andrew Gelman and Jonathan Katz (2007), Moderation in the pursuit of moderation is no vice: The clear but limited advantages to being a moderate for Congressional elections, <http://www.stat.columbia.edu/~gelman/research/unpublished/moderation5.pdf>.

<sup>17</sup>From Andrew Gelman and Antony Unwin (2013), Infovis and statistical graphics: Different goals, different looks (with discussion), *Journal of Computational and Graphical Statistics* 22, 2–49.

## 2. Graph criticism

For each of a series of graphs, identify at least one criticism. You can pick graphs from the internet or simply use the ones here:

### (a) Multiple time series:<sup>18</sup>



*Solution:* This graph is good, but it has a couple of small flaws. First, rather than using a legend at the top of the graph, it would be better to label the lines directly, which would convey the information without the reader needing to go back and forth. Second, the *x*-axis can be improved by putting year labels every ten years and tick marks every five years, which again gives the information without the reader needing to work to decode the *x*-axis.

### (b) Line plot:<sup>19</sup>

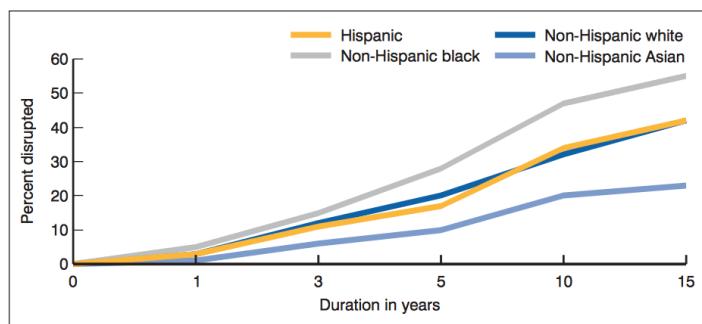


Figure 18. Probability that the first marriage breaks up by duration of marriage and race/ethnicity: United States, 1995

<sup>18</sup>From Signe-Mary McKernan, Caroline Ratcliffe, C. Eugene Steuerle, Caleb Quakenbush, and Emma Kalish, Nine charts about wealth inequality," Urban Institute, <https://apps.urban.org/features/wealth-inequality-charts/>.

<sup>19</sup>From Matthew Bramlett and William Mosher (2022), Cohabitation, marriage, divorce, and remarriage in the United States, *Vital Health Statistics* 23 (22), National Center for Health Statistics.

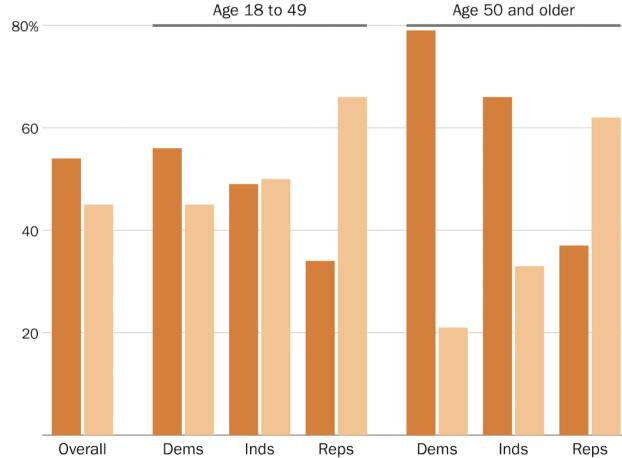
### 3.3. DATA COLLECTION AND VISUALIZATION

67

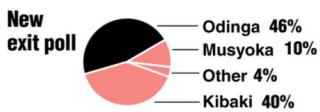
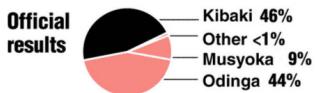
#### (c) Bar plot:<sup>20</sup>

How concerned are you that you or someone you know will be infected with the coronavirus:  
very concerned, somewhat concerned, not so concerned, or not concerned at all?

● Very/some      ● Not so/not



#### (d) Pie charts:<sup>21</sup>



### Discussion problems

#### 1. Tell stories with graphs

Find an example that interests you, and mock up a series of graphs to display what you'd like to see. For example, if you are interested in tracking the progress of students in a school and compare grades of students in different years of the program, you might make a set of time-series graphs and scatterplots.

#### 2. Name Grapher and plots of baby names

Go online to Name Grapher and play with some names.<sup>22</sup> This dynamic visualization, whose construction is far beyond the scope of this course, is a great demonstration of the possibilities of data exploration. Students should go on their computers in pairs and play with Name Grapher and see what interesting patterns they can find. We also point out that this program, while wonderful, does not do everything, and sometimes some coding is needed to extract and display aspects of the data, for example when graphing the trends of last letters of baby names shown in Section 2.1 of *Regression and Other Stories*.

<sup>20</sup>From Philip Bump (2020), Older Americans are more worried about coronavirus—unless they're Republican, *Washington Post*, 14 Mar, <https://www.washingtonpost.com/politics/2020/03/14/older-americans-are-more-worried-about-coronavirus-unless-theyre-republican/>.

<sup>21</sup>See Andrew Gelman (2017), You'll never guess this one quick trick to diagnose problems with your graphs and then make improvements, <https://statmodeling.stat.columbia.edu/2017/06/02/quick-trick/>.

<sup>22</sup>See <https://namerology.com/baby-name-grapher/>. For further background, see Andrew Gelman (2022), The Baby Name Voyager is back!, <https://statmodeling.stat.columbia.edu/2022/03/13/the-baby-name-voyager-is-back/>.

## 3.4 Review of mathematics and probability

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Death rate in the pandemic	Amoebas and exponential growth and decline	Matrix manipulations	Graphs of straight lines	College admissions
Galton's giants	Squares, cubes, a power-law growth	Compute weighted averages	Normal distribution	Probability of a rare event

### Reading

1. Chapter 3 of *Regression and Other Stories*: Some basic methods in mathematics and probability.
2. Appendix A of *Regression and Other Stories*: Computing in R, Sections A.4–A.5

### Pre-class warmup assignments

#### 1. Weighted averages

In R, create three vectors each of length 10, the first being the number 1 repeated 10 times, the second being the numbers 1 to 10, and the third being the sequence 1, 4, 9, . . . , 100. Compute the following weighted averages of these three vectors. In each case, your output should be a vector of length 10.

- (a) Weights 1/3, 1/3, 1/3
- (b) Weights 1, 1, 1
- (c) Weights 1, 2, 3
- (d) Weights 0, 1, 2

#### 2. Simulate from the binomial distribution

Express each of these lines of code in story form. For example, the first example could be the number of shots made by a basketball player who takes 20 shots and has a 30% chance of making each shot. Run each line of code in R on your computer to check that your story interpretation makes sense.

- (a) `rbinom(1, 20, 0.3)`
- (b) `rbinom(1, 20, 0.4)`
- (c) `rbinom(2, 20, c(0.3, 0.4))`
- (d) `rbinom(2, c(30, 20), c(0.3, 0.4))`

### Homework assignments

#### 1. (a) Data visualization (Exercise 2.6 of *Regression and Other Stories*)

Take data from some problem of interest to you and make several plots to highlight different aspects of the data, as was done in the baby names example in Figures 2.6–2.8 in *Regression and Other Stories*.

#### (b) Reliability and validity (Exercise 2.7 of *Regression and Other Stories*)

- i. Give an example of a scenario of measurements that have *validity* but not *reliability*.
- ii. Give an example of a scenario of measurements that have *reliability* but not *validity*.

### 3.4. REVIEW OF MATHEMATICS AND PROBABILITY

69

#### 2. (a) Weighted averages (Exercise 3.1 of *Regression and Other Stories*)

A survey is conducted in a certain city regarding support for increased property taxes to fund schools. In this survey, higher taxes are supported by 50% of respondents aged 18–29, 60% of respondents aged 30–44, 40% of respondents aged 45–64, and 30% of respondents aged 65 and up. Assume there is no nonresponse. Suppose the sample includes 200 respondents aged 18–29, 250 aged 30–44, 300 aged 45–64, and 250 aged 65+. Use the weighted average formula to compute the proportion of respondents in the *sample* who support higher taxes.

#### (b) Weighted averages (Exercise 3.2 of *Regression and Other Stories*)

Continuing the previous exercise, suppose you would like to estimate the proportion of all adults in the *population* who support higher taxes, so you take a weighted average as in Section 3.1 of *Regression and Other Stories*. Give a set of weights for the four age categories so that the estimated proportion who support higher taxes for all adults in the city is 40%.

#### (c) Probability distributions (Exercise 3.3 of *Regression and Other Stories*)

Using R, graph probability densities for the normal distribution, plotting several different curves corresponding to different choices of mean and standard deviation parameters.

#### (d) *In pairs:* Working through your own example (Exercise 2.10 of *Regression and Other Stories*)

Continuing the example from Exercise 1.10 of *Regression and Other Stories*, graph your data and discuss issues of validity and reliability. How could you gather additional data, at least in theory, to address these issues?

## Stories

### 1. Death rate in the pandemic

In April, 2021, the *New York Times* claimed that “the U.S. death rate in 2020 was the highest above normal since the early 1900s—even surpassing the calamity of the 1918 flu pandemic.”<sup>23</sup>

Really? No. The death rate increased by 15% from 2019 to 2020, but it jumped by 40% from 1917 to 1918. But, if so, why would anyone claim differently? Therein lies a tale.

Here’s the background, again from the news article:

“In the first half of the 20th century, deaths were mainly dominated by infectious diseases. As medical advancements increased life expectancy, death rates also started to smooth out in the 1950s, and the mortality rate in recent decades—driven largely by chronic diseases—had continued to decline.”

To continue the discussion, we follow the link to the news article and show three graphs.<sup>24</sup> By looking at these graphs, we will be able to figure out what is going on, how the wrong claim came to be, and why it is mistaken.

The first of the three graphs is labeled “Death rate above and below normal in the U.S.” and shows a time series from 1910 through 2020, with a peak at about +11% during the 1918 flu pandemic and another peak at +16% during the 2020 covid pandemic.

The second graph is labeled “Death rate in the U.S. over time” and shows a gradual decline in death rate from slightly above 2000 deaths per 100 000 at the beginning of the century to a level below 1000 in recent decades. We don’t know why they give deaths per 100 000, which is a scale that we have little intuition on. For some people, reporting the death rate as 2.6% (that is, 2.6 per hundred) would be more interpretable than a death rate of 2600 per 100 000.

<sup>23</sup>Denise Lu (2021), How COVID upended a century of patterns in U.S. deaths, *New York Times*, 23 Apr, <https://www.nytimes.com/interactive/2021/04/23/us/covid-19-death-toll.html>. The discussion here is taken from Andrew Gelman (2021), Is it really true that “the U.S. death rate in 2020 was the highest above normal since the early 1900—even surpassing the calamity of the 1918 flu pandemic”? <https://statmodeling.stat.columbia.edu/2021/04/25/is-it-really-true/>.

<sup>24</sup><https://www.nytimes.com/interactive/2021/04/23/us/covid-19-death-toll.html>.

The third graph is labeled “Total deaths in the U.S. over time” and shows a baseline level of 1 million in 1917 with a jump to 1.4 million in the flu year of 1918, and baseline level of 2.8 million in 2019 with a jump to 3.3 million in 2020.

To summarize, here’s 1917 and 1918, reading roughly off the posted graphs:

- 1917: 2300 deaths per 100 000 and a total of 1 million deaths,
- 1918: 2600 deaths per 100 000 and a total of 1.4 million deaths.

This is an increase of 13% in the rate but an increase of 40% in the total. But we looked up the U.S. population and it seems to have been roughly constant between 1917 and 1918, so these above numbers can’t all be correct!

According to Wikipedia, the United States had 103 million people in 1917 and 1918. One million deaths divided by 103 million people is 1%, not 2.3%. So it’s not clear what is meant by “death rate” in that article.

The problem also arises in other years. For example, the article says that 3.4 million Americans died in 2020. Our population is 330 million, so, again, that’s a death rate of about 1%. But the 2020 death rate in their “Death rate in the U.S. over time” chart is less than 1%.

We can guess that their death rate graph is some sort of age-adjusted death rate . . . ummmm, yeah, OK, there it is at the bottom of the page:

“Death rates are age-adjusted by the C.D.C. using the 2000 standard population.”

Compared to 1918, the 2000 population has a lot of old people, and we have an even older population today. So the age-adjusted death rate *overweights* the olds compared to 1918 and slightly *underweights* the olds compared to 2020. The age adjustment to the 2000 population makes 1918 look not so bad because the 1918 flu was killing lots of young people, and young people are a relatively small proportion of the 2000 population.

To put it another way, it seems wrong for them to say that the 1918 flu wasn’t so bad because it killed lots of young people, and young people get downweighted in their adjustment. To use the 2000 population to assess the impact of the 1918 flu would be like using a modern weighting of the Consumer Price Index and then saying that inflation in 1918 was really low because the prices of airline flights weren’t going up at all back then.

Also, one other thing. The note at the bottom of the article says, “Expected rates for each year are calculated using a simple linear regression based on rates from the previous five years.” One reason why 1918 is not more above normal than it is, is that there happens to be an existing upward trend during the five years preceding 1918, so the implicit model would predict a further increase even in the absence of the flu. It’s not clear how to think about that.

Age adjustment can be tricky, as we’ve already seen in Section 2.4 of *Regression and Other Stories*. In this case, there may also have been a political motivation to stress the seriousness of covid in comparison to earlier pandemics.

This example relates to the week’s reading because we are looking at the properties of weighted averages, which are discussed in Section 3.1 of *Regression and Other Stories*. It is relevant to the course more generally because it demonstrates the importance of understanding exactly how data summaries are formed; as noted above, the same issues of weighting arise in constructing the Consumer Price Index and other measures of the economy and society.

<sup>25</sup>Francis Galton (1869), *Heredity Genius*, London: Macmillan, <https://galton.org/books/hereditary-genius/text/pdf/galton-1869-genius-v3.pdf>. The discussion here is taken from Andrew Gelman (2006), Galton was a hero to most, [https://statmodeling.stat.columbia.edu/2006/10/23/galton\\_was\\_a\\_he/](https://statmodeling.stat.columbia.edu/2006/10/23/galton_was_a_he/). For more on historical trends in height and body size, see, for example, Roderick Flood (1988), Height, weight, and body mass of the British population since 1820, National Bureau of Economic Research historical working paper 1018, <https://www.nber.org/papers/h0108>.

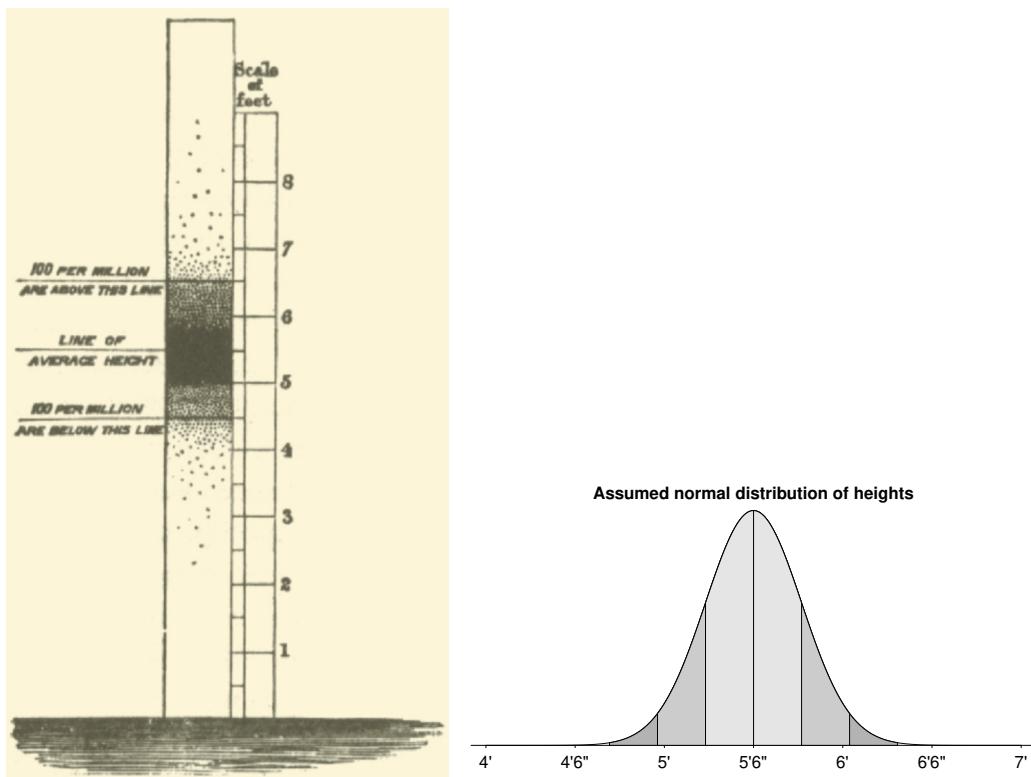


Figure 26 (a) From Francis Galton's 1869 book *Heredity Genius*, a visualization of the hypothesized distribution of heights of 1 million British men. The graph does not seem consistent with reality; it turned out that the problem was that Galton was unfamiliar with the tail properties of the normal distribution. (b) Normal distribution with mean 5'6" and standard deviation 3.23". Shaded areas show  $\pm 1$ , 2, and 3 standard deviations from the average. Under this distribution, fit to Galton's claim that 100 men per million would be below 5'6" and 100 per million above 6'6", we would not expect the extremes seen in the picture to the left.

## 2. Galton's giants

Figure 26a shows a graph from a classic book by Francis Galton, one of the fathers of applied statistics.<sup>25</sup> Galton, a cousin of Charles Darwin, is now controversial because of his racism and advocacy of eugenics; here, however, we focus on the more innocuous topic of the distribution of heights.

Figure 26a may look reasonable at first—it is centered at 5 feet 6 inches, the average height of men in Britain at that time—but a careful look shows problems. In particular, according to the picture, approximately one man in a million should be 9 feet tall! This doesn't make sense: if there were about 6 million men in England in Galton's time,<sup>26</sup> that would lead us to expect six 9-footers. As far as we know, there were no such giants roaming the land, and we assume Galton should have considered this when he was making the graph.

We asked two experts on the history of statistics to explain what happened. Antony Unwin wrote:

“Galton was postulating a hypothetical population with a normal distribution . . . he should have thought about the extreme values. On the other hand information in those days ways not so readily available as it is now. Might Galton have believed there were such people, he

<sup>26</sup>The population of the British Isles was about 30 million in 1869; excluding Ireland takes us to approximately 25 million; counting just males takes us down to about 12 million, and we can roughly assume half these were adult men.

just hadn't heard anything of them yet? The lower classes were not acknowledged and who knows what oddities might have been found amongst them?"

Stephen Stigler wrote:

"Galton was describing a hypothetical population, and he specifies for the illustration that the mean is 66 inches and that 100 per million exceed 78 inches. By my rough calculation that gives a standard deviation of about 3.2 inches. This was his earliest statistical book and Galton had more faith in the normal than later, but without good tables available (even though Laplace had given a continued fraction that would have given acceptable results) Galton did not appreciate how fast the tail comes down at the extremes. His view might have been colored by the fact that in the 1850s he had spent a couple of years in Africa, where there were and still are peoples of a quite wide variety of heights."

You can work out the calculations using R. Start with the claim that 100 men per million exceed 78 inches. To convert this into the normal distribution, we compute `qnorm(1e-4)`, which returns  $-3.71$ . The claim, then, is that 78 inches is 3.71 standard deviations higher than the mean of 66 inches; thus the standard deviation of heights would be  $(78 - 66)/3.71 = 3.23$ , as illustrated in Figure 26b. In that case, plugging in 9 feet (that is, 108 inches) yields  $(108 - 66)/3.23 = 13.0$ , so that a 9-footer would be 13 standard deviations above the mean. The probability of this occurring under the normal distribution is essentially zero: in R, `pnorm(-13.0)`, which returns the value  $6e-39$ , or  $6 * 10^{-39}$ .

What about seven-footers? If British men's heights were truly normally distributed with mean 66 inches and standard deviation 3.2 inches, then the proportion more than 7 feet tall would be  $1 - \text{pnorm}((84-66)/3.23)$ , or  $1.2 * 10^{-8}$ , which is still less than 1 in a million! In contrast, Galton's graph shows 20 men who are about 7 feet tall. This suggests that Galton did not know the tails of the normal distribution; he was just guessing and gave a distribution with tails that were too wide. In addition, his dots are too uniform in the range between 5 feet 10 inches and 6 feet 6 inches; in reality, there would be many more men at the lower end of that band and a lot fewer at the high end.

At this point, you might say we're just being picky, as Galton was just making his graph (using technology of the 1860s!) to make a general point about variation. The message we want to convey here is that assumptions have consequences. The normal distribution for men's heights might be a reasonable approximation (depending on your purpose); if you use this distribution, it has specific implications. Comparing the model's predictions to data can reveal problems with the model but only if we are careful about getting the predictions right.

The connection of this example to this week's reading is the use of R to calculate properties of the normal distribution. We did several calculations, first using the average and a quantile to deduce the standard deviation of the distribution, then using the average and the standard deviation to compute tail probabilities. The connection to the larger themes of the course is that probability models make specific predictions that can be checked.

### Class-participation activities

#### 1. Amoebas and exponential growth and decline

Suppose you have an amoeba that takes an hour to divide, and then the two amoebas each divide in one more hour, and so forth.<sup>27</sup> What is the equation of the number of amoebas,  $y$ , as a function of time,  $x$  (in hours)? It can be written as  $y = 2^x$  or, on the logarithmic scale,  $\log y = (\log 2) * x = 0.69x$ .

Suppose you have the same example, but the amoeba takes three hours to divide at each step.

<sup>27</sup>From Section 3.8.1 of *Teaching Statistics: A Bag of Tricks*.

### 3.4. REVIEW OF MATHEMATICS AND PROBABILITY

73

Then the number of amoebas  $y$  after time  $x$  has the equation,  $y = 2^{x/3} = (2^{1/3})^x = 1.26^x$  or, on the logarithmic scale,  $\log y = (\log 1.26) * x = 0.23x$ . The slope of 0.23 is one-third the earlier slope of 0.69 because the population is growing at one-third the rate.

We then ask students to work in pairs to come up with other examples of exponential growth or decline. They should come up with reasonable values for the parameters  $A, b$  or  $a, b$  in the equations  $y = A \exp(bx)$  or  $\log y = a + bx$ . They should also come up with examples of exponential decline.

For example, consider an asset that is initially worth \$1000 and declines in value by 20% each year. Then its value at year  $x$  can be written as  $y = 1000 * 0.8^x$  or, equivalently,  $y = 1000 e^{\log(0.8)x} = 1000 e^{-0.22x}$ . Taking the log of both sides yields  $\log y = \log 1000 - 0.22x = 6.9 - 0.22x$ .

#### 2. Squares, cubes, and power-law growth

The formula  $\log y = a + b \log x$  represents power-law growth if  $b > 0$  or decline if  $b < 0$ : Equivalently this model can be written as  $y = Ax^b$ , where  $A = \exp(a)$ . The parameter  $A$  is the value of  $y$  when  $x = 1$ , and the parameter  $b$  determines the rate of growth or decline. A one-unit difference in  $\log x$  corresponds to an additive difference of  $b$  in  $\log y$ .

Here are two examples:<sup>28</sup>

- *Power law.* Let  $y$  be the area of a square and  $x$  be its perimeter. Students should visualize this in pairs by sketching a square on paper. Then  $y = (x/4)^2$ , and you can take the log of both sides to get  $\log y = 2(\log x - \log 4) = -2.8 + 2 \log x$ .
- *Non-integer power law.* Let  $y$  be the surface area of a cube and  $x$  be its volume. Students should in pairs sketch a cube. If  $L$  is the length of a side of the cube, then  $y = 6L^2$  and  $x = L^3$ , hence the relation between  $x$  and  $y$  is  $y = 6x^{2/3}$ ; thus,  $\log y = \log 6 + \frac{2}{3} \log x = 1.8 + \frac{2}{3} \log x$ .

We then ask students to work in pairs to come up with social-science examples of power-law growth or decline. They should come up with reasonable values for the parameters  $A, b$  or  $a, b$  in the equations  $y = Ax^b$  or  $\log y = a + b \log x$ . An example is elasticity in economics. Suppose that if you increase the price of an item by 1%, its sales decline by 2%. That's  $\log y = a - 2 \log x$ . You can understand this further by graphing hypothetical data consistent with this or other elasticities.

The next step is for each student in a pair to make up and graph an example of power-law growth or decline and for the other student to figure out the mathematical relationship. Then they switch roles, giving each student the opportunity to simulate and decode the pattern.

### Computer demonstrations

#### 1. Matrix manipulations

You can predict the national popular vote in a presidential election conditional on economic growth, using the fitted the ElectionsEconomy model, using vectors and matrices.<sup>29</sup>

```
# Vectors
a_hat <- 46.3
b_hat <- 3.0
x <- c(-1, 0, 3)
y_hat <- a_hat + b_hat * x
print(y_hat)

# Matrices
X <- cbind(c(1, 1, 1), c(-1, 0, 3))
```

<sup>28</sup>From Section 3.8.1 of *Teaching Statistics: A Bag of Tricks* and Section 3.4 of *Regression and Other Stories*.

<sup>29</sup>From Section 3.2 of *Regression and Other Stories*.

```
coef_hat <- c(46.3, 3.0)
y_hat <- X %*% coef_hat
print(y_hat)
```

## 2. Compute weighted averages using vectors and matrices

One way to better understand weighted averages is to compute them in different ways.<sup>30</sup>

```
## Example 1: age in North America

population <- c(310, 112, 34) # US, Mexico, Canada
average_age <- c(36.8, 26.7, 40.7)
# Approach 1 (using raw quantities)
weighted_average <- sum(population * average_age) / sum(population)
print(weighted_average)
# Approach 2 (using relative weights and dot product)
weights <- population / sum(population)
weighted_average <- sum(weights * average_age)
print(weighted_average)
# Approach 3 (using matrix multiplication)
weighted_average <- weights %*% average_age
print(weighted_average)

## Example 2: age in the U. S. (using dot product)

weights <- c(0.51, 0.49) # female, male
average_age <- c(38.1, 35.5)
weighted_average <- sum(weights * average_age)
print(weighted_average)

## Example 3: teachers' salary in the U.S. (using matrix multiplication)

weights <- c(0.79, 0.21) # female, male
average_income <- c(45865, 49207)
weighted_average <- weights %*% average_income
print(weighted_average)

## Example 4: age, height, and weight in the U.S.

weights <- c(0.51, 0.49) # female, male
attributes <- cbind(c(38.1, 35.5), c(63.5, 69.1), c(170.6, 197.9))
  # age, height, weight
weighted_average <- weights %*% attributes
print(weighted_average)
```

## Drills

### 1. Graphs of straight lines

- (a) Write R code to graph the line,  $y = 250 + 0.35x$ , for  $x$  in the range 0 through 10.

*Solution:*

```
curve(250 + 0.35x, from=0, to=10)
```

- (b) Write R code to graph the three lines,  $y = 250 + bx$ , for  $x$  in the range 0 through 10, and for  $b$  taking on the values 0.1, 0.2, and 0.3.

<sup>30</sup>From Section 3.1 of *Regression and Other Stories*.

### 3.4. REVIEW OF MATHEMATICS AND PROBABILITY

75

- (c) Do the previous problem, using different colors for the three lines.
- (d) Repeat the previous problem, now making sure that your  $y$ -axis ranges from 249 to 255.

#### 2. Normal distribution

For each of the following questions, graph the distribution of  $y$ , estimate roughly the probability that  $y$  falls in the specified interval, and give R code to compute the probability that  $y$  falls in the specified interval. All distributions are normal distributions.

- (a) Distribution: mean 500 and standard deviation 100. Interval:  $(-\infty, 600)$

*Solution:* Sketch the normal distribution. 600 is 1 standard deviations above the mean, so the probability of  $y$  being more than 600 is 16%, hence the probability that  $y$  is less than 600 must be 84%.

In R, `pnorm(600, 500, 100)`

- (b) Distribution: mean 500 and standard deviation 100. Interval:  $(500, \infty)$

- (c) Distribution: mean 500 and standard deviation 100. Interval:  $(300, 500)$

- (d) Distribution: mean 0 and standard deviation 2. Interval:  $(-2, 2)$

- (e) Distribution: mean 0 and standard deviation 2. Interval:  $(-\infty, -4)$

- (f) Distribution: mean 0 and standard deviation 2. Interval:  $(-4, -2)$

#### Discussion problems

##### 1. College admissions and weighted averages

Admissions to university are based on many factors, two of which are the most important are the score on a standardized admissions test and high school grade point average. Suppose you want to combine these into a single score, using a weighted average of the form  $a * (\text{Test score}) + b * (\text{Grade point average})$ , where test scores range from 400 to 1600 and grade point averages range from 0 to 4. What would be reasonable values of  $a$  and  $b$  so that the two factors counts somewhat equally in the average?

##### 2. Estimate the probability of a rare event

Section 3.6 of *Regression and Other Stories* discusses how to estimate the probability of a tied election. Elections are sometimes close but are rarely exactly tied, and so the best way of estimating this probability is not purely empirically but rather using a combination of empirical frequencies and mathematical modeling to connect the rare event of interest to a more frequent “precursor event” which occurs more frequently. Instead of directly counting the number of times an election has been tied, you can see how often an election is decided within 10 000 votes and then divide that frequency by 10 000 to get an estimated probability of an exact tie.

For this discussion problem, give another example like this, estimating the probability of some rare event by connecting it to a more common precursor.

## 3.5 Statistical inference

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
They got the wrong standard error	Design a bogus study	Simulate fake data and conf interval	Binomial model for basketball	Confidence intervals and true values
Claims of implausibly large effects	Think about effect sizes	Proportions, means, and differences	Sample size and standard errors	Standard error for “feeling thermometers”

### Reading

Chapter 4 of *Regression and Other Stories*: Statistical inference

### Pre-class warmup assignments

#### 1. Confidence intervals from the binomial distribution

A sample of  $n$  people are selected at random from a large population and are asked a question, to which  $y$  reply Yes and  $n - y$  reply No. For each example below, give a 95% interval for the proportion in the population who would answer Yes if asked.

- (a)  $n = 1500, y = 750$
- (b)  $n = 1500, y = 900$
- (c)  $n = 100\,000, y = 51\,000$
- (d)  $n = 10, y = 6$
- (e)  $n = 3, y = 3$

#### 2. Coverage of confidence intervals

Based on the normal approximation, the estimate  $\pm 2$  standard errors gives an approximate 95% interval. Similarly:

- (a) What is the coverage of the estimate  $\pm 1$  standard error?
- (b) For what value  $x$  does the estimate  $\pm x$  standard error have 50% coverage?
- (c) For what value  $x$  does the estimate  $\pm x$  standard error have 80% coverage?

### Homework assignments

#### 1. (a) Probability distributions (Exercise 3.5 of *Regression and Other Stories*)

Using a bar plot in R, graph the binomial distribution with  $n = 20$  and  $p = 0.3$ .

#### (b) Linear transformations (Exercise 3.6 of *Regression and Other Stories*)

A test is graded from 0 to 50, with an average score of 35 and a standard deviation of 10. For comparison to other tests, it would be convenient to rescale to a mean of 100 and a standard deviation of 15.

- i. Labeling the original test scores as  $x$  and the desired rescaled test score as  $y$ , come up with a linear transformation, that is, values of  $a$  and  $b$  so that the rescaled scores  $y = a + bx$  have a mean of 100 and a standard deviation of 15.
- ii. What is the range of possible values of this rescaled score  $y$ ?
- iii. Plot the line showing  $y$  vs.  $x$ .

(c) Linear transformations (Exercise 3.7 of *Regression and Other Stories*)

Continuing the previous exercise, there is another linear transformation that also rescales the scores to have mean 100 and standard deviation 15. What is it, and why would you *not* want to use it for this purpose?

2. (a) Comparison of proportions (Exercise 4.1 of *Regression and Other Stories*)

A randomized experiment is performed within a survey. 1000 people are contacted. Half the people contacted are promised a \$5 incentive to participate, and half are not promised an incentive. The result is a 50% response rate among the treated group and 40% response rate among the control group. Give an estimate and standard error of the average treatment effect.

(b) Choosing sample size (Exercise 4.2 of *Regression and Other Stories*)

You are designing a survey to estimate the gender gap: the difference in support for a candidate among men and women. Assuming the respondents are a simple random sample of the voting population, how many people do you need to poll so that the standard error is less than 5 percentage points?

(c) Comparison of proportions (Exercise 4.3 of *Regression and Other Stories*)

You want to gather data to determine which of two students is a better basketball shooter. One of them shoots with 30% accuracy and the other is a 40% shooter. Each student takes 20 shots and you then compare their shooting percentages. What is the probability that the better shooter makes more shots in this small experiment?

(d) Designing an experiment (Exercise 4.4 of *Regression and Other Stories*)

You want to gather data to determine which of two students is a better basketball shooter. You plan to have each student take  $N$  shots and then compare their shooting percentages. Roughly how large does  $N$  have to be for you to have a good chance of distinguishing a 30% shooter from a 40% shooter?

## Stories

1. A consulting project where they got the wrong standard error

We once worked on a consulting project on a legal case involving sampling of insurance claims. We were asked to review a report written by a consultant on the other side. After a quick read of that other report, we realized its numbers were wrong, but it took us a few hours to figure out how they made the mistake that they made. It was important for us when writing our report to not just see the mistake that was made but to understand how it had happened, because without that full understanding we could not be sure that we hadn't missed anything.

The report involved a sample of 152 claims from a larger population, and the results were presented as a series of proportions with margins of error (that is,  $\pm 2$  standard errors); Figure 27 shows some of the results. For example, one of the estimates was 45.7% with a margin of error of 1.9%, thus a standard error of 0.9% or 1.0%. We knew this was wrong because the standard error of  $y/n$  for a simple random sample of size 130 with proportion 0.457 is  $0.5/\sqrt{152} = 0.041$ , that is, 4.1 percentage points. The reported standard error of 0.009 or 0.010 was too low by a factor of 4. The actual survey was more complicated, as it involved stratified sampling—we will not go into the details of the sampling here, except to say that in this case we would actually expect the standard error to be a bit higher, not lower, than 0.041.

So how did they get a standard error of 0.009 or 0.010 in that report? That sample of  $n = 152$  was taken from a larger population of size  $N$ . It seems they had mistakenly divided by  $\sqrt{N}$  instead of  $\sqrt{n}$  when computing the standard error, and we trace this error to the blind use of a formula, along with a lack of experience. It is standard in opinion polls to sample 1000–1500 people, which yields a margin of error (2 standard errors) of 3 percentage points; we've seen enough of these to know that a margin of error of 1.9 percentage points from a sample of 152 did not look right.

	Response Yes	Margin of Error + / -
Was the SIU assigned to the case?	4.1%	0.5%
Were other anti-fraud professionals assigned or alerted?	2.0%	0.3%
Was there an indication in file of suspected fraud, particularly with regard to a staged accident or exaggerated medical care, medical bills, and loss or earnings?	45.7%	1.9%

Figure 27 Snapshot from a report summarizing inferences from a sample of 152 insurance claims. The “margin of error” is intended to represent 2 standard errors; given the sample size, the numbers in this column are suspiciously low.

This story is relevant to the week’s reading as an example of computing uncertainty in an estimated proportion, and it relates to the course as a whole as a demonstration of the risks of mindless use of formulas.

## 2. Claims of implausibly large effects

Recent decades have seen many high-profile claims in the social and biological sciences that, in retrospect, would represent implausibly large effects. Figure 28 gives several examples.<sup>31</sup> Quickly going over these four stories can give some insight into effect sizes, statistical significance, and selection bias.

*The first example* is a study comparing a group of toddlers from low-income families who received psychosocial stimulation to a comparable group who received no such treatment. Both groups were followed up for twenty years. The study reports that early educational intervention increased average earnings by 42% when the participants were young adults. Moreover, the earnings of the stimulated underprivileged group caught up with those of their better-off peers, suggesting that early intervention is a key driver in reducing inequality. There are sweeping political implications here—namely, how much should a government invest in early childhood education? There remains a lot of controversy in education research about what works and what doesn’t, and for which students. Long-term experiments on children’s lives are costly in time, money, and human resources. As a result, major decisions on education policy can turn on the statistical interpretation of small, idiosyncratic data sets—in this case, a study of 129 Jamaican children.

Before considering the politics, though, we need to look at the methodology. The problem is that there are many ways of looking at the data in this sort of study, and this leads to overestimates of effect sizes. It’s just the nature of scientific reporting: estimates near zero remain unpublished or get adjusted higher (based on decisions arising from reasonable scientific judgments), while high

<sup>31</sup>Paul Gertler, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeesch, Susan Walker, Susan Chang, and Sally Grantham-McGregor (2013), Labor market returns to early childhood stimulation: A 20-year followup to an experimental intervention in Jamaica, National Bureau of Economic Research working paper 19185, <https://www.nber.org/papers/w19185>; see discussion in Section 1.5 of *Regression and Other Stories* and at Andrew Gelman (2013), Childhood intervention and earnings, *Symposium*, <http://www.symposium-magazine.com/childhood-intervention-and-earnings/>. Alec Beall and Jessica Tracy (2013), Women are more likely to wear red or pink at peak fertility, *Psychological Science*, 24, 1837–1841; see discussion at Andrew Gelman (2013), Too good to be true, *Slate*, 24 July, <https://slate.com/technology/2013/07/statistics-and-psychology-multiple-comparisons-give-spurious-results.html>. Lysann Damisch, Barbara Stoberock, and Thomas Mussweiler (2010), Keep your fingers crossed!: How superstition improves performance, *Psychological Science* 21, 1014–1020; see discussion at Andrew Gelman (2021), The so-called “lucky golf ball”: The Association for Psychological Science promotes junk science while ignoring the careful, serious work of replication, <https://statmodeling.stat.columbia.edu/2021/12/20/not-replicable-but-citable/>. Kristina Durante, Ashley Rae, and Vladas Griskevicius (2013), The fluctuating female vote: Politics, religion, and the ovulatory cycle, *Psychological Science* 24, 1007–1016; see discussion in Section 4.5 of *Regression and Other Stories* and at Andrew Gelman (2015), The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective, *Journal of Management* 41, 632–643.

- (a) “*Labor Market Returns to Early Childhood Stimulation*”:  
“We find large effects . . . from a randomized intervention that gave psychosocial stimulation to stunted Jamaican toddlers living in poverty. The intervention consisted of one-hour weekly visits from community Jamaican health workers over a 2-year period that taught parenting skills . . . We re-interviewed the study participants 20 years after the intervention. Stimulation increased the average earnings of participants by 42 percent . . . ”
- (b) “*Women are More Likely to Wear Red or Pink at Peak Fertility*”:  
“Across two samples ( $N = 124$ ), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk . . . ”
- (c) “*Keep Your Fingers Crossed!: How Superstition Improves Performance*”:  
“Experiments 1 through 4 show that activating good-luck-related superstitions via a common saying or action (e.g., ‘break a leg,’ keeping one’s fingers crossed) or a lucky charm improves subsequent performance in golfing, motor dexterity, memory, and anagram games . . . ”
- (d) “*The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle*”:  
“Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women’s politics, religiosity, and voting in the 2012 U.S. presidential election. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney. . . . ”

**Figure 28** *Titles and extracts from four published studies claiming to find implausibly large effects: (a) a claim that a weekly intervention on preschool children increases their later adult income by an average of 42%; (b) a claim that women are three times more likely to wear red or pink during certain times of the month; (c) a claim that a “lucky golf ball” increases putting accuracy by 35%; (d) a claim that single women were 20 percentage points more likely to support Barack Obama during certain times of the month.* We project this onto the screen during a discussion of inflated estimates and forking paths in statistical analyses.

estimates remain. We have every reason to think the effect of childhood stimulation is positive for most children, but we can only assume that the 42% number is an overestimate.

There are two issues with this study. First, given the uncertainty in the study, the bias in the estimate could be huge. Consider that published results typically are required to be “statistically significant,” that is, at least 2 standard errors from zero, and the estimated effect of 0.42 has a standard error of approximately 0.20.<sup>32</sup> Thus, in an experiment with this level of uncertainty, any statistically-significant estimate would have to be at least 0.40, that is, an estimated effect of at least 40% on adult earnings.

The second problem is that, given the selection involved in how to code and analyze the data and what summaries to report, it would not be difficult for the researchers to find a “statistically significant” result even in the absence of any effect. Just to be clear: we’re not saying we believe the treatment had no effect, just that, given the design of the study and the flexible analysis plan, we do not learn much from reading that a particular comparison is more than two standard errors away from zero. For example, if the true effect size—the average gain in adult earnings from the intervention—is 5%, then any estimate from this study will be essentially noise. We explain this point in further detail in Chapter 16 of *Regression and Other Stories*.

*The next example* in Figure 28 is a claim that women were three times more likely to wear red or pink during certain times of the month. The claimed effect is implausibly large, especially given that some people in this study might not regularly wear any red or pink clothing at all, others might be restricted in clothing choices by their regular routines. Even if clothing choices vary

<sup>32</sup>This sort of analysis would typically be done on the logarithmic scale so that the uncertainty would be multiplicative, not additive, a point we discuss on the top of page 15 of *Regression and Other Stories*, but this technical issue does not affect the point of this story, so we for simplicity set it aside here.

during the month for some subset of women, there is essentially no way to get to an *average* effect of a factor of 3. However, it is easy to obtain an *estimated* effect of this size with a noisy study.

It goes like this. First, the sample size is small and the measurement is highly variable, thus the estimated ratio has a high standard error, so just by chance it is possible to observe a ratio as high as 3. Second, there were several inconsistencies in the rules for which data to use in the analysis. For one of the samples in the study, 9 of the 24 women did not meet the inclusion criteria of being more than five days away from onset of menses, but they were included anyway. And even though another sample was supposed to be restricted to women younger than 40, the ages of the women included ranged up to 47. Out of all the women who participated across the two samples, 31% were excluded for not providing sufficient precision and confidence in their answers.

In addition, the authors found a statistically significant pattern after combining red and pink, but had they found it only for red, or only for pink, this would have fit their theories too. In their words: “The theory we were testing is based on the idea that red and shades of red (such as the pinkish swellings seen in ovulating chimpanzees, or the pinkish skin tone observed in attractive and healthy human faces) are associated with sexual interest and attractiveness.” Had their data popped out with a statistically significant difference on pink and not on red, that would have been news too. And suppose that white and gray had come up as the more frequent colors? One could easily argue that more bland colors serve to highlight the pink colors of a (European-colored) face.

This sort of study is like a lottery in which there are many different ways to win, and the winning conditions are not specified ahead of time. Combine this research flexibility with the variability in the estimate, and you obtain an estimate that is more than two standard errors away from zero, and thus will be unrealistically huge and provide essentially no information about any possible underlying effect. Indeed, a later study by the same authors failed to replicate the effect.

We are not saying that the scientific claims in these papers are necessarily wrong. What we are saying is that the evidence in these research papers is not as strong as stated. The scientific hypotheses in the papers at hand are general enough that they could have some validity even if the particular published claims do not hold.

We go through the third and fourth examples in Figure 28 more quickly, as they have the same issues as discussed above.

The “lucky golf ball” experiment posits an effect that could possibly be real (golfers performing better when they have more confidence) but could also plausibly be zero or even negative (golfers are already trying their best, and it is not clear that any positive effect of confidence is overwhelmed by negative effects of overconfidence), hence it seems reasonable to study the question experimentally. The problem is that the study in question is too small: the standard error of the resulting estimate is large (recall that the standard error scales like  $1/\sqrt{n}$  and so will be large when  $n$  is small), so the noise overwhelms any realistically-sized signal. In the event, the estimated effect was to improve the probability of success by 35%, an implausible amount given the difficulty of the task. Again, there is no reason to believe this as an estimate of the true effect, but it is easy to see how this estimate can arise from noise, given that there were only 28 people in the study. A later study failed to replicate the finding. We return to this example later in the course; see page 216.

The final example is the claim that 20% of women change their political views during their monthly cycle. This raises our skepticism, given that in polls that survey people more than once, the percentage of people who change their opinion during the final weeks of a campaign is closer to 2%.<sup>33</sup> As with the other examples we have just discussed, the way this unrealistically large estimate arose was that the study was small and noisy, hence the estimated effect was highly variable. And, again, there were so many options in the data coding and analysis that it would be

<sup>33</sup>See, for example, Andrew Gelman, Sharad Goel, Douglas Rivers, and David Rothschild (2016), The mythical swing voter. *Quarterly Journal of Political Science* 11, 103–130.

Rolf Zwaan's steps to produce a clickbait research finding:

1. The idea, based on some popular saying.
2. Theoretical background. Find some remotely relevant connection.
3. The manipulation. Take the expression literally.
4. Outcome measure. Use something fun like candy.
5. Participants in your experiment. Can be anyone.
6. Run experiment 1.
7. Analyze the results. Look for something big in the data.
8. Design experiment 2. Pick a new manipulation.
9. Pick a fun new outcome measure.
10. Repeat steps 5–7.
11. Write your general discussion.
12. Add a quirky celebrity quote.
13. Come up with an amusing title.
14. Hype your findings by overgeneralizing.

**Figure 29** Compressed version of satirical advice for designing a successful social science experiment. The instructor should display this on the screen and then talk through the “half empty or half full” example, in preparation for students working together to come up with their own pseudoscience studies.

possible to find some comparison that is two standard errors away from zero. We discuss this in detail in Section 4.5 of *Regression and Other Stories*.

These examples are relevant to the week’s reading in showing how context can limit what can be learned from any given study; this connects statistical modeling and scientific modeling for causal inference, as will be discussed later in the course. It relates to the course as a whole in demonstrating the challenges of extrapolating from data analysis to the real world, even in the case of randomized experiments. We give four different examples so as to get a sense of the generality of the problem.

### Class-participation activities

#### 1. Design a bogus social science study

In 2013, researcher Rolf Zwaan came up with a facetious plan to design and conduct a bogus psychology experiment.<sup>34</sup> The instructor can display Figure 29, which gives an abbreviated version of his list, and then explain that students will be asked to follow this template to come up with their own ideas for bogus research projects. To give them a sense of how to do this, the instructor can talk through the particular example used by Zwaan:

1. The idea. “All you need to do is take an idiomatic expression and run with it. Here we go: the glass is half-full or the glass is half-empty.”
2. Theoretical background. “Surely there is some philosopher (preferably a Greek one) who has said something remotely relevant about optimists and pessimists while staring at a wine glass. Include him. . . . Google is your friend here.”
3. The manipulation. “All you need to do is take the expression literally. . . . The subject is in a room. In the glass-full condition, a confederate comes in with an empty glass and a bottle

<sup>34</sup>Rolf Zwaan (2013), How to cook up your own social priming article, <https://rolfzwaan.blogspot.com/2013/09/how-to-cook-up-your-own-social-priming.html>.

of water. She then pours the glass half full and leaves the room. In the glass-half-empty condition, she comes in with a full glass and a bottle. She then pours half the glass back into the bottle and leaves.”

4. Outcome measure. “Let’s say the subjects get to choose ten pieces of differently colored pieces of candy from a container that has equal numbers of orange and brown M&Ms. Your prediction here is that people in the half-full condition will be more likely to pick the cheery orange M&Ms than those in the half-empty condition, who will tend to prefer the gloomy brown ones.”
5. Participants in your experiment. “About 30 students from a nondescript university will do nicely.”
6. Run experiment 1. “Don’t fuss about . . . details of the procedure; you won’t be reporting them anyway.”
7. Analyze the results. “Normally, you’d worry that you might not find an effect. But this is social priming remember? You are guaranteed to find an effect. In fact, your effect size will be around 0.8. That’s social priming for you!”
8. Design experiment 2. “Come up with a new manipulation. What’s wrong with the glass and bottle from Experiment 1?, you might wonder. Are you kidding? . . . How about balloons? In the half-full condition, the confederate walks in with an inflated balloon and lets half the air out in front of the subject. In the half empty condition, she half-inflates a balloon.”
9. Pick a fun new outcome measure. “Why not have the subjects list their favorite TV shows? Your prediction here is that the half-full condition will list more sitcoms like Seinfeld and Big Bang Theory than the half-empty condition, which will list more crime shows like CSI and Law & Order. . . . How about having subjects indicate how much they identify with Winnie the Pooh characters? Your prediction here is obvious: the half full condition will identify with Tigger the most while the half empty condition will prefer Eeyore by a landslide.”
10. Repeat steps 5–7.
11. Write your general discussion. “Don’t be shy here. Talk about the major implications for business, health, education, and politics this research so evidently has.”
12. Add a quirky celebrity quote. “Just go to [www.goodreads.com](http://www.goodreads.com) to find a quote. Here, I already did the work for you: ‘Some people see the glass half full. Others see it half empty. I see a glass that’s twice as big as it needs to be.’—George Carlin. Just say something clever like: Unless you are like George Carlin, it does make a difference whether the glass is half empty or half full.”
13. Come up with an amusing title. “Just use the expression from Step 1 as your main title, describe your (huge) effect in the subtitle and you’re done: ‘Is the glass half empty or half full? The effect of perspective on mood.’”
14. Hype your findings by overgeneralizing. “Like all social priming research, your work has profound consequences for all aspects of society.”

Zwaan concludes, “Once you’ve worked through this example, you might try your hand at more advanced topics like coming out of the closet. Imagine all the fun you’ll have with that one!”

After hearing the half-full/half-empty story, it’s time for the students to pair up and come up with their own examples. The instructor can walk around the room, listening to the students’ ideas and helping them get unstuck as needed, and then choose one pair’s plan and go through it with the entire class. Yes, such a study would be a joke, but the point is to understand how it is possible to construct an experiment with no scientific value but which could still get publication and publicity. The class can follow up with a discussion of the ways in which these bogus studies differ from serious research.

1. Consider a topic of interest
2. Consider an outcome measure and hypothesize a treatment effect
3. Construct a hypothetical experiment
4. Specify sample size
5. Hypothesize distribution of outcomes under control and treatment
6. Figure out estimate and standard error
7. Will the experiment give a reliable estimate?

Figure 30 *Steps for the activity on quantitative thinking about effect sizes. This should be projected on the screen or written on the board to guide students in constructing their examples.*

This activity relates to the week’s reading as an explanation of how misinterpretation of statistical significance, as discussed in Section 4.5 of *Regression and Other Stories*, can lead to problems in the scientific literature. It relates to the course as a whole by connecting statistical design and analysis to practices of scientific and popular communication.

2. Think about effect sizes in the context of a social science example

The point of this activity is to get students thinking about effect sizes and statistical parameters in a quantitative way: not just that an effect is positive or negative, but how large it is and how large a study would be needed to estimate it reliably. To do this, the instructor should project Figure 30 on the screen and explain that students will be expected to do this sequence of steps in pairs.

First it is best to go through for a simple example. Consider a new way to teach reading to first-graders, where the hypothesized effect is to raise test scores by 2 points on a 0–100 scale. A simple experiment would randomly assign students into the treatment or control group. Suppose an experiment were to be conducted on 100 students, with 50 in each group.

Will this be enough data to reliably estimate the effect? To answer this question we need some sense of the distribution of test scores. Suppose that, under the control, scores have a distribution with mean 60 and standard deviation 15 (so that approximately two-thirds of students have scores between . . . ask the students to figure this out . . . it’s “between 45 and 75”), and further suppose scores under the treatment will follow a distribution with mean 62 and standard deviation 15. Then the difference between the average of 50 treated students and 50 control students will be approximately normally distributed with mean 2 and standard deviation  $\sqrt{15^2/50 + 15^2/50} = 3$ . In this case, the standard error of the estimate is larger than the effect size, which tells us that the experiment is not precise enough to reliably estimate the treatment effect.

As discussed in Chapter 16 of *Regression and Other Stories*, the usual rule is that the treatment effect should be more than 2.8 standard errors away from zero. In addition, we would recommend that such an experiment begin with a pre-test on each student, which can then be included in a regression model to allow the treatment to be estimated more precisely. But here is no need to go into all this at this point in the course; the point of this exercise is, first, to get students thinking quantitatively about effect sizes, averages, and standard deviations; and second, to use and understand the distribution of the difference between two averages.

Once the instructor has gone through an example, the students can work in pairs to construct their own examples of realistic effect sizes and distributions, following the steps in Figure 30.

This activity relates to the discussions of statistical inference in Chapter 4 of *Regression and Other Stories*, and it is relevant to the course as a whole in connecting statistical methods to substantive questions.

### Computer demonstrations

#### 1. Simulate fake data and compute confidence interval, plus looping

This demonstration teaches three lessons: simulation from the binomial distribution, computing the standard error and confidence interval from a proportion, and looping, as discussed in Section 4.2 of *Regression and Other Stories*.

```
# Generate fake data
p <- 0.3
n <- 20
data <- rbinom(1, n, p)
print(data)

# Estimate proportion and calculate confidence interval
p_hat <- data/n
se <- sqrt(p_hat * (1-p_hat) / n)
ci <- p_hat + c(-2, 2) * se
print(ci)

# Put it in a loop
reps <- 100
for (i in 1:reps){
  data <- rbinom(1, n, p)
  p_hat <- data/n
  se <- sqrt(p_hat * (1-p_hat) / n)
  ci <- p_hat + c(-2, 2) * se
  print(ci)
}
```

#### 2. Statistical inference: proportions, means, and differences of means

We use a Pew Research survey to estimate a proportion, a mean, and a difference in means.

```
# Read data from here: https://github.com/avehtari/ROS-Examples
library("foreign")
library("dplyr")
pew_pre <- read.dta(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Pew/data/",
  "pew_research_center_june_elect_wknd_data.dta"))
pew_pre <- pew_pre %>% select(c("age", "regicert")) %>%
  na.omit() %>% filter(age != 99)
n <- nrow(pew_pre)

# Estimate a proportion (certain to have registered for voting?)
registered <- ifelse(pew_pre$regicert=="absolutely certain", 1, 0)
p_hat <- mean(registered)
se_hat <- sqrt((p_hat * (1 - p_hat)) / n)
round(p_hat + c(-2, 2) * se_hat, 4) # ci

# Estimate an average (mean age)
age <- pew_pre$age
y_hat <- mean(age)
se_hat <- sd(age) / sqrt(n)
round(y_hat + c(-2, 2) * se_hat, 4) # ci

# Estimate a difference of means
```

```
age2 <- age[registered==1]
age1 <- age[registered==0]
y_2_hat <- mean(age2)
se_2_hat <- sd(age2) / sqrt(length(age2))
y_1_hat <- mean(age1)
se_1_hat <- sd(age1) / sqrt(length(age1))
diff_hat <- y_2_hat - y_1_hat
se_diff_hat <- sqrt(se_1_hat^2 + se_2_hat^2)
round(diff_hat + c(-2, 2) * se_diff_hat, 4) # ci
```

## Drills

### 1. Binomial distribution (example of basketball shots)

A basketball player takes  $n$  shots. The shots are independent and she has a 30% chance of making each shot. Let  $y$  be the number of shots she makes. What are the mean and standard deviation of  $y$ ? Sketch the distribution of  $y$ .

(a)  $n = 20$

*Solution:* Mean is  $0.3 * 20 = 6$ , standard deviation is  $\sqrt{0.3 * 0.7 * 20} = 2.0$ . Sketch a bell-shaped curve centered at 6 with standard deviation 2, then sketch the bar graph corresponding to the discrete distribution with possible values 0, 1, 2, etc.

(b)  $n = 50$

(c)  $n = 100$

(d)  $n = 0$

(e)  $n = 1$

(f)  $n = 2$

### 2. Sample size and standard errors

(a) In a national survey of  $n$  people, how large does  $n$  have to be so that you can estimate presidential approval to within a standard error of  $\pm 3$  percentage points?

*Solution:* Standard error is  $0.5/\sqrt{n}$ . Try  $n = 100$ , then the standard error is  $0.5/\sqrt{100} = 0.05$ . We want 0.03, so we need to increase sample size by a factor of  $(0.5/0.3)^2 = 2.78$ . So we need a sample size of 278.

(b)  $\pm 1$  percentage point?

(c) How large does  $n$  have to be so that you can estimate the gender gap in approval to within a standard error of  $\pm 3$  percentage points?

(d)  $\pm 1$  percentage point?

(e) What if you're estimating something more rare, such as the percentage of people who have ever run for office?

## Discussion problems

### 1. Confidence intervals and true parameter values

Suppose you do 1000 experiments and, from each, you get a 95% interval. You'd expect 950 of these intervals to contain the true parameter values. Assuming your statistical model is correct, would it be a surprise if only 925 of these intervals contained the true parameter values?

### 2. Approximate standard error for average “feeling thermometer” ratings

The American National Election Study and other surveys ask this sort of “feeling thermometer” question: “I’d like to get your feelings toward some of our political leaders and other people who are in the news these days. I’ll read the name of a person and I’d like you to rate that person using

something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person."

In a survey of 1500 people, what might you expect the standard error for the average feeling thermometer rating to be, approximately? You should figure this out in three steps: first, sketch a plausible distribution of feeling thermometer ratings among survey respondents; second, estimate the standard deviation of this distribution using the rule that approximately two-thirds of the responses should be in the range of the mean  $\pm$  one standard deviation; third, compute the standard error of the average as the standard deviation divided by  $\sqrt{n}$ . This standard error is a relevant baseline when comparing average ratings from surveys taken at different times.

## 3.6 Simulation

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Proportion of identical twins	Real vs. fake coin flips	Break R functions	Random sampling and looping in R	Discrete / continuous distribution
Simulate a process of innovation	Simulate a probability process	Simulate 100 coin flips	Propagate uncertainty	Simulate clustering of buses

### Reading

1. Chapter 5 of *Regression and Other Stories*: Simulation
2. Appendix A of *Regression and Other Stories*: Computing in R, Sections A.6–A.7

### Pre-class warmup assignments

1. Simulate from probability models

Write R code to do the following:

- (a) Simulate a random number representing the number of shots that a basketball player makes in 10 tries, if the results of the shots are independent and she has a 40% chance of making each shot.
- (b) Simulate a random number representing the number of shots that a basketball player makes in 10 tries, if the results of the shots are independent and her chance of making a shot is 10% for the first shot, 20% for the second shot, 30% for the third, etc.
- (c) Suppose the distance that you throw a ball has a normal distribution with mean 20 meters and standard deviation 5 meters. Simulate the results of three independent throws.

2. Simulate data collection

Write R code to do the following:

- (a) 100 people are asked how many pets they have. The respondents are randomly sampled from a population where 30% of the people have no pets, 40% have one pet, 20% have two pets, and 10% have three pets. Simulate the 100 responses and compute the average number of pets among the respondents.
- (b) 100 students are randomly divided into two groups: 50 take a new experimental math course and 50 take the usual class. They are then all given a test. Suppose that test scores are normally distributed with mean 60 and standard deviation 10 after taking the usual class, or mean 70 and standard deviation 15 after taking the experimental course. Simulate the 100 test scores from the experiment and compute the estimated treatment effect and standard error.

### Homework assignments

1. (a) Inference from a proportion with  $y = 0$  (Exercise 4.7 of *Regression and Other Stories*)

Out of a random sample of 50 Americans, zero report having ever held political office. From this information, give a 95% confidence interval for the proportion of Americans who have ever held political office.

(b) Survey weighting (Exercise 4.10 of *Regression and Other Stories*)

Compare two options for a national opinion survey: (a) a simple random sample of 1000 Americans, or (b) a survey that oversamples Latinos, with 300 randomly sampled Latinos and 700 others randomly sampled from the non-Latino population. One of these options will give more accurate comparisons between Latinos and others; the other will give more accurate estimates for the total population average.

- i. Which option gives more accurate comparisons and which option gives more accurate population estimates?
  - ii. Explain your answer by computing standard errors for the Latino/other comparison and the national average under each design. Assume that the national population is 15% Latino, that the items of interest are yes/no questions with approximately equal proportions of each response, and (unrealistically) that the surveys have no problems with nonresponse.
2. (a) Discrete probability simulation (Exercise 5.1 of *Regression and Other Stories*)
- Suppose that a basketball player has a 60% chance of making a shot, and he keeps taking shots until he misses two in a row. Also assume his shots are independent (so that each shot has 60% probability of success, no matter what happened before).
- i. Write an R function to simulate this process.
  - ii. Put the R function in a loop to simulate the process 1000 times. Use the simulation to estimate the mean and standard deviation of the total number of shots that the player will take, and plot a histogram representing the distribution of this random variable.
  - iii. Using your simulations, make a scatterplot of the number of shots the player will take and the proportion of shots that are successes.

(b) Continuous probability simulation (Exercise 5.2 of *Regression and Other Stories*)

The logarithms of weights (in pounds) of men in the United States are approximately normally distributed with mean 5.13 and standard deviation 0.17; women's log weights are approximately normally distributed with mean 4.96 and standard deviation 0.20. Suppose 10 adults selected at random step on an elevator with a capacity of 1750 pounds. What is the probability that their total weight exceeds this limit?

## Stories

### 1. The proportion of identical twins in the population

In Section 5.1 of *Regression and Other Stories*, we state that the probability of a girl birth is 48.8%. It's clear enough how to estimate this from population data; for example in 2020 there were 3,745,540 babies born in the United States, of whom 48.9% were girls. The percentage varies a bit over time, and we took 48.8% as a reasonable consensus value.<sup>35</sup>

We also say there's a 1/125 chance of fraternal twins and a 1/300 chance of identical twins, with about 49.5% of twins being girls. Where we get those numbers is a bit more interesting. To start with, we're working with old data, before the advent of fertility treatments which have resulted in a near-doubling of the rate of twins in recent decades.<sup>36</sup> In the data we saw, 1.13% birth events (thus, approximately 2.26% of babies) were twins, and 49.5% of those twin babies were girls.

Fine. But how can we estimate the proportion of identical and fraternal twins? We let the students discuss this in pairs for two minutes. Here's a relevant piece of information: back in the day,

<sup>35</sup>T. J. Mathews and Brady Hamilton (2005), Trend analysis of the sex ratio at birth in the United States, *National Vital Statistics Reports* 53 (20), Centers for Disease Control, [https://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53\\_20.pdf](https://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_20.pdf).

<sup>36</sup>U.S. Centers for Disease Control and Prevention (2016), Assisted reproductive technology and multiple births, <https://www.cdc.gov/art/key-findings/multiple-births.html>, and Joyce Martin, Brady Hamilton, Michelle Osterman, and Anne Driscoll (2021), Births: Final data for 2019, *National Vital Statistics Reports* 70 (2), <https://www.cdc.gov/nchs/data/nvsr/nvsr70/nvsr70-02-508.pdf>.

approximately 65% of twins were same-sex pairs and 35% were opposite-sex pairs. Why this imbalance?

We can make use of the fact that identical twins must be of the same sex. On the board, we draw a probability tree, showing the options for a birth—single birth (b or g), identical twin (bb or gg), or fraternal twin (bb, bg, gb, or gg)—and the probabilities for each. Let the relative probabilities of identical and fraternal twins be  $X$  and  $1 - X$ , respectively. Then, if you have twins, the probability of a same-sex pair is  $X + 0.5 * (1 - X)$  and the probability of an opposite-sex pair is  $0.5 * (1 - X)$ . Setting these to 0.65 and 0.35 yields  $X = 1 - 0.35/0.5 = 0.3$ . Multiplying this by a 1.13% rate of twins yields a 0.34% rate of identical twins and a 0.79% rate of fraternal twins. Finally,  $0.0034 = 1/294.1$  and  $0.0079 = 1/126.6$ , which we rounded to 1/300 and 1/125 for convenience.

This example relates to the week's reading because it shows where the data came from in an example in the book. It relates to the course as a whole in demonstrating an example of statistical estimation that is not a simple average, difference, or regression.

## 2. Simulate a process of innovation and improvement

In medicine, there are new drugs and new treatments; in education, improved teaching methods; in business, new techniques to improve sales and market share. In all these arenas, policymakers will have a stream of possible new ideas to be tested and possibly implemented on a large scale. When possible, they will do controlled experiments—called “clinical trials” in medicine or “A/B tests” in industry. In such an experiment, a group of people are randomly assigned either the existing treatment (the “control”) or the new treatment that is being tested. The experiment is done, data are collected and analyzed, and the result is an estimate (with uncertainty) of the efficacy of the treatment compared to the control.

We discuss a simulation we did to estimate the effect, not just of a single intervention, but of a system with a stream of potential interventions. In this story, we won't give the code for our simulation but we'll go through the steps of setting it up.

The scenario is a hypothetical company with a stream of potential interventions coming at the average rate of one per month over a period of two years. Each intervention is tested in a randomized experiment, the result of which is used to decide whether to implement the new idea in production. The goal is to increase profitability, so the outcome being measured is projected total sales minus costs. This could just as well be put in a medical context, where the outcome is survival probability; sports, where the outcome is expected wins; or any other setting where there is a clearly-defined target for improvement.

To perform this simulation it is necessary to have a random process for interventions to come in (similar to the sorts of models used for waiting times at a bus stop, where we know the average rate at which buses arise, but each bus arrives at a random time) and a distribution of effect sizes of the treatments. This distribution will include the possibility of positive or negative effects. If effects were always positive, we would just implement every proposed intervention, and there would be no need to run any experiments. The purpose of gathering experimental data is to estimate the treatment's effect relative to control, and this effect could be either positive or negative.

Next we need to be able to simulate data from the experiments. A single simulation of the process produces a series of times when interventions appear, a true effect size for each intervention, and an estimated effect size and uncertainty from each experiment. Simulating this process will then require us to specify certain “hyperparameters” describing the system, including the average rate of innovations and the mean and standard deviation of the distribution of effect sizes.

Once the process has been simulated, we can look at the effect on total profitability over the two-year period, given different decision rules about when to implement proposed interventions. The simplest decision rule is to do nothing, in which case profitability (which we have defined relative to the control or status quo) would be constant. The next simplest rule is to do all

interventions, in which case profitability goes up or down as new interventions come in. The ideal strategy would be to implement only the interventions that have positive effects; such a strategy is unattainable in practice because the decision maker would observe the estimate, not the true value, of each intervention. In the simulation, we can show what would happen under the unattainable ideal, but we can also show the result under the strategy of implementing all interventions with positive estimates. The result would show some instances of improvement and some instances of decline, but hopefully more pluses than minuses. Finally, the whole simulation can be repeated to get a sense of different possibilities.

The relevance of to the week's reading is that it demonstrates the use of simulation to address an applied problem. It is relevant to the course more generally in being an example where we need to speculate and hypothesize many details about a process in order to simulate it. Even before coding the simulation, a key part of the problem is setting up its assumptions in the first place.

### Class-participation activities

#### 1. Real vs. fake coin flips

Students often have difficulty thinking about summary statistics as random variables with probability distributions. This demonstration, which also alerts students to misconceptions about randomness, motivates the concept of the sampling distribution.<sup>37</sup>

People generally believe that a sequence of coin flips should have a haphazard pattern, including frequent (but not regular) alternations between heads and tails. In fact, it is common for sequences of random coin flips to have long runs of heads and tails.

The demonstration proceeds as follows. The instructor picks two students to be “judges” and one to be the “recorder” and divide the others in the class into two groups. One group is instructed to flip a coin 100 times, or flip 10 coins 10 times each, or follow some similarly defined protocol, and then to record the results, in order, on a sheet of paper, writing heads as “1” and tails as “0” (because “H” and “T” look similar and can be confused when reading them off a sheet of paper). The second group is told to create a sequence of 100 “0’s and “1’s that are intended to *look like* the result of coin flips—but they are to do this without flipping any coins or using any randomization device (or consulting with the other group of students)—and to write this sequence on a sheet of paper. The recorder is then to copy these sequences onto two separate places on the board.

The instructor and the two judges leave the room for five minutes while the two groups of students in the class create their sequences, and then they return and guess which sequence is from actual coin flips and which was made up; see Figure 31 for an example. After the judges guess, the instructor should try guessing. When we do this, we almost always correct, and the students are impressed.

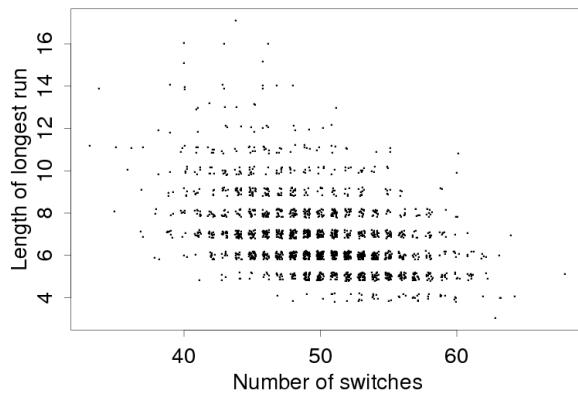
How did we do it? Unless the students are particularly clever, well informed, or lucky, the sequence of fake coin flips will look “random” in an orderly sort of way, with frequent switches between 0’s and 1’s, and the sequence of real coin flips will have a “streaky” look to it, with one or more long runs of successive 0’s or 1’s. One could distinguish between real and fake sequences using a formal rule based on the longest run length, but we find that we can make the distinction more effectively based on a visual inspection of the sequences, which implicitly takes into account much more information.

We can reliably pick out the real sequence using our collective experience and knowledge of coin flips and human guesses. How can this reasoning be formalized? For each of the two sequences on the board, one can count the number of runs (sequences of 0’s and 1’s) and the length of the longest run. If there is interest in formal testing, the instructor can project on the screen or

<sup>37</sup>From Section 8.3.2 of *Teaching Statistics: A Bag of Tricks*.

00111000110010000100	01000101001100010100
00100010001000000001	11101001100011110100
00110010101100001111	01110100011000110111
11001100010101100100	10001001011011011100
100010000001111001	01100100010010000100

**Figure 31** Two binary sequences from a demonstration of real and fake coin flips. After two groups of students produced these data, a pair of students who had not been involved was asked to figure out which of these is an actual sequence of 100 random coin flips and which is fake. The answer is that the sequence on the left is real and the one on the right is fake. The left sequence has a few long sequences of 0's and 1's, which are typical in real coin flips but are rarely produced by students trying to produce realistic fake sequences.



**Figure 32** Length of longest run (sequence of successive heads or successive tails) vs. number of runs (sequences of heads or tails) in each of 2000 independent simulations of 100 coin flips. Each dot on the graph represents a sequence of 100 coin flips; the points are jittered so they do not overlap. When plotted on this graph, the results from an actual sequence of 100 coin flips will most likely fall in an area with a large number of dots. In contrast, a sequence of heads and tails that is artificially created to look “random” will probably have too many switches and no long runs, hence will fall on the lower right of this graph.

hand out copies of Figure 32, which shows the *probability distribution* of these two statistics, as simulated from 2000 independent computer simulations of 100 coin flips. Students should circle on the scatterplot the locations of the values for the sequences on the board. Most of the times we have used this example in class, the sequence of real coin flips is near the center of the scatterplot, and the sequence of fake coin flips has too many switches and too short a longest run, compared to this distribution.

This demonstration illustrates an important point for the interpretation of data: seemingly surprising patterns (long sequences of heads or tails) can occur entirely at random, with no external cause. Long runs in real coin-flip data surprise students because they expect that any part of a random sequence will itself look “random”—that is, typical of the whole. This can motivate a discussion of the general phenomenon that small samples can be unrepresentative of a population. Familiar examples include biological data (for example, a family can have several boys or girls in a row) and sports (a player can have hot and cold streaks that are consistent with random fluctuation).<sup>38</sup>

<sup>38</sup>But see Joshua Miller and Adam Sanjurjo, Surprised by the hot hand fallacy? A truth in the law of small numbers, *Econometrica* 86, 2019–2047, and the discussion in Andrew Gelman (2015), Hey—guess what? There really is a hot hand!, <https://statmodeling.stat.columbia.edu/2015/07/09/hey-guess-what-there-really-is-a-hot-hand/>, and Andrew Gelman (2018), A couple of thoughts regarding the hot hand fallacy fallacy, <https://statmodeling.stat.columbia.edu/2018/12/14/couple-thoughts-regarding-hot-hand-fallacy-argument/>.

The demonstration doesn't always work—sometimes it's hard to tell the two sequences apart—so more recently when doing this activity we have made our odds easier by dividing into four groups, creating two sequences of real coin flips and two of fake coin flips, and having them write both real sequences on one board and both fake sequences on the other. Now it's easier to figure out which sequences are which—all that is necessary is to identify one of the four sequences that is clearly real and one that is clearly fake, and with two of each there are more chances to figure this out.

After surprising students by identifying the real and fake sequences of coin flips, it is useful to develop their intuition as to why real sequences would be expected to have some long runs of heads or tails, as is indicated by Figure 32. We ask the students what is the probability of having six straight heads? They quickly calculate that it is  $1/2$  to the sixth power, or  $1/64$ . A sequence of 100 coin flips includes 95 sequences of length six, and thus one would expect to see one or two runs of six heads, as well as one or two runs of six tails. A sequence of seven heads occurs with probability  $1/128$ , which is certainly a possibility given that there are 94 chances.

This activity relates to the week's readings because the real sequences are a product of random simulation—using coins rather than a computer, but simulation nonetheless. The relevance to the course as a whole is that the simulation shows how simple randomness can yield surprising and counterintuitive patterns.

## 2. Simulating a probability process

For this activity, the class will together come up with an example of simulating a process and its measurement. Students should begin by working in pairs to come up with a real-world example of interest to them. Then the instructor should pick one of the ideas, take some interesting aspect of the problem and discuss what to simulate and how it would be measured, and then go into R and program it. During the programming process, the instructor should discuss and perturb to give a sense of what each line of code is doing. This activity demonstrates the open-ended nature of programming and how coding can be used to explore.

For example, in a recent class a pair of students came up with the problem of studying the percentage of people who own guns in each of the 50 states. We picked this as our example and started by asking what sort of graphs they might use to display the process we will be simulating. One student suggested a plot of gun ownership rate vs. time, which could look like a graph showing 50 time series. To keep things simple, we asked how one might make a scatterplot with 50 dots showing gun ownership in each state for a single year. If gun ownership rate is plotted on the  $y$ -axis, what would we show on the  $x$ -axis? A student suggested Joe Biden's vote share in the 2020 presidential election. From that we made a rough sketch, with  $x$  ranging from 0.3 to 0.7 (the rough range of Biden's share of the two-party vote in the states) and  $y$  ranging from 0.2 to 0.4 (based on the assumption that approximately 30% of American adults own guns, and that this percentage varies a bit from state to state). We then put a bunch of dots on the graph showing a negative correlation with states like Wyoming on the upper left (high gun ownership rate, low Biden vote share) and states like New York on the lower right.

We stated our goal as being able to simulate this pattern and then simulate hypothetical survey data to produce estimated rates of gun ownership.

We started by simulating Biden's vote share in the states. We could just look this up, but for demonstration purposes we'd like to simulate the whole thing. What's a good distribution to use that matches our sketched graph? We entered the following code snippet into the R text window that is projected onto the screen:

```
J <- 50  
biden <- rnorm( , , )
```

We explained that we use  $J$  for the number of states because we are reserving  $n$  for the sample

size of our survey. We then asked the students to spend a minute in pairs figuring out the correct numbers to put in for the three arguments of the `rnorm` function. Reasonable values are:

```
biden <- rnorm(J, 0.5, 0.1)
```

This simulates 50 values from a distribution with mean 0.5, about 68% of the cases falling within one standard deviation (that is, between 0.4 and 0.6), and about 95% falling in the range (0, 3, 0.7), so that seems reasonable.

We next did the same thing for rates of gun ownership:

```
guns <- rnorm(J, 0.3, 0.6)
```

This simulates 50 values centered around 0.5, with 95% of the points within two standard deviations, that is, the range (0.2, 0.3). There's nothing magic about these numbers; we were just trying to come up with values that roughly matched the range in the scatterplot we'd drawn on the board.

We then took a look at the simulated values:

```
plot(biden, guns)
```

The resulting scatterplot doesn't look right at all: there's no downward correlation as in our sketch, and students realized the problem:  $x$  and  $y$  were simulated independently. We moved the discussion along by suggesting that we simulate  $x$  first and then set  $y$  accordingly to get the desired pattern.

We went to the scatterplot on the board and drew a downward-sloping line going through the points (0.4, 0.3) and (0.2, 0.6), roughly capturing the pattern of lower gun-ownership rates among states that showed more support for Biden. We can write this line as  $y = \_\_ + \_\_x$ . We gave the students two minutes in pairs to work out the intercept and slope of this line.

The way to do it is to first figure out the slope: the line drops by 0.1 in  $y$  while  $x$  increases by 0.2, so the slope is  $-0.1/0.2 = -0.5$ . We can then use the fact that the line goes through (0.4, 0.3) and has a slope of  $-0.5$  to write it as  $y = 0.4 - 0.5(x - 0.3)$ , which comes to  $y = 0.4 - 0.5x + 0.15$ , or  $y = 0.55 - 0.5x$ , which we then code and check:

```
guns <- 0.55 - 0.5*biden  
plot(biden, guns)
```

The resulting graph still looks wrong! The 50 points all fall exactly on the line, which doesn't look right. So we add some variation and check the plot again:

```
guns <- 0.55 - 0.5*biden + rnorm(J, 0, 0.03)  
plot(biden, guns)
```

This looks more like it. Again, there's nothing magic about the value 0.03; decrease it and the points will be closer to the line, or increase it to see more spread.

We then continued with the second step, which was to simulate surveys in each of the 50 states. First we need to pick a sample size:

```
n <- 1000
```

Then we simulated a random sample survey in each state:

```
y <- rbinom(J, n, guns)
```

This code is a bit tricky for newcomers to R, as `rbinom` is taking a mixture of scalar and vector arguments. To clarify the output, we printed  $J$ ,  $n$ ,  $\text{guns}$ , and  $y$  on the R console so students could see what was produced. We have simulated for each state the number of Yes responses to the hypothetical survey; to get estimated gun ownership rates we divide by sample size in each state:

```
guns_estimate <- y/n
```

And then we could finally make our plot:

```
plot(biden, guns_estimate)
```

### Computer demonstrations

#### 1. Break R functions

We start with basic simulations:

```
rbinom(5, 10, 0.5)  
rnorm(5, -3, 2)
```

Then we play around with each of these, trying different numbers, and we go through some experimentation, each time asking students to discuss in pairs and guess what will happen before we try it out. What does it take to get the functions to produce errors or meaningless results? Try setting some of the arguments to fractional or zero or negative values. What happens if you set the probability for the binomial distribution to be greater than 1?

#### 2. Simulate 100 coin flips

We first simulate a series of coin flips and display and summarize them. When that is clear, we loop it 1000 times to get a sense of what is possible, what might happen with 100 flips. When we prepared this for our class, we wanted to be able to quickly compute the length of the runs in the simulated coin flips, and it seemed mildly annoying to code this ourselves, so we googled r longest run of coin flips, which led us to a blog from 2009 that presented the R function rle (Run Length Encoding), which does exactly what we wanted.<sup>39</sup> So here goes:

```
n <- 100  
flips <- rbinom(n, 1, 0.5)  
print(flips)  
print(rle(flips))  
print(rle(flips)$lengths)  
longest_run <- max(rle(flips)$lengths)  
print(longest_run)  
  
# Loop it and print  
for (i in 1:20){  
  n <- 100  
  flips <- rbinom(n, 1, 0.5)  
  longest_run <- max(rle(flips)$lengths)  
  print(longest_run)  
}  
  
# Loop it and show distribution  
n_loop <- 1000  
longest_run <- rep(NA, n_loop)  
for (i in 1:n_loop){  
  n <- 100  
  flips <- rbinom(n, 1, 0.5)
```

<sup>39</sup>See Erik Iverson (2009), R function of the day: rle, <https://blogisticreflections.wordpress.com/2009/09/22/r-function-of-the-day-rle/>, and rle, Run length encoding, R Documentation, <https://stat.ethz.ch/R-manual/R-devel/library/base/html/rle.html>.

```
longest_run[i] <- max(rle(flips)$lengths)
}
hist(longest_run)
hist(longest_run,
  breaks=seq(min(longest_run) - 0.5, max(longest_run) + 0.5, 1))
```

This demonstration relates to the earlier class-participation activity, to the week's reading on simulation, and more generally to the practice of understanding random processes by simulating multiple realizations on the computer.

Students can explore the different statistics (numbers of switches, longest streak, etc.), analytically and visually. They can also come up with their own questions and then modify the analysis accordingly.

## Drills

### 1. Random sampling and looping in R

- (a) Write an R function to simulate the outcome of two basketball players shooting  $n_1$  and  $n_2$  baskets, with probability  $p_1$  and  $p_2$  of success. The function should take  $n_1$ ,  $n_2$ ,  $p_1$ , and  $p_2$ , as arguments, simulate the shots, calculate the proportions of shots made for each player, and return the difference in proportions.

*Solution:*

```
shots <- function(n1, n2, p1, p2){
  y1 <- rbinom(1, n1, p1)
  y2 <- rbinom(1, n2, p2)
  y1/n1 - y2/n2
}
```

- (b) Write R code for running the above function with  $p_1 = 0.3$ ,  $p_2 = 0.4$ , and  $n_1 = n_2 = 20$ .  
(c) Write a loop to evaluate this 1000 times, and plot the distribution of results.

### 2. Simulate propagation of uncertainty

- (a) A man applies for  $n$  jobs. For each job he has a  $p_1$  chance of getting an interview. If he is interviewed, he has a  $p_2$  chance of getting an offer. Write an R function to simulate this process and compute the number of offers he gets. The function should take  $n$ ,  $p_1$ , and  $p_2$  as inputs and return a single number.

*Solution:*

```
offers <- function(n, p1, p2){
  interviews <- rbinom(1, n, p1)
  offers <- rbinom(1, interviews, p2)
  offers
}
```

- (b) Write R code for running the above function with  $p_1 = 0.2$ ,  $p_2 = 0.4$ , and  $n = 10$ .  
(c) Write a loop to evaluate this 1000 times and plot the distribution of results.

## Discussion problems

### 1. Simulate a mixed discrete/continuous distribution

Simulate the incomes of a hypothetical set of 100 people where there is a probability of zero income and a lognormal distribution otherwise.

```
# Simulate 100 people
n <- 100
p <- 0.1 # Set probability of zero income
zero_income <- rbinom(n, 1, p) # Generate cases with zero income
income <- ifelse(zero_income == 1, 0,
                 exp(rnorm(n, log(50e3), 0.5))) # Generate income for nonzero cases

# Display results
hist(income, breaks = 40)
print(mean(income==0))

# Simulate probability of getting max income higher than 300K
n_rep <- 1000
max_income <- rep(NA, n_rep)
for (i in 1:n_rep) {
  zero_income <- rbinom(n, 1, p)
  income <- ifelse(zero_income == 1, 0, exp(rnorm(n, log(50e3), 0.5)))
  max_income[i] <- max(income)
}
print(mean(max_income > 300e3))
```

Now play around with the numbers in this simulation and see how the results change.

## 2. Simulate clustering of buses

A famous real-world example of a stochastic process is the clustering of buses along a route. Suppose that buses start out equally spaced in time and then have to stop for passengers. The bus in front will pick up the first set of passengers, allowing the next bus to skip some stops if no new passengers arrive. This random process will, on average, lead to the clumping of buses: that annoying phenomenon whereby you have to wait a long time for a bus, and then two or three arrive together.

How can this process be simulated on the computer? You need some schedule of buses, a sequence of stops, and a process by which passengers arrive at the bus stops at random. The program would track positions of the buses over time, and you can see how long it would take for the buses to start clumping. Setting up a full simulation of this process would be too difficult for 10 minutes in class, but something can be learned from discussing how each of the steps could be programmed, how these steps could be put together in a big loop, and how the program could be evaluated.

## 3.7 Background on regression modeling

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Slope when predicting elections from the economy	Simulate fake data and fit a regression	Play with a simulated regression	Regression to the mean	Examples of regression to the mean
Clinton/Trump vote vs. polls, and predictions	Memory quiz and regression to the mean	Challenges in setting up a simulation	Scatterplots, lines, and regression	Uniform partisan swing

### Reading

Chapter 6 of *Regression and Other Stories*: Background on regression modeling

### Pre-class warmup assignments

#### 1. Fit a regression to fake data

- Write R code to simulate 100 data points from a linear model with intercept 1, slope 2, and residual standard deviation 3, where the predictors are sampled at random uniformly from the range (0, 4). Fit and print a linear regression to these data, and report whether the true values of the intercept and slope fall within 1 standard error of the true parameter values.
- Make a plot with the simulated data and the fitted regression line, including the formula for the fitted line as in Figure 6.1 of *Regression and Other Stories*.

#### 2. Choose reasonable parameter values

For each of the following examples, sketch with pen on paper a rough graph of a plausible scatterplot of  $y$  vs.  $x$  (include axes on your sketch) and then come up with reasonable values of  $a$ ,  $b$ ,  $\sigma$ , where  $a$  is the intercept of the regression of  $y$  on  $x$ ,  $b$  is the slope, and  $\sigma$  is the residual standard deviation.

- $x$  = midterm exam scores (on a 0–50 scale),  $y$  = final exam scores (on a 0–100 scale) for a class
- $x$  = year (from 1900 to 2000),  $y$  = United States population
- $x$  = age in years,  $y$  = vote in recent election (coded as 1 for Republican, -1 for Democrat, 0 for other)
- $x$  = education level (coded as 1 for less than high school, 2 for high school graduate, 3 for some college, 4 for college graduate, 5 for postgraduate degree),  $y$  = vote in recent election (coded as 1 for Republican, -1 for Democrat, 0 for other)

### Homework assignments

#### 1. (a) Binomial distribution (Exercise 5.3 of *Regression and Other Stories*)

A player takes 10 basketball shots, with a 40% probability of making each shot. Assume the outcomes of the shots are independent.

- Write a line of R code to compute the probability that the player makes exactly 3 of the 10 shots.
- Write an R function to simulate the 10 shots. Loop this function 10 000 times and check that your simulated probability of making exactly 3 shots is close to the exact probability computed in (a).

2. (a) Data and fitted regression line (Exercise 6.1 of *Regression and Other Stories*)

A teacher in a class of 50 students gives a midterm exam with possible scores ranging from 0 to 50 and a final exam with possible scores ranging from 0 to 100. A linear regression is fit, yielding the estimate  $y = 30 + 1.2 * x$  with residual standard deviation 10. Sketch (by hand, not using the computer) the regression line, along with hypothetical data that could yield this fit.

(b) Programming fake-data simulation (Exercise 6.2 of *Regression and Other Stories*)

Write an R function to: (i) simulate  $n$  data points from the model,  $y = a + bx + \text{error}$ , with data points  $x$  uniformly sampled from the range (0, 100) and with errors drawn independently from the normal distribution with mean 0 and standard deviation  $\sigma$ ; (ii) fit a linear regression to the simulated data; and (iii) make a scatterplot of the data and fitted regression line. Your function should take as arguments,  $a, b, n, \sigma$ , and it should return the data, print out the fitted regression, and make the plot. Check your function by trying it out on some values of  $a, b, n, \sigma$ .

(c) *In pairs:* Working through your own example (Exercise 5.13 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, construct a probability model that is relevant to your question at hand and use it to simulate some fake data. Graph your simulated data, compare to a graph of real data, and discuss the connections between your model and your larger substantive questions.

## Stories

1. Slope when predicting elections from the economy

Figure 33 is an update of Figure 7.2 from *Regression and Other Stories* including data through 2020, showing a prediction of U.S. presidential election outcomes (incumbent party's percentage of the two-party vote) given a measure of recent economic growth. Unsurprisingly, stronger growth predicts better electoral performance, and here is the fitted model:

	Median	MAD_SD
(Intercept)	46.7	1.4
growth	2.8	0.6

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 3.7 0.7

This was fit to all the elections since 1948. But something has been changing in American politics. There has been an increase in political polarization: Democrats vote for Democratic candidates and Republicans vote for Republican candidates, with not much crossover voting.

The modern era of political polarization has been dated to the post-1990 period, starting with Bill Clinton's first election campaign.<sup>40</sup> We ask students to discuss in pairs what they would expect to see if they were to fit separate regression to the elections before and after 1990. They should sketch in their notebooks and write their guesses of the regression lines.

Figure 34 shows the results. The slope is lower in recent elections, which makes sense in the context of recent political trends. A more polarized electorate will be less sensitive to the economy. Here are the numerical results for the regression fit to elections from 1948 through 1988:

	Median	MAD_SD
(Intercept)	44.8	2.7
growth	3.5	1.0

<sup>40</sup>See Andrew Gelman, David Park, Boris Shor, and Jeronimo Cortina (2009), *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*, Princeton University Press.

### 3.7. BACKGROUND ON REGRESSION MODELING

99

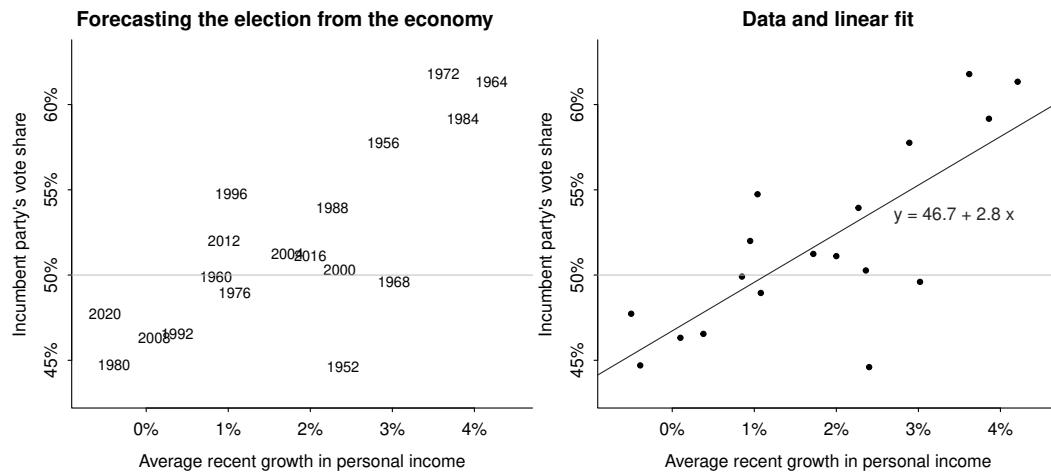


Figure 33 Forecasting U.S. presidential elections from the economy: an update of Figure 7.2 from Regression and Other Stories including data through 2020.

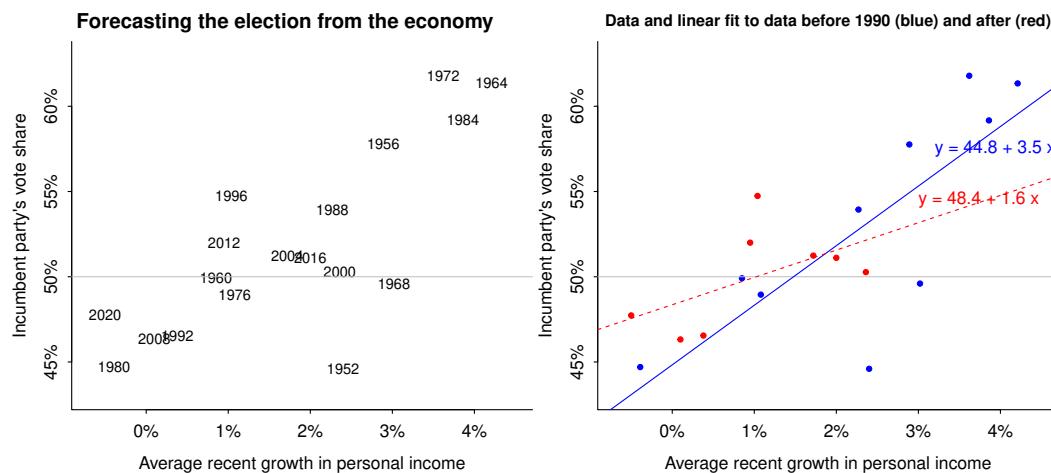


Figure 34 Forecasting U.S. presidential elections from the economy, with separate lines fit to elections before and after 1990. The dotted line, corresponding to more recent elections, has a shallower slope, which makes sense in the context of increasing political polarization.

Auxiliary parameter(s):

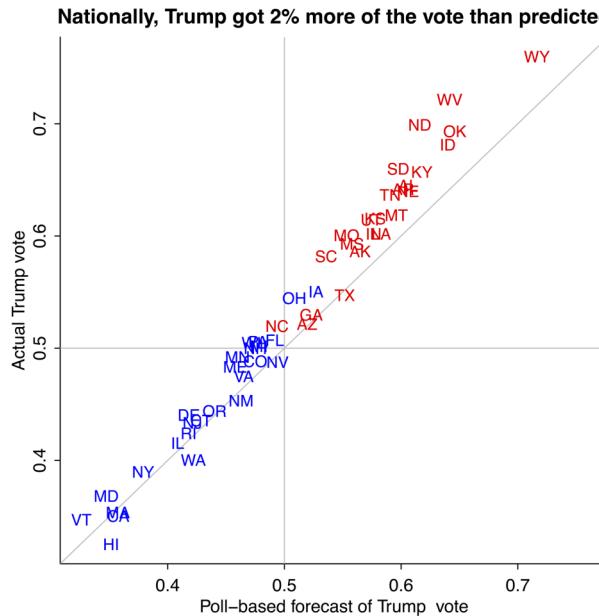
Median MAD\_SD  
 sigma 4.5 1.2

And here's the regression fit to elections from 1992 through 2020:

Median	MAD_SD
(Intercept)	48.4 1.5
growth	1.6 1.1

Auxiliary parameter(s):

Median MAD\_SD  
 sigma 2.8 0.8



**Figure 35** Donald Trump's share of the two-party vote, by state, compared to a polls-based forecast constructed before the election.

This example relates to the week's reading in that it is about understanding the intercept and slope of a linear regression. It relates to the course more generally in connecting these models to social science theory.

## 2. Clinton/Trump vote vs. polls and predictions

The 2016 and 2020 U.S. presidential elections were similar: both times the Democratic candidate was leading in the polls, and both times the Democrat unambiguously won the popular vote, but with a lesser margin than was expected from public polling data. After 2016 we investigated what went wrong and considered more general implications for politics and polling.<sup>41</sup> Here we talk about one part of these analyses: the comparison of the state-by-state outcome in 2016 to poll-based forecasts.

Figure 35 shows the results for the 50 states. We ask students in pairs to summarize the main messages from this graph, which we would take to be, first that the forecasts were highly accurate for most of the states; second that Trump did better most places than forecast; and finally that the discrepancies were larger in more Republican-leaning states.

Figure 36 focuses on the differences, and it shows these patterns more clearly. We ask the students to discuss why this graph is more effective for that purpose.

Then ask about regression lines. If the regression line for the first graph has intercept 0 and slope 1, then what's the regression line for the second graph? If  $y = 0 + 1 * x + \text{error}$ , then we can write  $(y - x) = 0 + 0 * x + \text{error}$ . More generally, if the line is  $y = a + b * x + \text{error}$ , then  $(y - x) = a + (b - 1) * x + \text{error}$ .

What is the regression line for the second graph? Draw a line through the points and see what you get. You also should notice out that the line doesn't fit the data! Not all relationships are linear.

<sup>41</sup> See Andrew Gelman and Julia Azari (2017), 19 things we learned from the 2016 election (with discussion), *Statistics and Public Policy* 4 (1), 1–10.

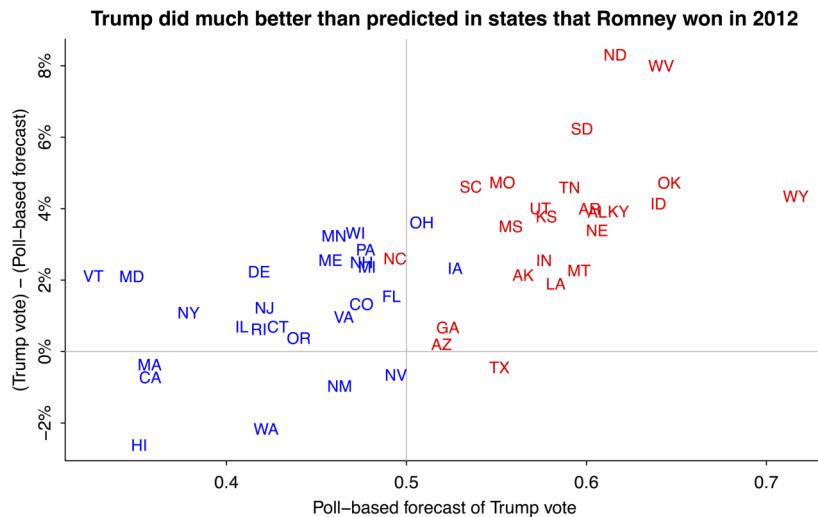


Figure 36 *Trump share of the two-party vote, by state, minus polls-based forecast, plotted vs. the polls-based forecast. Trump outperformed the forecast the most in highly Republican states.*

This example is relevant to the week’s reading because we’re working through a simple regression model. It’s relevant to the class as a whole because it illustrates the general principle of regression being a prediction of  $y$  given  $x$ .

### Class-participation activities

#### 1. Simulate fake data and fit a regression

In discussion, students come up with a “cover story”—a scenario in which before-after data could be approximated by a linear model with variation—and then the instructor leads the class through the process of choosing the parameters (the intercept, slope, and residual standard deviation), simulating data from the model, graphing the data and seeing if they make sense given the cover story, and then altering the parameters and checking again.

#### 2. Before-after memory quiz demonstrating regression to the mean

You can use the computer to simulate 20 random nouns.<sup>42</sup> Here’s what came up when we did this:

brother house bat theory beginner train boy prose run government experience art end song wheel nation jeans baseball reward flock

We tell the students that we will be displaying 20 randomly chosen nouns on the screen for 30 seconds; they should try to memorize as many as they can. After displaying the words for 30 seconds, we discuss other things for three minutes (answering questions on the homework or starting the computer demonstration), and then give the students a minute to individually write down as many words as they remember. We then display the words again and ask them each to count how many they got correct.

We tell the students we will be giving them another memory quiz with 20 new words, and ask them each to predict the score they will get.

We next repeat the exercise with 20 new nouns, again chosen at random; for example,

<sup>42</sup>For example, <https://wordcounter.net/random-word-generator>.

cloth boundary lizard drain hook health wheel wax school car fight lace string  
class wave woman garden army division fold

We then gather three pieces of data from each student: score on first quiz, guessed score on second quiz, actual score on second quiz. Students enter the data on a Google form we have already prepared. We download the data and make a scatterplot of quiz 2 vs. quiz 1 scores and fit the regression line. That line can be used to get a prediction for each student of score 2 given score 1, and students can then compare their guesses with the predicted scores from the regression. The predicted scores typically are closer than the guesses.

This example is relevant to the week's reading as it is an example of "regression to the mean." It is relevant to the course as a whole as an example of the effectiveness of statistical prediction.

### Computer demonstrations

#### 1. Experiment with a simulated regression

Go through the example of Section 6.2 of *Regression and Other Stories*, which gives a chance to discuss and play with the code.

```
library("rstanarm")
x <- 1:20
n <- length(x)
a <- 0.2
b <- 0.3
sigma <- 0.5
y <- a + b*x + sigma*rnorm(n)
fake <- data.frame(x, y)

fit_1 <- stan_glm(y ~ x, data=fake, refresh=0)
print(fit_1, digits=2)

plot(fake$x, fake$y, main="Data and fitted regression line")
a_hat <- coef(fit_1)[1]
b_hat <- coef(fit_1)[2]
abline(a_hat, b_hat)
x_bar <- mean(fake$x)
text(x_bar, a_hat + b_hat*x_bar,
     paste("y =", round(a_hat, 2), "+", round(b_hat, 2), "* x"), adj=0)
```

Some things to do here are to change the sample size, the values of  $x$ , the intercept  $a$  and the slope  $b$ , and the error standard deviation  $\sigma$ . Some things to explore: the variation around the line in the graph, and the discrepancies between the true parameters and their estimate from the regression. Before doing each change (altering the range of  $x$ , increasing the sample size, decreasing the sample size, shifting the intercept, changing the slope, changing the standard deviation of the errors), the instructor can write the change on the board, then ask the students in pairs to write what they expect to see, and then run the code to see what happens. We like this general strategy of having students guess the results of a computation before seeing it appear.

#### 2. Challenges in setting up the regression-to-the-mean simulation

When setting up the fake data example in Section 6.5 of *Regression and Other Stories*, our original plan was to (i) start with random pre-test data, (ii) hypothesize a regression model linking pre-test and post-test results, and (iii) create post-test data using the regression model.

However, it seemed to make more sense to model the underlying process by starting with a latent ability variable, and to generate pre-test as well as post-test data based on it. We do that here.

```
# Setup
library("rstanarm")

# Simulate fake data
n <- 1000
true_ability <- rnorm(n, 50, 10)
noise_1 <- rnorm(n, 0, 10)
noise_2 <- rnorm(n, 0, 10)
midterm <- true_ability + noise_1
final <- true_ability + noise_2
exams <- data.frame(midterm, final)

# Run linear regression
fit_1 <- stan_glm(final ~ midterm, data=exams, refresh=0)
print(fit_1)

# Plot midterm and final exam scores
par(mar=c(3, 3, 2, 1), mgp=c(1.7, .5, 0), tck=-.01)
plot(midterm, final, xlab="Midterm exam score", ylab="Final exam score",
     xlim=c(0,100), ylim=c(0,100), pch=20, cex=.5)
abline(coef(fit_1))
```

This demonstration relates to the week's reading by giving insight into linear regression with a single predictor, and it is relevant to the course as a whole as an example of putting together the steps of simulating a before-after study.

## Drills

### 1. Regression to the mean

- (a) Consider a pre-test, post-test situation where scores on both tests have mean 50, and the regression of post-test on pre-test has a slope of 0.6. A student scores 70 on the pre-test. What is that student's expected score on the post-test?

*Solution:*  $y = 50 + 0.6*(x - 50) + \text{error}$ , so if  $x = 70$ , the predicted value of  $y$  is  $50 + 0.6*20 = 62$ .

- (b) A student scores 80 on the pre-test. What is that student's expected score on the post-test?  
(c) Same, but with a regression slope of 0.1.  
(d) Same, but now the average score is 30 on the pre-test and 65 on the post-test.

### 2. Scatterplots, regression lines, and regression functions

Here are several scatterplots of  $y$  vs.  $x$ . For each, draw the fitted regression line and a curve of  $E(y|x)$ :

- (a) Normal with correlation 0.5 (simulate directly from computer).  
*Solution:* We draw the line on the board.  
(b) Normal with correlation 0.  
(c) Quadratic relationship of  $E(y|x)$  vs.  $x$   
(d) Asymptotic relationship of the form,  $E(y|x) = a(1 + e^{-bx})$ .

## Discussion problems

### 1. Other examples of regression to the mean

Section 6.4 of *Regression and Other Stories* discusses the classic example of parents' and children's heights. When using mother's height to predict daughter's height, the best predictor is to "regress"

toward the mean. Section 6.5 of the of *Regression and Other Stories* shares two other examples: midterm and final exams, and performance in successive maneuvers in a flight school. Can you come up with another example in which “regression to the mean” arises?

2. Understanding uniform partisan swing (considering regression to the mean)

National elections approximately follow uniform partisan swing at the national and local levels, typically with only small changes from year to year. But over a 20-year period there can be big changes. How can these patterns in the United States and elsewhere be understood? Is the concept of regression to the mean relevant here?

## 3.8 Linear regression with a single predictor

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
$5^2 + 12^2 = 13^2$	African countries in the United Nations	Regression, transformations, and sample size	Sketch a fitted model	Predict elections given incumbency
Regression of earnings on height	Socioeconomic status and political ideology	Take average or regress on a constant term	Probabilities from regression	How large was the sample size?

### Reading

Chapter 7 of *Regression and Other Stories*: Linear regression with a single predictor

### Pre-class warmup assignments

1. Check regression using fake-data simulation
  - (a) Write an R function that does the following: (i) simulate  $n$  data points from a linear model with intercept  $a$ , slope  $b$ , and residual standard deviation  $\sigma$ , where the predictors are sampled at random uniformly from the range  $(x_{lo}, x_{hi})$ ; (ii) fit a linear regression to these data; and (iii) check to see if the estimated slope is within 1 standard error of its true value.
  - (b) Check your function by running it with the values  $n = 50$ ,  $a = 10$ ,  $b = -20$ ,  $\sigma = 30$ .
  - (c) Put your function in a loop, run it 100 times, and make a plot with the simulated data and the fitted regression line, including the formula for the fitted line as in Figure 6.1 of *Regression and Other Stories*.
2. Formulate comparisons as regression models  
Suppose you have a vector `test_scores` of exam grades for a class of students, and a vector `level` taking on the values 1 for undergraduate and 2 for graduate students.
  - (a) Explain why the estimated coefficient in the linear regression `test_scores ~ 1` gives the average grade in the class. Give R code to perform this regression and extract the coefficient estimate.
  - (b) Write the code for the linear regression that gives the difference between the average grades for graduate and undergraduates in the class.

### Homework assignments

1. (a) Simulating regression with measurement error
  - i. Simulate 100 data points with measurements  $x$  and  $y$  as follows. Let  $x$  be uniformly distributed between 0 and 10, and simulate  $y$  from the model,  $y = a + bx + \text{error}$ , where  $a = 2$ ,  $b = 0.5$ , and the error is normally distributed with mean 0 and standard deviation  $\sigma = 2$ . Give your code.
  - ii. Use `stan_glm` to fit a linear regression of  $y$  on  $x$ . Display the result. Discuss the discrepancies between your estimates of  $a$ ,  $b$ ,  $\sigma$  and the true values, in comparison to the standard errors from the regression. Give your code as well as your R output.
  - iii. Plot the data and the fitted regression line. Give your code as well as your graph.
  - iv. Now suppose  $x$  is measured with error: Create a new variable `x_obs` by taking  $x$  and adding normally distributed errors with mean 0 and standard deviation 5. Now fit the

linear regression predicting  $y$  from  $x_{\text{obs}}$ . How do the coefficient estimates compare to the result from regressing  $y$  on  $x$  as you did earlier?

2. (a) Regression predictors (Exercise 7.1 of *Regression and Other Stories*)

In the election forecasting example of Section 7.1 of *Regression and Other Stories*, we used inflation-adjusted growth in average personal income as a predictor. From the standpoint of economics, it makes sense to adjust for inflation here. But suppose the model had used growth in average personal income, not adjusting for inflation. How would this have changed the resulting regression? How would this change have affected the fit and interpretation of the results?

(b) Fake-data simulation and regression (Exercise 7.2 of *Regression and Other Stories*)

Simulate 100 data points from the linear model,  $y = a + bx + \text{error}$ , with  $a = 5$ ,  $b = 7$ , the values of  $x$  being sampled at random from a uniform distribution on the range  $[0, 50]$ , and errors that are normally distributed with mean 0 and standard deviation 3.

- i. Fit a regression line to these data and display the output.
- ii. Graph a scatterplot of the data and the regression line.
- iii. Use the `text` function in R to add the formula of the fitted line to the graph.

## Stories

1.  $5^2 + 12^2 = 13^2$

Many areas of statistics involve adding components of variation. The mathematical rule is that standard deviations are added on the squared scale; that's why we have formulas such as  $\text{sd}(u + v) = \sqrt{\text{sd}(u)^2 + \text{sd}(v)^2}$  for independent random variables  $u$  and  $v$ . We don't attempt to explain or derive this Pythagorean formula in class, but we can give several clean examples:

- A survey is done in 2021 to estimate opinion on some issue, and a new group of people is surveyed in 2022 in order to estimate the change in opinion on some issue. If the sample sizes of the surveys are  $n_1$  and  $n_2$ , then the standard error of the difference is  $\sqrt{0.5^2/n_1 + 0.5^2/n_2}$ . Here we are making the usual assumption that the proportion  $p$  is close to 0.5 so that  $p(1-p)$  is close to 0.25; see pages 51–52 of *Regression and Other Stories*.
- A survey is done of  $n_1$  women and  $n_2$  men, with the goal being to estimate the gender gap: the difference in support for some candidate comparing the two sexes. The standard error of the difference is  $\sqrt{0.5^2/n_1 + 0.5^2/n_2}$ .
- An A/B test: an experiment is performed in which  $n_1$  people get the treatment and  $n_2$  get the control, and some outcome is measured (for example, click-through rate or sales or time spent on a website). The experimenters then compute the average difference between treated and control groups; the standard error of this difference is  $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ .

All these examples involve two errors that are squared, added, and square-rooted. This mathematical operation has a diminishing-returns feature that you can see numerically and then with an applied example.

First the numerical demonstrations.

Start with two independent errors, each with standard deviation 10. Add these up and you get  $\sqrt{10^2 + 10^2} = 14.1$ . What if you start with a standard deviation of 10 and add a standard deviation of 5? You get  $\sqrt{10^2 + 5^2} = 11.2$ . This is barely more than the original value of 10. In short: when adding two numbers in this way, the larger value dominates.

Here's another example. We write this expression on the board:

$$\sqrt{50^2 + 10^2}$$

and ask students in pairs to guess, without direct calculations, what it will be. We ask some for their guesses and then write the answer on the board. It's 51, which is lower than typical guesses. Once you have an error with standard deviation 50, adding a new error term with standard deviation 10 is pretty much irrelevant.

To put it another way, suppose you start with these two error terms and you have the opportunity to decrease one of them. You start with  $\sqrt{50^2 + 10^2}$ , and you can either decrease the first term by one-fifth (so you'll have  $\sqrt{40^2 + 10^2}$ ) or decrease the second term by one-fifth (to get  $\sqrt{50^2 + 8^2}$ ). Which is better? The answer is that it's better—much better—to reduce the larger error:  $\sqrt{40^2 + 10^2} = 41.2$ , whereas  $\sqrt{50^2 + 8^2} = 50.6$ . You get much more bang for your buck by reducing the larger error term than by reducing the smaller one.

This is a principle of diminishing returns: if you have multiple sources of error, you should work to reduce the largest term.<sup>43</sup> As a mnemonic, we summarize this as  $5^2 + 12^2 = 13^2$ .

Here is an example from Section 4.3 of *Regression and Other Stories*. Consider a survey with  $n = 600$  respondents and thus a sampling error of  $\pm 0.5/\sqrt{600} = 0.02$ , or 2 percentage points. Now suppose you were to redo the survey with 60 000 respondents. What would the standard error be? With 100 times the sample size, the standard error will be divided by 10, thus the estimated proportion will have a standard error of 0.002, or  $\pm 0.2$  percentage points.

But imagine a poll were reported in this way: the support for some issue is  $52.5\% \pm 0.2\%$ . This would not make sense. Why not?

The problem here is that 0.002 is the scale of the sampling error, but there are many other sources of error arising from nonresponse, inaccuracy in predicting who will vote, variation in opinion over time, and inaccurate survey responses.

One way to account for these sources of uncertainty is to add an error term representing nonsampling error. For example, in state polling for U.S. elections, nonsampling error has been estimated to have a standard error of approximately 2.5 percentage points. So if we want to account for total uncertainty in our survey of 600 people, we would use a standard error of  $\sqrt{2^2 + 2.5^2} = 3.2$  percentage points, and for the survey of 60 000 people, the standard error would be  $\sqrt{0.2^2 + 2.5^2} = 2.51$  percentage points. This formula shows that not much would be gained by increasing the sample size to 60 000 because it only serves to decrease the first term inside the square root; it doesn't effect the second term at all.

This story is relevant to the week's reading because it helps us put uncertainties and standard errors in context. It connects to the course more generally by addressing a limitation of the usual statistical summaries.

## 2. Interpret the regression of earnings on height

Taller people make more money, on average. We fit this model to survey data from 1990:

```
fit0 <- stan_glm(earn ~ height, data=earnings, refresh=0)
print(fit0)
```

And here's what we saw:

```
stan_glm
family: gaussian [identity]
formula: earn ~ height
observations: 1816
predictors: 2
-----
```

<sup>43</sup> Andrew Gelman (2008), The most important formula in statistics, [https://statmodeling.stat.columbia.edu/2008/10/25/the\\_most\\_import/](https://statmodeling.stat.columbia.edu/2008/10/25/the_most_import/).

```
Median MAD_SD
(Intercept) -85000  9000
height       1600   100
```

```
Auxiliary parameter(s):
Median MAD_SD
sigma 22000   400
```

The instructor can display this output on the screen and ask the students in pairs to go through each of the numbers in the output. How to understand the result? One thing we know is that men make more money than women, on average. So maybe the positive coefficient for height is just coming from the comparison of women to men.

We tried fitting the model just to men:

```
fit1 <- stan_glm(earn ~ height, data=earnings, subset=(male==1), refresh=0)
print(fit1)
```

And here's what came out:

```
stan_glm
family:      gaussian [identity]
formula:     earn ~ height
observations: 675
predictors:  2
-----
Median MAD_SD
(Intercept) -39000  26000
height       1000   400

Auxiliary parameter(s):
Median MAD_SD
sigma 29000   800
```

How to understand this? Students in pairs should sketch, with pen on paper, the fitted regression line with  $x$  in the range (60, 80)—that's from 5 feet to 6 feet 8 inches tall. They should then sketch dotted lines at  $\pm \sigma$  above and below the line. This shows that the predictive value of height is small compared to the variation in the data.

The instructor can then ask the students to consider some possible explanations for the height-and-earnings pattern and to consider what data they could gather to explore their hypotheses. Some theories in the literature include the psychological advantage of being taller as a child, the advantage of being taller as a young adult, and height being correlated with social class, intelligence, and other attributes that are predictive of earnings.<sup>44</sup>

This story is relevant to the week's reading because it is about the interpretation of a regression with one predictor (or two, if you count the constant term). It relates to the course more generally because the first step in understanding any regression is to understand one coefficient at a time.

### Class-participation activities

#### 1. African countries in the United Nations

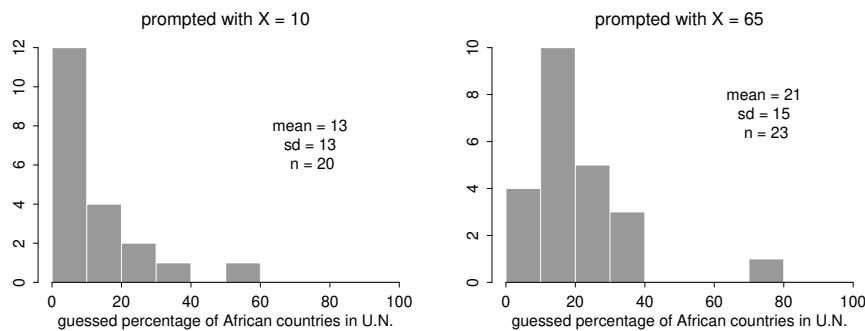
Ahead of time the instructor should prepare survey forms, one for each student in the class, as shown in Figure 37. When it is time in class to perform the activity, the instructor should pass out

<sup>44</sup>There's a large literature on this topic; see for example Andreas Schick and Richard Steckel (2015), Height, human capital, and earnings: The contributions of cognitive and noncognitive ability, *Journal of Human Capital* 9, 94–115.

We assigned a number to you between 0 and 100. It is  $X = \underline{\hspace{2cm}}$ .

1. Do you think the *percentage* of countries, among all those in the United Nations, that are in Africa is **higher** or **lower** than  $X$ ?
2. Give your best estimate of the *percentage* of countries, among all those in the United Nations, that are in Africa.

**Figure 37** Handout for the United Nations activity. The instructor should make copies of this form for all students in the class and write, by hand, “10” in the blank space for half the forms and “65” on the others. The instructor should then fold each of the forms, shuffle them, and hand them out to the students, telling them to answer the survey questions individually, without discussing with their neighbors.



**Figure 38** Responses of students in a small introductory probability and statistics class to the question, “Give your best estimate of the percentage of countries, among all those in the United Nations, that are in Africa.” The students were previously asked to compare this percentage with a specified value  $X$ ; histogram (a) displays responses for students given  $X = 10$ , and histogram (b) displays responses for students given  $X = 65$ . The students were told that the value of  $X$  was chosen at random, and yet it has an effect (on average) on their responses. Similar results were obtained when the experiment was repeated in other classes.

the folded forms, ask the students to answer the two questions independently (without discussing it with their neighbors), and then fold the forms back up and return them. Because each student must separately fill out a survey form, this is one of the few demonstrations in the course in which the students do *not* work in pairs or groups.

After the students have answered the questions and returned the forms, the instructor should explain that only two values of  $X$  were actually assigned and that the point of the activity is to find the relation between  $X$  and the students’ responses on the second question. In statistical terminology, this is an experiment in which the units are the students, the treatments are the hand-written values of  $X$  (10 or 65), and the outcome of interest is the response to the second question.

This activity is adapted from a published study that reported the median responses to the second question as 25 or 45, given  $X = 10$  or 65, respectively. This result has been characterized as an example of the “anchoring heuristic,” in which an estimate of an unknown quantity is influenced by a previously supplied starting point. In this example, the value of  $X$  should not affect the outcome (after all, the students were told that  $X$  was randomly generated), yet it does!<sup>45</sup>

<sup>45</sup>Amos Tversky and Daniel Kahneman (1974), Judgment under uncertainty: Heuristics and biases, *Science* 185, 1124–1131. The class-participation activity is taken from Section 6.4.1 of *Teaching Statistics: A Bag of Tricks*.

Now is a good time to discuss the principles of randomization and blindness in experimentation, now that the students have been subjects in an experiment. Incidentally, the actual value of the unknown quantity is irrelevant for this example—we are only studying the differences in responses between the two groups of students.

In performing this example in our classes, we have replicated the anchoring effect, although its magnitude has not been so dramatic as in the published literature; for example, histograms of responses for a class of 43 students are shown in Figure 38. When the data have been collected, the two groups can immediately be compared graphically using histograms; typically, as in Figure 38, the histograms overlap considerably but clearly differ.

Other examples in which an experiment is embedded in a survey include studies of question wording, question ordering, and other causes of survey response bias. In addition, the experiment given here could be made more complex in various ways, for example by randomly ordering the options “higher” and “lower” in the first question of the survey, thus giving two factors and four possible experimental conditions.

This activity is relevant to the week’s readings in that the comparison between the two groups can be framed as a regression on an indicator—a variable that equals 0 or 1 if the anchor was set to 10 or 65, respectively. If you label that indicator variable as  $z$ , then the difference between the two groups is the slope of the regression of  $y$  on  $z$ . The activity is relevant to the course as a whole because it is an example of a controlled experiment, a form of data collection we often see in social and medical sciences.

## 2. Socioeconomic status and political ideology, studied using a class-designed questionnaire

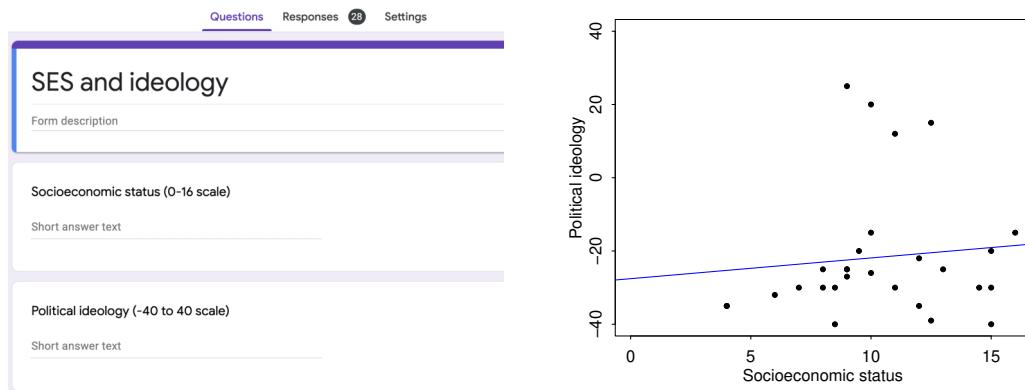
Various socioeconomic variables are known to be predictive of political attitudes.<sup>46</sup> The instructor can ask students to list some of these variables (for example, race/ethnicity, sex, age, income, education, and marital status). For this activity, students should consider a variable that would predict political ideology in this class. First we need an outcome measure. For a class at Columbia University, simple party identification would not have worked well because there are so few Republicans at the university, so instead we used a left-right ideology score on a scale from  $-5$  (far left) to  $+5$  (far right). Students then need to come up with a good socioeconomic predictor. A challenge for our class was to come up with a measure that would work for students coming from many different countries.

The instructor can set up a Google questionnaire (see Figure 39a for an example), write the URL on the board, and have students fill out the form on their phones or computers. Save the data as a .csv file, read it into R, and fit a regression. Before they see the result, students should work in pairs to guess what the data and regression output will look like. They can then be shown the regression fit, along with a scatterplot and fitted line, and have students discuss how this differed from their expectations. Figure 39b illustrates.

Next the class can consider how this could be done better. What would be a stronger predictor of the outcome? How could you know it was stronger? Do you compare the intercepts? The slopes? The residual standard deviation?  $R^2$ ?

This activity is relevant to the week’s readings because it is about the interpretation and understanding of a linear regression. This time, instead of starting with the data and moving to the interpretation, students starting with the problem and then guess what might happen in the data. This is how things go when you are designing a study rather than analyzing existing data. The activity relates to the course as a whole as it is an example of using a survey to address a social science question.

<sup>46</sup>See, for example, Andrew Gelman, David Park, Boris Shor, and Jeronimo Cortina (2009), *Red State, Blue State, Rich State, Poor State*, Princeton University Press.



**Figure 39** (a) A Google form we created in class for students to enter data on socioeconomic status (SES) and sociology. The SES measure was created by adding four items, each on a 0–4 scale, representing private education, parents’ education, parents’ income, and size of city where the student grew up. The political ideology measure was anchored by the positions of various politicians (Bernie Sanders at −30, Hillary Clinton at −10, Mitt Romney at +10, and Ted Cruz at +30), with responses restricted to fall between −40 and +40. (b) Data and fitted regression line. Before displaying this graph, we asked students to guess the intercept, slope, and residual standard deviation of the regression line. For the data shown, it is  $y = -28 + 0.6x$  with residual standard deviation 18.

### Computer demonstrations

#### 1. Regression, linear transformation, and sample size

Here you can work out some basic operations on a fitted regression, using the example of height and earnings for men, as before using data from the 1990 Work, Family, and Well-being Survey. To get to the data, start with the webpage for *Regression and Other Stories*,<sup>47</sup> then click on Examples, Examples alphabetically, Earnings, data, earnings.csv, Raw, and then copy and paste the URL.<sup>48</sup>

```
library("rstanarm")
earnings <- read.csv(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Earnings/data/",
  "earnings.csv"
))
head(earnings)

fit0 <- stan_glm(earn ~ height, data=earnings, refresh=0)
print(fit0)

earnings$earnk <- earnings$earn/1000
fit1 <- stan_glm(earnk ~ height, data=earnings, refresh=0)
```

Before showing the result, we ask students in pairs to fill in the regression output. They should be able to do this. We then continue:

```
print(fit1)
par(mar=c(3,3,1,1), mgp=c(1.5,.5,0), tck=-.01)
plot(earnings$height, earnings$earnk,
  pch=20, cex=.5)
```

<sup>47</sup><http://www.stat.columbia.edu/~gelman/regression/>.

<sup>48</sup><https://raw.githubusercontent.com/avehtari/ROS-Examples/master/Earnings/data/earnings.csv>.

```
abline(coef(fit1))

R2 <- 1 - sigma(fit1)^2 / sd(earnings$earnk)^2
print(R2)
```

What would the results look like with a different sample size? Repeat the above steps on a random sample of 1/4 of the data points:

```
n <- nrow(earnings)
earnings$select <- sample(c(0,1), n, replace=TRUE, prob=c(0.75, 0.25))
fit2 <- stan_glm(earnk ~ height, data=earnings, refresh=0,
subset=(select==1))
print(fit2)
```

The sample size is divided by 4, so the standard errors should be about twice as large.

2. Take the average or regress on a constant term

As discussed in Section 7.3 of *Regression and Other Stories*, performing a regression with just the constant term is equivalent to estimating a mean. We demonstrate here with a simulation.

```
library("rstanarm")

# Create fake data
n <- 30
y <- rnorm(n, 40, 20)

# Estimate mean by hand
y_bar <- mean(y)
se <- sd(y) / sqrt(length(y))
print(c(y_bar, se))

# Estimate mean using regression on constant term
data <- data.frame(y)
fit <- stan_glm(y ~ 1, data=data, refresh=0)
print(fit)
```

## Drills

1. Sketch a fitted regression model

For each example, sketch the regression line and data that match both the fitted model and the residual standard deviation.

- (a)  $y = 2.5 + 0.4x + \text{error}$ , and  $\sigma = 3$  (from  $x = 0$  to  $10$ )

*Solution:* Draw the axes and the range of  $x$ , then figure out the range of  $y$  (from  $2.5 + 0.4*0$  to  $2.5 + 0.4*10$ , that is, from  $2.5$  to  $6.5$ , so draw a  $y$ -axis roughly from  $0$  to  $10$ ), then draw the line, then draw dotted lines parallel to the regression line at  $\pm\sigma$ , then sketch a bunch of points, with about two-thirds within the dotted lines and one-third outside.

- (b)  $y = -150 + 20x + \text{error}$ , and  $\sigma = 100$  (from  $x = 10$  to  $20$ )

- (c)  $y = 80 + 3 * (x - 1900) + \text{error}$ , and  $\sigma = 10$  (from  $x = 1900$  to  $2000$ )

2. Predict probabilities using regression

The following regression output describes the relationship of midterm and final exam scores, based on fake data (compare Section 6.5 of *Regression and Other Stories*).

Calculate the predicted final score given the stated midterm score; and calculate the probability that the student will score in the range specified.

```
Median  MAD_SD
(Intercept) 24.8    1.4
midterm      0.5    0.1
```

Auxiliary parameter(s):

```
Median  MAD_SD
sigma   11.6    0.3
```

- (a) Midterm score 60; probability to score higher than 50?

*Solution:* Prediction is  $24.8 + 0.5 * 60 = 54.8$ , and errors have standard deviation 11.6, so probability of scoring higher than 50 is  $1 - \text{pnorm}(50, 54.8, 11.6)$ , which comes to 0.66, or 66%. Or, equivalently,  $1 - \text{pnorm}((50 - 54.8)/11.6)$ . We also graph the normal distribution centered at 54.6 and with standard deviation 11.6, then draw a vertical line at 50 and shade the area on the right.

- (b) Midterm score 100; probability to score higher than 50?  
(c) Midterm score 60; probability to score lower than 60?  
(d) Midterm score 100; probability to score lower than 90?  
(e) Midterm score 50; probability to score between 40 and 60?

### Discussion problems

1. Regression line and modeling for future elections given incumbency

The biggest U.S. presidential electoral landslides in modern memory (see Figure 33 in this book or Figure 7.2 in *Regression and Other Stories*) came in 1964, 1972, and 1984, years when incumbents were running for reelection. More generally, we might expect the model predicting election outcomes from the economy to be different when incumbents are running for reelection, compared to when new candidates are running. How might this change the regression line? How might you model this?

2. How large was the sample size?

Consider this regression predicting final exam scores on midterm exam scores:

```
Median  MAD_SD
(Intercept) 24.8    1.4
midterm      0.5    0.1
```

Auxiliary parameter(s):

```
Median  MAD_SD
sigma   11.6    0.3
```

In the output, we didn't include  $n$ . How can you approximately figure out  $n$  here? You should be able to attack this problem by simulating data and a regression in R and playing around with the inputs until you get something that looks roughly like the result above.

## 3.9 Least squares and fitting regression models

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Ronald Reagan and the evangelical vote	Simulate and recover regression lines	Play with the regression errors	Sample size and standard estimate	From inference to decision
Does having a girl make you more conservative?	Move a point and shift the regression line	Compare <code>lm</code> and <code>stan_glm</code>	Averages and comparisons as regressions	Sample size and statistical significance

### Reading

Chapter 8 of *Regression and Other Stories*: Fitting regression models

### Pre-class warmup assignments

1. Linear regression fit to three data points
  - (a) Without using the computer, figure out the slope,  $b$ , of the least-squares line,  $y = a + bx$ , for the three points  $(x, y) = (0, 0), (1, 2), (2, 2)$ .
  - (b) Without using the computer, figure out the intercept,  $a$ , of this fitted line.
  - (c) Suppose you fit a linear regression to world population during the twentieth century, using the following data points: 1.5 billion in 1900, 2.5 billion in 1950, 6 billion in 2000. Without using the computer, figure out the least-squares line,  $y = a + bx$ , for these three data points.
2. Compare `lm` and `stan_glm`
  - (a) Simulate 100 data points from a linear regression model with two predictors (that is,  $y = b_0 + b_1x_1 + b_2x_2 + \text{error}$ ). Fit the model to the simulated data using `lm` and then fit it using `stan_glm` with its default settings. Compare the results and discuss how they differ.
  - (b) Repeat, but just simulating 10 data points instead of 100.

### Homework assignments

1. (a) Fake-data simulation and fitting the wrong model (Exercise 7.3 of *Regression and Other Stories*)  
Simulate 100 data points from the model,  $y = a + bx + cx^2 + \text{error}$ , with the values of  $x$  being sampled at random from a uniform distribution on the range  $[0, 50]$ , errors that are normally distributed with mean 0 and standard deviation 3, and  $a, b, c$  chosen so that a scatterplot of the data shows a clear nonlinear curve.
  - i. Fit a regression line `stan_glm(y ~ x)` to these data and display the output.
  - ii. Graph a scatterplot of the data and the regression line. This is the best-fit linear regression. What does “best-fit” mean in this context?
- (b) Formulating comparisons as regression models (Exercise 7.6 of *Regression and Other Stories*)  
Take the election forecasting model and simplify it by creating a binary predictor defined as  $x = 0$  if income growth is less than 2% and  $x = 1$  if income growth is more than 2%.
  - i. Compute the difference in incumbent party’s vote share on average, comparing those two groups of elections, and determine the standard error for this difference.
  - ii. Regress incumbent party’s vote share on the binary predictor of income growth and check that the estimate and standard error are the same as your earlier result.

2. (a) Least squares (Exercise 8.1 of *Regression and Other Stories*)

The folder `ElectionsEconomy` in the repository at <https://avehtari.github.io/ROS-Examples/examples.html> contains the data for the example in Section 7.1 of *Regression and Other Stories*. Read in these data, type in the R function `rss()` from page 104 of the book, and evaluate it at several different values of  $(a, b)$ . Make two graphs: a plot of the sum of squares of residuals as a function of  $a$ , with  $b$  fixed at its least squares estimate given in Section 7.1 of the book, and a plot of the sum of squares of residuals as a function of  $b$ , with  $a$  fixed at its least squares estimate. Confirm that the residual sum of squares is indeed minimized at the least squares estimate.

(b) *In pairs:* Working through your own example (Exercise 7.10 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, fit a linear regression with a single predictor, graph the data along with the fitted line, and interpret the estimated parameters and their uncertainties.

## Stories

1. No, Ronald Reagan did not win “overwhelming support from evangelicals”

We were reading an article in the *New Yorker* magazine and noticed the following claim that made us suspicious.<sup>49</sup>

“[In 1980], Ronald Reagan won the Presidency, with overwhelming support from evangelicals. The evangelical vote has been a serious consideration in every election since.”

Figure 40 shows the support of Republican presidential candidates across different religious groups.<sup>50</sup> According to the National Election Study, Reagan did a bit *worse* among evangelical Protestants than among voters as a whole—no surprise, really, given that Reagan was not particularly religious and his opponent, Jimmy Carter, was an evangelical himself.

It was 1992, not 1980, when evangelicals really started to vote Republican.

Numbers are important, and we worry about misconceptions of American politics—for example, the idea that Reagan won “overwhelming support from evangelicals.” A big motivation of our research in political science is to show people how all sorts of things they thought they knew about politics were actually false.

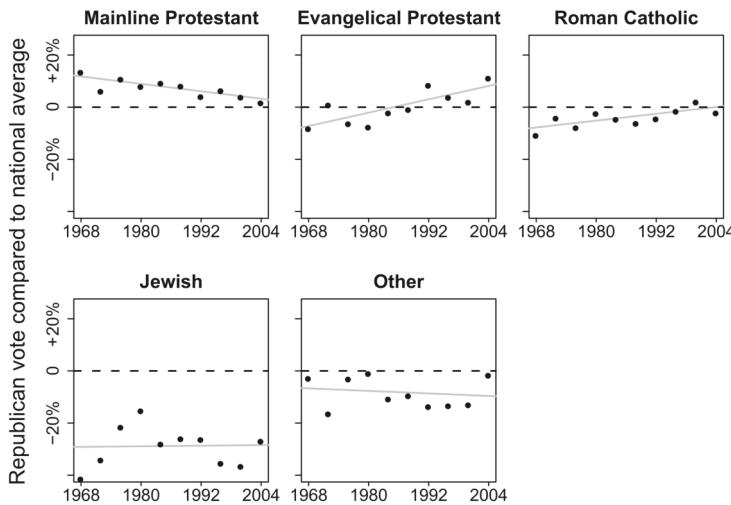
Perhaps the *New Yorker* and other similar publications should hire a statistical fact checker or copy editor? Maybe this is the worst time to suggest such a thing, with the collapsing economics of journalism and all that. Still, we think the magazine could hire someone at a reasonable rate who could fact check their articles. This would free up their writers to focus on the storytelling that they are good at without having to worry about getting the numbers wrong.

Reagan did do well among *white* evangelicals, though.

To connect to the week’s reading on least squares regression: Figure 40 shows how regression can be useful for purely descriptive purposes. In this case, the expression  $y = a + bx + \text{error}$  (where  $x$  is year and  $y$  is Republican vote share relative to the national average, with a separate model fit to each denomination) does not represent any sort of model of the world—it’s not like we’re saying that the gods of elections compute  $a + bx$  and then add an error to determine the election outcome—but the fitted lines are convenient summaries.

<sup>49</sup>Ariel Levy (2010), *Prodigal son: Is the wayward Republican Mike Huckabee now his party’s best hope?*, *New Yorker*, <https://www.newyorker.com/magazine/2010/06/28/prodigal-son/>. The discussion here is taken from Andrew Gelman (2010), Statistical fact checking needed, or, No, Ronald Reagan did not win “overwhelming support from evangelicals,” [https://statmodeling.stat.columbia.edu/2010/07/12/statistical\\_fac/](https://statmodeling.stat.columbia.edu/2010/07/12/statistical_fac/).

<sup>50</sup>Figure 6.9 from Andrew Gelman, David Park, Boris Shor, and Jeronimo Cortina (2009), *Red State, Blue State, Rich State, Poor State*.



**Figure 40** Time series plots of Republican vote share within various religious denominations. In each year, we have subtracted the national vote so that the graph shows each denomination compared to the national average. Points that are above the zero line correspond to groups that are more Republican than the country as a whole. Over these decades, mainline Protestants became less Republican, evangelical Protestants moved from the Democrats to the Republicans, and Catholics lost their Democratic affiliation.

To connect with some of the larger themes in the class, this example shows the value of looking at data, which then allows us to refine our understanding of the world.

## 2. Does having a girl make you more conservative or more liberal?

In 2010 a study was released involving children and political orientation.<sup>51</sup> From a study in the United States, sociologists Dalton Conley and Emily Rauscher wrote:<sup>52</sup>

“Using nationally-representative data from the [1994] General Social Survey, we find that female offspring induce *more conservative political identification*. We hypothesize that this results from the change in reproductive fitness strategy that daughters may evince.”

This surprised us, because an earlier study by economists Andrew Oswald and Nattavudh Powdthavee had come out with the exact opposite finding:<sup>53</sup>

“We document evidence that having daughters leads people to be *more sympathetic to left-wing parties*. Giving birth to sons, by contrast, seems to make people more likely to vote for a right-wing party. Our data, which are primarily from Great Britain, are longitudinal. We also report corroborative results for a German panel.”

This is a fun problem: we have two different studies, both by reputable researchers, with opposite

<sup>51</sup>The discussion here is taken from Andrew Gelman (2010), Having daughters makes you more liberal. No, it makes you more conservative. No, it . . . ?, [https://statmodeling.stat.columbia.edu/2010/04/06/having\\_daughter\\_1/](https://statmodeling.stat.columbia.edu/2010/04/06/having_daughter_1/), Andrew Gelman (2013), Don’t doubt that, man! Please give this fallacy a name, <https://statmodeling.stat.columbia.edu/2013/12/20/discovered-new-fallacy-now-need-name/>, Andrew Gelman (2014), Into the thicket of variation: More on the political orientations of parents of sons and daughters, and a return to the tradeoff between internal and external validity in design and interpretation of research studies, <https://statmodeling.stat.columbia.edu/2014/01/31/thicket-variation/>, and Andrew Gelman (2014), One-way street fallacy again! in reporting of research on brothers and sisters, <https://statmodeling.stat.columbia.edu/2014/02/03/one-way-street-fallacy-reporting-research-brothers-sisters/>.

<sup>52</sup>Dalton Conley and Emily Rauscher (2013), The effect of daughters on partisanship and social attitudes toward women, *Sociological Forum* 28, 700–718.

<sup>53</sup>Andrew Oswald and Nattavudh Powdthavee (2010), Daughters and left-wing voting, *Review of Economics and Statistics* 92, 213–227.

results! We took a look at both papers and couldn't immediately see a resolution, but we will offer some speculations, followed by some scattered comments.

From a political perspective, we can see the sense in either of these results. Start with the finding that having a girl makes you, on average, more conservative. Having a child should make you much more focused on family issues, and perhaps having a daughter (as compared to a son) might make this effect even larger. To the extent that the Republicans were seen as the family-values party, this could swing some votes (or, at least, some party identification). From the other direction, having a daughter could make you more sympathetic to left-wing views on reproductive rights and economic redistribution and thus push you to the left. In either case we're only talking about average behavior; individuals could go in either direction.

Perhaps unsurprisingly, Oswald and Powdthavee, being economists, focus on economic issues:

"The interesting recent work of Campbell (2004) documents systematic gender differences in modern British political attitudes. The author tabulates answers given in the British Election Survey of 2001. She shows that the single most important concern to males is that of low taxes. For females, by contrast, it is the quality of the National Health Service.

Similarly, left-wing parties tend to be more supportive of women-friendly policies such as family leave; and right-wing parties tend to favor increased spending on the military, which is of course mostly staffed by men."

Conley and Rauscher, being sociologists, center their story on social issues:

"Males' optimal reproductive strategy is to sire many offspring with a range of mates and push the parenting requirements onto the mothers. Meanwhile, the mother seeks to maximize not only the genetic fitness of the sire, but also to induce more post-conception investment in rearing the offspring from the father. Seen in this light, more conservative policies that increase the cost of promiscuity—particularly for males—will enhance the reproductive bargaining power of women. . . . The conservative emphasis on family, traditional values and gender roles, and prolife/anti-abortion sentiments all stress investment in children—for both men and women. Conservative policies mirror the genetic interests of women, writ large. They attempt to promote paternal investment in offspring."

To us, this seems to be a lot to hang on a fragile thread. When it comes to economic policy, Democrats are pushing higher taxes and public services while Republicans want lower taxes and public services—and we don't see how these map at all on to Conley and Rauscher's categories. Even when you come to the particular issue of abortion, we don't see how an anti-abortion policy is an "attempt to promote paternal investment in offspring." There are some deadbeat-dad laws out there, but we didn't think of them as being associated more with one party than the other.

Conley and Rauscher seem to have found a real pattern, so we don't want to dismiss their explanations too quickly. But we see their discussions of "reproductive bargaining power" and the like to be a bit of a distraction. To us it's simpler and clearer to think of liberalism and conservatism, or Democratism and Republicanism, to be linked to general attitudes about the family.

How did the two studies end up with opposite findings? Our guess is that it's differences between different countries, or maybe just noise in the data. Conley and Rauscher suggest that it's data issues (in particular, the inclusion or exclusion of adopted children or children who are no longer living at home), but both of the articles here consider various alternative specifications and the results don't change, so we doubt this is what's going on.

But for a completely different angle on the problem, consider how these findings were reported in the news media:

"The Effect of Daughters on Partisanship and Social Attitudes Toward Women"

- “Does Having Daughters Make You More Republican?”
- “Parents With Daughters Are More Likely To Be Republicans, Says New Study”
- “Parents Of Daughters Lean Republican, Study Shows”
- “The Daughter Theory: Does Raising Girls Make Parents Conservative?”

To their credit, the study’s authors and many of the journalists make it clear the claims are speculative (consider, for example, the question marks in two of the above headlines).

But here’s our question: Why is it all about “the effect of daughters”? Why not “Does having sons make you support the Democrats?” Having sons is considered the default. Okay, sure, a bit over 51% of babies are boys. Really, though, you can have a boy or a girl, and the whole discussion of these claims in the media is a bit distorted by the implicit attitude that the boy is a default. Lots of discussion about how you, as a parent, might change your views of the world if you have a girl. But not so much about how you might change your views if you have a boy. Lots of discussion of how having a girl might affect your attitudes on abortion, not so much discussion about how having a boy might affect your attitudes on issues such as gun control or war, which disproportionately involve young men. This is a real problem, when issues of girls and boys, men and women, are treated asymmetrically.

An example of this asymmetry came from *New York Times* columnist Ross Douthat, who expressed pleasure about the headline, “Study: Having daughters makes parents more likely to be Republican,” and wrote:<sup>54</sup>

“Why pleasure? Well, because previous research on this question had suggested the reverse, with parents of daughters leaning left and parents of sons rightward. And those earlier findings dovetailed neatly with liberal talking points about politics and gender: Republicans make war on women, Democrats protect them, so it’s only natural that raising girls would make parents see the wisdom of liberalism . . .

But the new study undercuts those talking points. Things are more complicated than you thought, liberals! You can love your daughters, want the best for them, and find yourself drawn to . . . conservative ideas! Especially if you’re highly educated, which is where the effect was strongest!”

Douthat gets credit for referring to the earlier study and writing that things are complicated, but the fallacy is that he is thinking unidirectionally. He’s all about what happens if you have a girl. But what happens if you have a boy? Parents of boys are drawn to . . . liberal ideas! Especially if you’re highly educated, which is where the effect was strongest! This would seem to contradict theories of the feminization of America, the “war on men,” and so forth.

This fallacy is not special to Douthat or to conservative columnists. Indeed, the political “gender gap” in America is typically framed as an advantage for Democrats who get these extra votes from women. It could be just as well be framed as a male gender gap in favor of the Republican party, but you don’t usually hear it that way. For another example, liberal columnist Charles Blow wrote about yet another study, this one by Andrew Healy and Neil Malhotra, finding that “men with sisters are more likely to be Republican. . . . A report from Stanford about the study concluded, ‘Watching their sisters do the chores ‘teaches’ boys that housework is simply women’s work, and that leads to a traditional view of gender roles—a position linked to a predilection for Republican politics.’”<sup>55</sup>

In all these cases, the fallacy is that the comparison could go either way, but people think about it only in one direction, thus not fully understanding the implications (in this case, thinking that a

<sup>54</sup>Ross Douthat (2013), The daughter theory, *New York Times*, 14 Dec, <https://www.nytimes.com/2013/12/15/opinion/sunday/douthat-the-daughter-theory.html/>.

<sup>55</sup>Charles Blow (2014), The masculine mistake, *New York Times*, 31 Jan, <https://www.nytimes.com/2014/02/01/opinion/the-masculine-mistake.html/>.

connection between daughters and Republican voting is good news for conservatives, because having sons is implicitly considered as the default case).

This story connects to the previous week's reading in the idea of a regression being a comparison. When comparing two groups using an indicator that takes on the value 0 or 1, the coefficient compares the 1 case to the default 0 case. Switching the labels on the indicator flips the coefficient. The story connects to the course more generally through the steps we take to understand regression coefficients and the implicit causal implication. There is a symmetry in causal effects: the effect of the treatment is relative to the control, and we could just as well speak of the effect of control relative to the treatment. In the case of the sex of a baby's sex—which is, essentially, a randomly-assigned treatment—it does not make sense to think of either sex as the default; the 0/1 labeling is arbitrary and we should recognize that.

### Class-participation activities

#### 1. Simulate and recover regression lines

Students will work in pairs. The first student should open R and simulate data from a regression line, creating a data frame with predictor  $x$  and outcome  $y$ . After creating the data, the student should type control-L to clear the R console and then print out the range of  $x$ . Then the computer is passed to the other student, who should fit the regression in R and then with pen on paper sketch a guess of the data that is consistent with this fitted model and range of  $x$ . Instructions for the activity are in Figure 41.

- Student #1:
  - (a) Create fake data from the model,  $y = a + bx + \text{error}$
  - (b) Put  $x$  and  $y$  in a data frame called `data`
  - (c) Type `library("rstanarm")`
  - (d) Type `ctrl-L` to clear the R console
  - (e) Type `range(data$x)`
- Student #2:
  - (a) Take the computer
  - (b) Fit the regression of  $y$  on  $x$  using `stan_glm`
  - (c) Sketch (not on the computer) your guess of the scatterplot of  $x$  and  $y$

Figure 41 Instructions for the activity of simulating and recovering regression lines. We project this on the screen or write on the board, and students can then follow along as they work on the activity.

This activity relates to the week's readings by decomposing the information in the fitted model. It is relevant to the course as a whole in that, in the real world, we often need to reconstruct the story without access to raw data, just based on a model that someone else has fit.

#### 2. How much do you have to move a point to shift the fitted line by a specified amount?

In the first part of this demonstration, students work in pairs. One student simulates data from a regression model and plots the data and fitted line. The other student shifts individual data points up and down to see what happens to the regression line. What does it take to make the slope change sign?

We project the following code onto the screen so that students can focus on the concepts of regression fitting and not have to struggle with coding up the simulation and graphs:

```
n <- 9
x <- 1:n
y <- 5 + 3*x + rnorm(n, 0, 10)
```

```
fake <- data.frame(x, y)

fit <- stan_glm(y ~ x, data=fake, refresh=0)
plot(fake$x, fake$y, ylim=c(0,100))
abline(coef(fit))

fake$y[9] <- fake$y[9] + 50
fit_2 <- stan_glm(y ~ x, data=fake, refresh=0)
points(fake$x[9], fake$y[9], col="blue")
abline(coef(fit_2), col="blue")
```

This activity relates to the week's reading on the influence of a data point in a regression, and it is relevant to the class more generally in allowing an example where we apply linear regression to data that do not fit the model.

### Computer demonstrations

#### 1. Play with the linear regression estimate

Here you can see how the regression line changes if you move points around in the data. Start with data in a straight line with no error:

```
# Prepare a plotting grid
par(mfrow=c(2,2))

# Generate fake data
n <- 11
x <- 1:n
y <- 0 + 1*x + rnorm(n, 0, 0.1)
fake <- data.frame(x, y)

# Plot the data
par(mar=c(3,3,2,1), mgp=c(1.5,.5,0), tck=-.01)
plot(fake$x, fake$y, pch=20, bty="l")
# Fit the model
fit <- stan_glm(y ~ x, data=fake, refresh=0)
abline(coef(fit), col="blue")
print(fit)
a_hat <- coef(fit)[1]
b_hat <- coef(fit)[2]
mtext(paste("y =", round(a_hat, 1), "+", round(b_hat, 1), "x"),
      side=3, line=1, col="blue")
```

Next we put the plotting code into a function:

```
simple_plot <- function(fake){
  plot(fake$x, fake$y, pch=20, bty="l")
  fit <- stan_glm(y ~ x, data=fake, refresh=0)
  abline(coef(fit), col="blue")
  a_hat <- coef(fit)[1]
  b_hat <- coef(fit)[2]
  mtext(paste("y =", round(a_hat, 1), "+", round(b_hat, 1), "x"),
        side=3, line=1, col="blue")
}
```

And now you can see what happens when you add 20 to the first point in the data:

```
new <- fake
new$y[1] <- new$y[1] + 20
```

### 3.9. LEAST SQUARES AND FITTING REGRESSION MODELS

121

```
simple_plot(new)
```

Let's try adding 20 to all the points:

```
new <- fake
new$y <- new$y + 20
simple_plot(new)
```

What would it take to get the line to have a negative slope? How much would you need to pull down one of the points to get that to happen? What happens if you move up or down the point in the middle? The instructor should ask these questions aloud and try these and other experiments, fielding different suggestions from students. In doing this, keep the values of  $x$  fixed, as there's enough to learn just from manipulating  $y$ .

This demonstration is relevant to the week's reading on the influence of individual points in a fitted regression (Section 8.3 of *Regression and Other Stories*), and it relates to the course more generally in showing the connection between individual data points and the fitted line. This demonstration also shows how it is possible to fit a linear regression to data that do not follow any linear pattern (as occurs when we move individual data points up and down).

#### 2. Compare `lm` and `stan_glm`

In *Regression and Other Stories*, we use the `stan_glm` function to fit linear regression. As discussed in Section 8.4 of that book, we use this Bayesian method for two reasons: first because it automatically returns simulations that we can use to summarize and propagate uncertainty in the fit, and second because it allows the use of prior information, a topic discussed further in Chapter 9 of *Regression and Other Stories*. The more standard approach is least squares regression, implemented in R using the `lm` function. In this demonstration, you can compare the two fitting functions using a simulated-data example:

```
library("rstanarm")
n <- 10
x <- 1:n
y <- 2 + 0.3*x + rnorm(n, 0, 1)
fake <- data.frame(x, y)
fit <- stan_glm(y ~ x, data=fake, refresh=0)
print(fit, digits=3)
fit_lm <- lm(y ~ x, data=fake)
print(fit_lm, digits=3)
fit_flat <- stan_glm(y ~ x, data=fake, prior_intercept=NULL, prior=NULL,
prior_aux=NULL, algorithm="optimizing", draws=0)
print(fit_flat, digits=3)
```

There are essentially no differences.

Next try with just 3 data points: just change `n <- 10` in the above code to `n <- 3` and re-run. The differences between the estimates are larger, although in this case there are still no real practical consequences. The differences become more important when using `stan_glm` with strong priors.

This demonstration is relevant to the week's reading on understanding the differences between least squares and Bayesian inference, and it is useful for the course as a whole in connecting to other regression software.

### Drills

#### 1. Sample size and standard errors

Here's a fitted regression:

```
stan_glm
  family: gaussian [identity]
  formula: earn ~ height
  observations: 1816
  predictors: 2
  -----
          Median MAD_SD
(Intercept) -85000   9000
height        1600    100

Auxiliary parameter(s):
  Median MAD_SD
  sigma 22000    400
```

- (a) Suppose this model were re-fit with a sample size of 7000 people, that is, approximately four times the sample size above. What would you expect the standard errors to look like?

*Solution:* Standard errors are proportional to  $1/\sqrt{n}$ , so you'd expect them to be multiplied by  $1/2$ : so, 4500, 50, and 200.

- (b) Suppose this model were re-fit with a sample size of 1000 people. What would you expect the standard errors to look like?  
(c) 100 people?  
(d) 1 million people?

## 2. Averages and comparisons as regression models

For each statement, express it as a regression in R code and algebra, and give the estimated regression coefficients.

- (a) 42% of people approve of the president's job performance.

*Solution:* Data points are survey respondents, and the outcome is  $y = 1$  if respondent approves or 0 if respondent does not approve. The regression is `stan_glm(y ~ 1)`, or  $y = a$ , with the estimate  $a = 0.42$ .

- (b) 46% of women and 38% of men approve of the president's job performance.  
(c) Average height is 63.7 inches for women and 69.1 inches for men.  
(d) 59% of Americans have received at least two covid vaccines.

## Discussion problems

### 1. From inference to decision

You run a regression on 100 students studying the effect of an educational intervention on a test score that has a population mean of 500 and standard deviation 100. You get an estimated treatment effect of 20 points with a standard error of 15 points. What does this tell you? Would you do the intervention? What might you do next?

### 2. Sample size and statistical significance

You run an experiment on 200 people and get an estimated treatment effect of 0.20 with standard error 0.15—so, not “statistically significant.” What might you expect to see if you re-ran with 400 people? Would you expect statistical significance then?

## 3.10 Prediction and Bayesian inference

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Fairness of random exams	Coverage of prediction intervals	Different forms of predictive uncertainty	Fairness of random exams	Coverage of prediction intervals
Uncertainties in election forecasts	Prior for a real-world parameter	Bayes estimate of childhood intervention	Uncertainties in election forecasts	Prior for a real-world parameter

### Reading

Chapter 9 of *Regression and Other Stories*: Prediction and Bayesian inference

### Pre-class warmup assignments

1. Different sorts of predictive uncertainty
  - (a) Simulate 100 data points from the regression model,  $y = 0.2 + 0.3x + \text{error}$ , with  $x$  drawn at random from the range (10, 20) and errors drawn from a normal distribution with mean 0 and standard deviation 0.4. Make a plot that includes the simulated data and a fitted regression line.
  - (b) Do `predict`, `posterior_linpred`, and `posterior_predict` for a new data point at the value  $x = 20$ . For the last two of these, summarize the distribution of predictions by their median and 50% predictive intervals. Discuss how these intervals differ.
  - (c) Do the same thing but for a new data point at the value  $x = 100$ .
2. Regression with informative priors

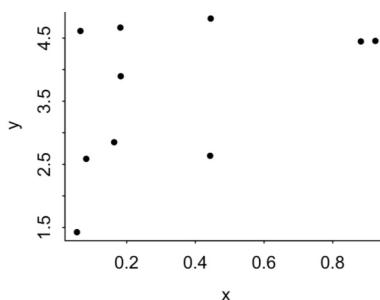
A randomized experiment is performed in a number of cities, estimating the effect of an experimental program to increase quality of life. At the beginning of the study, a survey is conducted in each city measuring residents' satisfaction on a 1–10 scale. Label  $x_i$  as the average satisfaction of residents surveyed in city  $i$ . Then some cities get the program ( $z_i = 1$ ) and some do not ( $z_i = 0$ ), and a year later the quality-of-life survey is repeated. Label  $y_i$  as the result for this new survey. A regression is fit predicting  $y$  given  $x$  and  $z$ .

  - (a) Write R code for fitting the regression using default priors.
  - (b) Write R code for fitting the regression with a flat prior on the effect of  $z$ .
  - (c) Write R code for fitting the regression with a prior that says the effect of  $z$  to be equally likely to be positive or negative, with an approximate 68% chance that the effect is in the range  $(-1.5, +1.5)$ .

### Homework assignments

1. (a) Influence of individual data points (Exercise 8.5 of *Regression and Other Stories*)

A linear regression is fit to the data below. Which point has the most influence (see Section 8.2 of *Regression and Other Stories*) on the slope?



2. (a) Prediction for a comparison (Exercise 9.1 of *Regression and Other Stories*)

A linear regression is fit on high school students modeling grade point average given household income. Write R code to compute the 90% predictive interval for the difference in grade point average comparing two students, one with household incomes of \$40 000 and one with household income of \$80 000.

(b) Predictive simulation for linear regression (Exercise 9.2 of *Regression and Other Stories*)

Using data of interest to you, fit a linear regression. Use the output from this model to simulate a predictive distribution for observations with a particular combination of levels of all the predictors in the regression.

(c) *In pairs*: Working through your own example (Exercise 8.11 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, take your fitted regression line from Exercise 7.10 of *Regression and Other Stories* and compute the influence of each data point on the estimated slope. Where could you add a new data point so it would have zero influence on the estimated slope? Where could you add a point so it would have a large influence?

## Stories

### 1. Fairness of random exams

We once did the following experiment in our class.<sup>56</sup> Without the knowledge of the students, we prepared two versions of the midterm examination, identical in all respects except that the order of the questions was reversed. We prepared equal numbers of the two versions and mixed them randomly before handing out one to each student for the exam. Each exam question was on a separate page and we graded the questions one at a time, so there's no reason to think the grading was influenced by the order of the questions. We then tallied the grades achieved by the two groups of students.

After returning the graded exams to the students, we revealed that there were two forms of the exam, shared the aggregate scores, and then discussed these results with the students. Reactions varied depending on which exam a student received. For example, suppose the average score was 65 for exam A and 71 for exam B. Should we adjust the scores of the "exam A" students upward (or adjust the "exam B" grades) to reflect that exam A seems more difficult, in retrospect? A student who took exam B might object, noting that the two exams had identical questions—just the order was different. But the order could have an effect, right? What if the two forms had been randomly given to 1000 students and this difference had been observed—would it be "real" then? The goal in this discussion was to get the "exam A" students and "exam B" students all fired up and holding opposite positions, to give them some skin in the game, as it were, in this debate about statistical uncertainty.

The data from the exams were being used to compare two different hypotheses or models of the

<sup>56</sup>Section 6.4.2 of *Teaching Statistics: A Bag of Tricks*.

world: one hypothesis is that the two exams are functionally identical and the observed difference was just due to chance, and the other hypothesis is that the order of the questions does make a difference. The choice of model makes a difference: under the first model, no adjustment should be made, whereas the second model implies that it's only fair for the students who took exam A to get a bonus.

How could we address the question of whether the observed difference is due to the exams or just because, say, the better students happened to take exam B? We can consider this as an experiment designed to measure the difference in exam difficulties.

If the true difference between the exams is approximately zero and the standard deviation of exam scores across students is 15, then with a class of 50 students the observed difference in means has an expected value of zero and a standard deviation of  $\sqrt{15^2/25 + 15^2/25} = 4.2$ . An observed difference of 6 points is then 1.4 standard errors away from zero.

Should we adjust the scores and, if so, how much? One way to handle this is using a Bayesian analysis, for which we would need a prior distribution on the effect size or population average difference in scores comparing the two exam forms. What if the exams differed not just in their ordering, but in the questions themselves? How would/should our statistical methods change? This is a subtle question with no easy answers. Students have also raised ethical questions about basing grades on different forms of the exam.

This example is relevant to the week's reading because it demonstrates the relevance of prior expectations in interpreting data. It relates to the course as a whole as an example of an experiment with ambiguous results.

## 2. Uncertainties in election forecasts

Elections are a high-profile event involving uncertainty, data, and probabilistic forecasting. Back in the old days, each pre-election poll would be reported on its own (for example, Reagan has 56% support with a margin of error of 3%), but in recent years there have been so many public polls that analysts and news organizations have performed "poll aggregation," whereby a series of polls are plotted and averaged to give a sense of public opinion and how it is changing over time. For a presidential election, national and state polls can be combined with the so-called "fundamentals" (such as the state of the economy; see Section 7.1 of *Regression and Other Stories*) to produce a state-by-state forecast of the election.

For this story you do not need to talk about how these forecasts are constructed; instead, the topic is how their uncertainties are displayed in graphics.

Figure 42 shows several visualizations of election forecasts from the *Economist* magazine.<sup>57</sup> We were involved in the *Economist* forecast, but we're not claiming that its uncertainty display is better than those of other media outlets.<sup>58</sup> Rather, the point of this story is that there is no unique way to display forecast uncertainty, and it can be useful to visualize it in different ways.<sup>59</sup>

In particular, when considering an election campaign, we need to juggle three sources of uncertainty and variation:

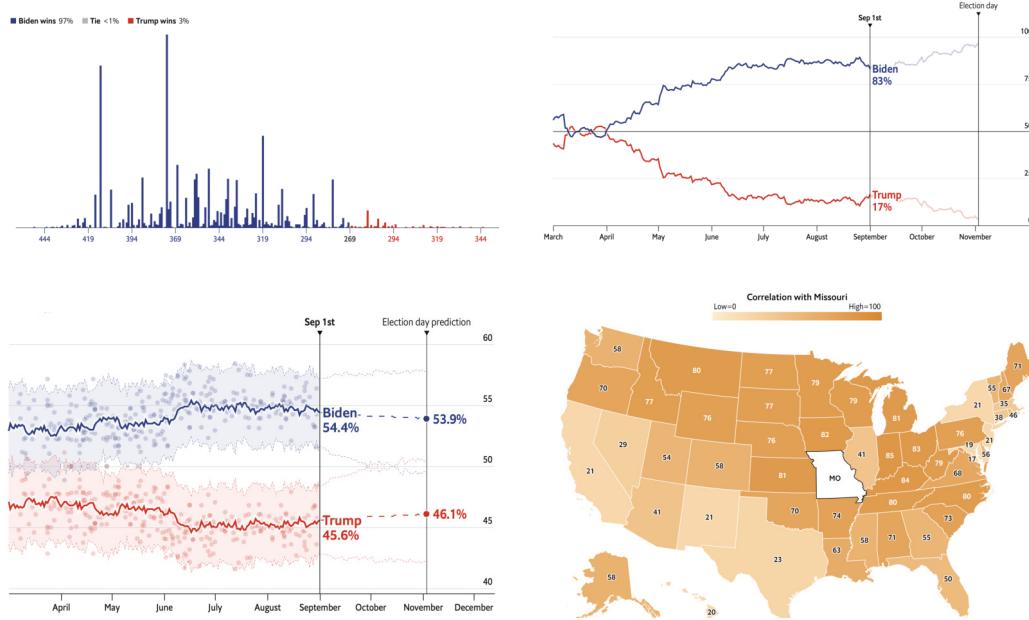
- Uncertainty in the national election outcome (as expressed both by popular and electoral vote)
- Variation over time as new information comes in
- Variation across the 50 states.

<sup>57</sup><https://projects.economist.com/us-2020-forecast/president/>.

<sup>58</sup>See, for example, <https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html> and <https://projects.fivethirtyeight.com/2020-election-forecast/>.

<sup>59</sup>Andrew Gelman, Jessica Hullman, Christopher Wlezien, and George Elliott Morris (2020), Information, incentives, and goals in election forecasts, *Judgment and Decision Making* 15, 863–880.

### 3. WEEK BY WEEK: THE FIRST SEMESTER



**Figure 42** Four visualizations of uncertainty in the Economist's 2000 presidential election forecast. Top row: Forecast of possible electoral vote outcomes and time series of Biden's and Trump's win probabilities during the course of the campaign. Bottom row: Time series of popular vote projections and an interactive display for examining between-state correlations in the forecast. No single visualization captures all aspects of uncertainty, but a set of thoughtful graphics can help readers grasp uncertainty and learn about model assumptions over time.

Different displays get at different aspects of this uncertainty. It's hard to visualize all of the variation at once, hence the value of multiple visualizations.

The upper-left plot in Figure 42 shows the predictive distribution of Donald Trump's electoral votes for the 2000 election, with 270 needed to win. The histogram represents forecast uncertainty: either candidate had a chance of winning, but Biden's chance was bigger, and he was likely to win by a lot of electoral votes.

The upper-right plot shows the time series of Biden and Trump's win probabilities, as estimated from the *Economist*'s model during the course of the campaign. We placed the slider of this interactive graph to the start of September, 2020, to show the two candidates' modeled chances of winning at that point.

The lower-left plot shows the corresponding time series forecast of the two-party popular vote. Each dot on the graph shows Biden's or Trump's support from a national poll, and the lines show the estimated path of vote intentions during the months leading up to that point, along with the forecast of the election outcome. Shaded areas and dotted lines show 80% uncertainty bounds; that is, according to this forecast, there was an 80% chance that Biden would receive between 49.5% and 58% of the popular vote. (He ended up with about 52%.) This plot is appealing in that it presents predictive uncertainty, changes over time, and (some of the) data used to construct the forecast. It does not, however, show what is happening in each state, and for that the *Economist* had similar state-by-state plots showing the polls and the forecast time series for each state.

They also included an interactive plot, one setting of which is shown in the lower right of Figure 42, showing the correlations between the predictions in different states. These correlations are a sort of insider knowledge that won't be of interest to most readers but can help build trust and understanding among the most sophisticated people in the audience.

As this story demonstrates, different graphs display different aspects of data and uncertainty and can also serve different readers. This example is relevant to the week’s reading because of the focus on predictive uncertainty: the model used to construct these forecasts is more complicated than simple linear regression, but the principles of propagation of uncertainty still apply. The story relates to the course more generally as a real-world example of statistical modeling and communication.

### Class-participation activities

#### 1. Coverage of prediction intervals

This is an activity demonstrating the challenge of expressing uncertainty.<sup>60</sup> We start by asking a student to guess some unknown quantity, for example the number of active duty U.S. military personnel. The actual value is 1.35 million,<sup>61</sup> but we do not reveal this number until the class has gone through an iterative guessing procedure as described below.

It is important to use a quantity whose true value can be looked up ahead of time. For example, in preparing this activity many years ago at Smith College, we first counted the number of people named Smith in the local telephone book. Then in class we asked the students to guess this number out loud. We wrote the first several guesses on the board: 156, 250, 72, 210, 150, 120, 200, 35, 76, 49, 50. Substantial differences were revealed.

We then asked one student to give a 50% probability interval (more precisely, the 25% and 75% quantiles) for the uncertain quantity—as we explained, this is an interval for which the student believes there is a  $\frac{1}{2}$  chance the true value is inside the interval, a  $\frac{1}{4}$  chance the interval is too high, and  $\frac{1}{4}$  chance the interval is too low. This interval is intended to summarize the student’s uncertainty. For example, when we did the demonstration at Smith, a student volunteered the interval [100, 200].

At this point, we invited the other students to comment. When students in the class would place more or less than 50% probability on the stated interval, we pointed out the opportunity for a bet that both parties should accept. We discussed with the class and altered the interval on the board until the students generally consider it reasonable. For example, most of the students in the class at Smith thought [100, 200] was too high, so we considered [75, 175]. There was a general consensus that 75 was too low—that is, the class judged that there was less than a  $\frac{1}{4}$  chance that the true value was below 75—so we adjusted again to [85, 175], which most of the class found reasonable.

We then revealed that the local telephone book had 458 listings of people named Smith. This result—a true value that was well outside the 50% interval—is typical. People’s uncertainty intervals tend to be too narrow—that is, they are overconfident. Again, when doing this activity, just about any unknown quantity will do, as long as it is something that students can guess and which the instructor can reveal the true value. The students may be reassured to see the well-known example displayed in Figure 43, in which a set of internationally known geotechnical engineers show overconfidence (in retrospect) in their probabilistic forecasts of the failure height of an embankment.<sup>62</sup>

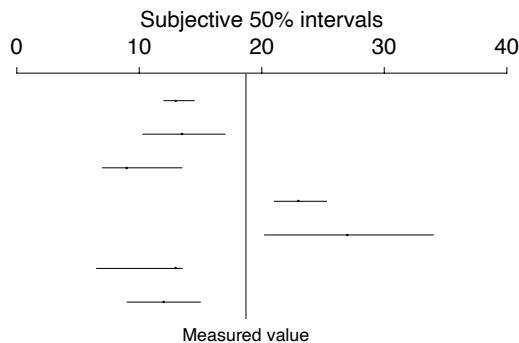
We continue with an adaptation of a classic classroom demonstration of subjective probability intervals.<sup>63</sup> We hand out Figure 44 and ask each of the students to fill out the form, giving a

<sup>60</sup>Section 17.2.2 of *Teaching Statistics: A Bag of Tricks*.

<sup>61</sup>See U.S. House of Representatives Committee on Appropriations (2021), Appropriations committee releases fiscal year 2022 defense funding bill, <https://appropriations.house.gov/news/press-releases/appropriations-committee-releases-fiscal-year-2022-defense-funding-bill/>.

<sup>62</sup>Mary Ellen Hynes and Erik Vanmarcke (1977), Reliability of embankment performance predictions, in *Mechanics in Engineering*, 367–384. University of Waterloo Press.

<sup>63</sup>Marc Alpert and Howard Raiffa (1982), A progress report on the training of probability assessors, in *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 294–305, Cambridge University Press.



**Figure 43** Classic example of overconfidence: experts' predictions and 50% predictive intervals of the height at which an embankment would fail, along with the true value. None of the predictions included the true value.

Uncertain quantity	25% lower bound	75% upper bound
% Black		
# eggs		
# airline deaths		
% girl births		
# babies born		
# abortions		
% degrees in CS		
# degrees		
# Super Bowl watchers		
\$ median income		

Give 25% and 75% probability bounds for each of these quantities. You should specify the bounds so that, for an unknown quantity  $x$ , there should be a 50% chance that  $x$  is between your upper and lower bounds. Fill in all the blanks on the table. You will then be told the true values of these quantities.

- (a) The percentage of people in the United States identifying as Black or African American (from the 2020 Census)
- (b) The total egg production in the United States in 1965 (in number of eggs)
- (c) The number of airline passengers worldwide who died in plane crashes in 2020
- (d) The percentage of babies born in the United States in 2000 that were girls
- (e) The number of babies born in the United States in 2000
- (f) The number of abortions in the United States in 2000
- (g) The percentage of bachelor's degrees in the United States in 2019 that were in the fields of computer and information sciences
- (h) The number of bachelor's degrees awarded in the United States in 2019
- (i) The number of people in the United States who watched the Super Bowl in 2020
- (j) The median household income in the United States in 2020

**Figure 44** A handout for demonstrating the difficulty of calibrating subjective probability intervals. The true values of the uncertain quantities are 12.4, 64.6 billion, 299, 48.8, 4.06 million, 857 000, 4.4, 2.01 million, 101.3 million, and 67 500. There is nothing special about this list; we encourage you to pick questions on topics that interest you.

50% subjective probability interval for each of the 10 unknown quantities on the list. Before announcing the true values of the 10 quantities, we ask each student to guess how many of the values fell within their intervals. If their uncertainty statements are calibrated, you'd expect to see approximately 5 out of 10 intervals contain the true values. We then announce the true values and ask students to count the number of their intervals that contains the truth. Invariably, the coverage is less than 50%; it is typically more like one-third, which is what has been seen in the literature.

This example demonstrates two points that are relevant both for the week's reading and for the course more generally. First, it is hard to give calibrated uncertainties: our bounds are typically too narrow. Second, it's better to try to express uncertainty than just to give point estimates.

## 2. Prior distribution for a real-world parameter

How does one come up with a prior distribution for a parameter in a real-world social science model? You can do this together with the class, using some example of interest. Our class had many political science students, and so we considered a hypothetical experiment of a treatment to increase voter turnout by knocking on doors. You first need to clearly define the treatment, the outcome of interest, and how it will be measured. Then, to build a prior distribution, start with a guess of the treatment effect. Next you need to assess uncertainty. One way to do this is to consider the prior probability that the treatment will have a positive effect. In our example, if the prior guess of the treatment effect is 0.01 (an increase in voter turnout of 1 percentage point) and the prior standard deviation is 0.01 also, then you can graph the prior distribution, and there is an 84% prior probability that the treatment effect is positive. Does this make sense? Sketch on the board and discuss with the class.

## Computer demonstrations

### 1. Different forms of predictive uncertainty

In this demonstration, based on Section 9.2 of *Regression and Other Stories*, you simulate fake data on IQ scores and course grades and then generate a point prediction for a new observation with IQ of 105:

```
n <- 100
x <- rnorm(n, 100, 15) # IQ, roughly
y <- 2.5 + 0.02*(x - 100) + rnorm(n, 0, 0.5) # GPA, roughly
fake <- data.frame(x, y)

# Fit regression
library("rstanarm")
fit <- stan_glm(y ~ x, data=fake, refresh=0)
print(fit)
plot(fake$x, fake$y)
abline(fit$coef)

# Generate point prediction
new <- data.frame(x=105)
y_point_pred <- predict(fit, newdata=new)
print(y_point_pred)

# Calculate point prediction by hand
a_hat <- coef(fit)[1]
b_hat <- coef(fit)[2]
y_point_pred_byhand <- a_hat + b_hat * 105
print(y_point_pred_byhand)
```

```
# Predict _average GPA_ among all students with IQ of 105
y_linpred <- posterior_linpred(fit, newdata=new)
hist(y_linpred)
print(c(mean(y_linpred), sd(y_linpred)))

# Predict GPA for a single student with IQ of 105
y_pred <- posterior_predict(fit, newdata=new)
hist(y_pred)
print(c(mean(y_pred), sd(y_pred)))
```

Then, some questions: What would happen to the predictions if  $n$  were increased from 100 to 400, or decreased from 100 to 25? How can we get a predicted probability that the student will have a GPA of greater than 3? You can answer that question using the simulations from `posterior_predict`:

```
mean(y_pred > 3)
```

You can also approximate using the regression parameters:

```
1 - pnorm(3, a_hat + b_hat*105, sigma(fit))
```

This last bit ignores the uncertainty in the estimated coefficients, which turns out to not be a big deal in this example.

## 2. Bayesian estimate of the effect of early childhood intervention

Section 1.5 of *Regression and Other Stories* discusses a study performed with Jamaican children that estimated the effect of early childhood intervention on earnings to be 42% with a standard error of 20%, thus a 95% confidence interval of [2%, 82%] (or something slightly different on the log scale, but here we will keep things simple and work on the untransformed scale). Here you can consider how the estimate changes with different priors.

An effect of 42% on earnings sounds pretty large, maybe even implausible. Suppose our prior on the true effect is normal with mean 0 and standard deviation 10%. Then what are the posterior mean and standard deviation of the effect?

```
post_mean <- (0/0.10^2 + 0.42/0.20^2) / (1/0.10^2 + 1/0.20^2)
post_se <- sqrt(1 / (1/0.10^2 + 1/0.20^2))
print(c(post_mean, post_se))
```

Write this as a function, entering code directly into the text editor (as always, in class the instructor should read aloud while typing) and then copying code one line at a time into the R console:

```
post_mean_and_se <- function (prior_est, prior_se, data_est, data_se) {
  post_mean <- (prior_est/prior_se^2 + data_est/data_se^2) /
    (1/prior_se^2 + 1/data_se^2)
  post_se <- sqrt(1 / (1/prior_se^2 + 1/data_se^2))
  c(post_mean, post_se)
}
```

And try it out:

```
post_mean_and_se(0, 0.10, 0.42, 0.20)
```

Then try other priors: flat prior, spike near zero, and a prior where the effect is expected to be positive but small:

```
post_mean_and_se(0, Inf, 0.42, 0.20)
post_mean_and_se(0, 0.01, 0.42, 0.20)
post_mean_and_se(0.05, 0.10, 0.42, 0.20)
```

You can also play around with the data and see how the inferences change.

## Drills

### 1. Prediction

Here's a fitted regression:

```
stan_glm
family: gaussian [identity]
formula: earn ~ height
observations: 1816
predictors: 2
-----
      Median MAD_SD
(Intercept) -85000    9000
height        1600     100

Auxiliary parameter(s):
  Median MAD_SD
sigma 22000    400
```

- (a) Approximately what is the predictive distribution from this regression of the earnings of a person who is 66 inches tall?

*Solution:* Point prediction is  $-85\,000 + 1600 * 66 = 20\,600$ . Predictive standard deviation is approximately 22 000 (actually, it's slightly higher). So from the regression the approximate predictive distribution is normal with mean 20 600 and standard deviation 22 000.

Actual earnings can't be negative, though, so the predictive distribution from the linear regression model is not so realistic in this case.

- (b) Approximately what is the predictive distribution from this regression of the earnings of a person who is 66 inches tall and male?  
(c) Approximately what is the predictive distribution from this regression of the earnings of a man who is randomly sampled from the population? Assume men's heights are normally distributed with mean 69.1 inches and standard deviation 2.9 inches.

### 2. Bayesian combination of information

- (a) Using a regression model, you forecast that a certain candidate will have 45% support, with forecast standard deviation of 5%. You then do a simple random sample survey of 100 people, of whom 50 support the candidate. What is the Bayesian posterior mean of the candidate's support in the population?

*Solution:* The prior estimate is  $0.45 \pm 0.05$ ; the data estimate is  $0.50 \pm 0.10$ . The prior is more precise than the data, so the weighted average is closer to the prior. In the weighted average, the prior estimate gets weight  $1/0.05^2$  and the data gets weight  $1/0.10^2$ . So the weights are in the ratio 4 to 1; the posterior mean is 0.46. Or use the formula,  $\frac{\frac{1}{0.05} \cdot 0.45 + \frac{1}{0.10} \cdot 0.50}{\frac{1}{0.05^2} + \frac{1}{0.10^2}} = 0.46$ .

The posterior standard deviation is  $\sqrt{\frac{1}{\frac{1}{0.05^2} + \frac{1}{0.10^2}}} = 0.044$ . The posterior probability that this candidate has at least 50% support is  $1 - \text{pnorm}(0.50, 0.46, 0.044)$ , equivalently,  $1 - \text{pnorm}((0.50 - 0.46) / 0.044)$ ; either way, it's 0.18.

- (b) Same problem, but the survey estimate is 55%.  
(c) Same problem, but the survey estimate is 50% with a sample size of 1000.  
(d) Same problem, but the survey estimate is 50% with a sample size of 1 million.

Then discuss how this last estimate is unreasonable because a survey can have bias.

### Discussion problems

#### 1. Interpreting statistically significant results given huge sample sizes

Suppose you run a regression with a huge sample size, for example a mega-poll with 100 000 respondents or an A/B test at a big company. With a large enough sample size, the standard error will be small, and even a small effect can be statistically significant. How can you interpret such a result?

For a related problem, suppose you are using a customer survey to forecast demand for some new service, and you are concerned that the survey has nonsampling error through some combination of the sample not being representative of the population and the responses being not completely trustworthy. How could you account for this in constructing a forecast with uncertainty?

#### 2. Everything that's not in the model

One of us bought a cheap bathroom scale.<sup>64</sup> We took the scale home and zeroed it—there's a little gear in front to turn. We tapped our foot on the scale, it went to  $-1$  kg, we turned the gear a bit, then it went up to  $+2$ , then we turned a bit back to get it exactly to zero, and tapped again . . . it was back at  $-1$ .

That was frustrating, but we still wanted to estimate our weight. So we got on and off the scale multiple times. The first few measurements were 66 kg, 65.5 kg, 68 kg, and 67 kg. A lot of variation! To get a good estimate in the presence of variation, it is recommended to take multiple measurements. So we did so. After 46 measurements, we got bored and stopped. The resulting measurements had mean 67.1 with standard deviation 0.7, hence standard error of  $0.7/\sqrt{46} = 0.1$ .

Do we really believe the resulting 95% confidence interval,  $67.1 \pm 0.2$ ? Of course not! The whole scale is off. Even if we could ever consistently zero it—which we can't—there's the trust factor. If your weight can be 66 on one weighing and 68 on another, what can you do with that?

One approach is to calibrate, either using a known object that weighs in the neighborhood of 67 kg or else our own weight measured on an accurate instrument. If that is not possible, then we would want a wider uncertainty interval to account for the uncertainty in the scale's bias. This sort of adjustment is discussed in Section 4.3 of *Regression and Other Stories*.

Here is the discussion problem. Based on the measurements summarized above, what is a reasonable uncertainty estimate for our weight? The variation in the measurements should give some upper bound on the uncertainty in the bias. There is no single correct answer here; it's a good example for considering how to handle the gaps between statistical models and reality.

<sup>64</sup>Andrew Gelman (2023), God is in every leaf of every tree (bathroom scale edition), <https://statmodeling.stat.columbia.edu/2023/01/06/god-is-in-every-leaf-of-every-tree-bathroom-scale-edition/>.

## 3.11 Linear regression with multiple predictors

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Incumbency advantage in elections	Memory quiz with treatment and pre-test	Regression with interactions	Causal regressions	Regression adjustment
Beauty and teaching evaluations	Design a study with regression in mind	Adding interactions to a model	Interpret interaction coefficients	What is gained by looking at a pre-test?

### Reading

Chapter 10 of *Regression and Other Stories*: Linear regression with multiple predictors

### Pre-class warmup assignments

#### 1. Indicator variables

A study is done predicting middle-school students' test scores given pre-test score, sex, and grade (6, 7, or 8).

- Explain why it is better when fitting the regression to code sex as a variable called `male` that equals 1 for men and 0 for women, rather than a variable `sex` that equals 1 for women and 2 for men.
- Write R code for fitting the regression, including grade as a linear predictor.
- Write R code for fitting the regression, including indicators for each grade.

#### 2. Linear transformations and interactions

Consider the following model predicting earnings given height and sex:

	Median	MAD_SD
(Intercept)	-9.3	15.2
height	0.4	0.2
male	-29.3	24.3
height:male	0.6	0.4

Auxiliary parameter(s):

	Median	MAD_SD
sigma	21.4	0.3

- Define the centered variable `male_c <- male - 0.5`. Give the estimated coefficients from regressing earnings on `height`, `male_c`, and their interaction.
- Define the centered variable `height_c <- height - 66`. Give the estimated coefficients from regressing earnings on `height_c`, `male`, and their interaction.
- Give the estimated coefficients from regressing earnings on `height_c`, `male_c`, and their interaction.

### Homework assignments

- (a) Uncertainty in the predicted expectation and the forecast (Exercise 9.3 of *Regression and Other Stories*)

Consider the economy and voting example from Section 7.1 of *Regression and Other Stories*. Fit the linear regression model to the data through 2012; these are available in the folder `ElectionsEconomy`. Make a forecast for the incumbent party's share of the two-party vote in a future election where economic growth is 2%.

- i. Compute the point forecast, the standard deviation of the predicted expectation from formula (9.1) of *Regression and Other Stories*, and the standard deviation of the predicted value from formula (9.2) of *Regression and Other Stories*.

Relevant passage from *Regression and Other Stories*:

In the regression model  $y = a + bx + \text{error}$ , the standard deviation for the linear predictor  $a + bx$  is

$$\hat{\sigma}_{\text{linpred}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \quad (3.1)$$

The standard deviation for the predicted value  $a + bx + \epsilon$  is

$$\hat{\sigma}_{\text{prediction}} = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \quad (3.2)$$

- ii. Now compute these using the relevant prediction functions discussed in Section 9.2 of *Regression and Other Stories*. Check that you get the same values as in part (i) of this problem.
2. (a) Regression with interactions (Exercise 10.1 of *Regression and Other Stories*)  
 Simulate 100 data points from the model,  $y = b_0 + b_1x + b_2z + b_3xz + \text{error}$ , with a continuous predictor  $x$  and a binary predictor  $z$ , coefficients  $b = c(1, 2, -1, -2)$ , and errors drawn independently from a normal distribution with mean 0 and standard deviation 3, as follows. For each data point  $i$ , first draw  $z_i$ , equally likely to take on the values 0 and 1. Then draw  $x_i$  from a normal distribution with mean  $z_i$  and standard deviation 1. Then draw the error from its normal distribution and compute  $y_i$ .
    - i. Display your simulated data as a graph of  $y$  vs.  $x$ , using dots and circles for the points with  $z = 0$  and 1, respectively.
    - ii. Fit a regression predicting  $y$  from  $x$  and  $z$  with no interaction. Make a graph with the data and two parallel lines showing the fitted model.
    - iii. Fit a regression predicting  $y$  from  $x$ ,  $z$ , and their interaction. Make a graph with the data and two lines showing the fitted model.
  - (b) Regression with interactions (Exercise 10.2 of *Regression and Other Stories*)  
 Here is the output from a fitted linear regression of outcome  $y$  on pre-treatment predictor  $x$ , treatment indicator  $z$ , and their interaction:

	Median	MAD_SD
(Intercept)	1.2	0.2
$x$	1.6	0.4
$z$	2.7	0.3
$x:z$	0.7	0.5

Auxiliary parameter(s):

Median	MAD_SD
sigma	0.5 0.0

- i. Write the equation of the estimated regression line of  $y$  on  $x$  for the treatment group and the control group, and the equation of the estimated regression line of  $y$  on  $x$  for the control group.
  - ii. Graph with pen on paper the two regression lines, assuming the values of  $x$  fall in the range (0, 10). On this graph also include a scatterplot of data (using open circles for treated units and dots for controls) that are consistent with the fitted model.
- (c) *In pairs:* Working through your own example (Exercise 10.11 of *Regression and Other Stories*)  
Continuing the example from the final exercises of the earlier chapters, fit a linear regression with multiple predictors and interpret the estimated parameters and their uncertainties. Your regression should include at least one interaction term.

## Stories

### 1. Incumbency advantage in congressional elections

Section 10.6 of *Regression and Other Stories* considers the problem of estimating the effects of incumbency in elections for the U.S. House of Representatives. Start with the example from the book, using incumbency status to predict the election outcome in 1988. Here's the regression on incumbency alone:

Median	MAD_SD	
(Intercept)	0.50	0.00
inc88	0.17	0.01

Auxiliary parameter(s):

Median	MAD_SD	
sigma	0.08	0.00

The outcome here is the Democratic share of the two-party vote in 1988, and `inc88` is equal to 1 if a Democrat is running for reelection, -1 if a Republican is running for reelection, and 0 if the district is an open seat. The coefficient for incumbency is 0.19: comparing two districts, one with no incumbent running and one with an incumbent running, on average you'd predict the party with the incumbent to get 19 percentage points more of the vote. This analysis only includes districts that were contested by both parties in both elections.

Next we adjust for previous vote:

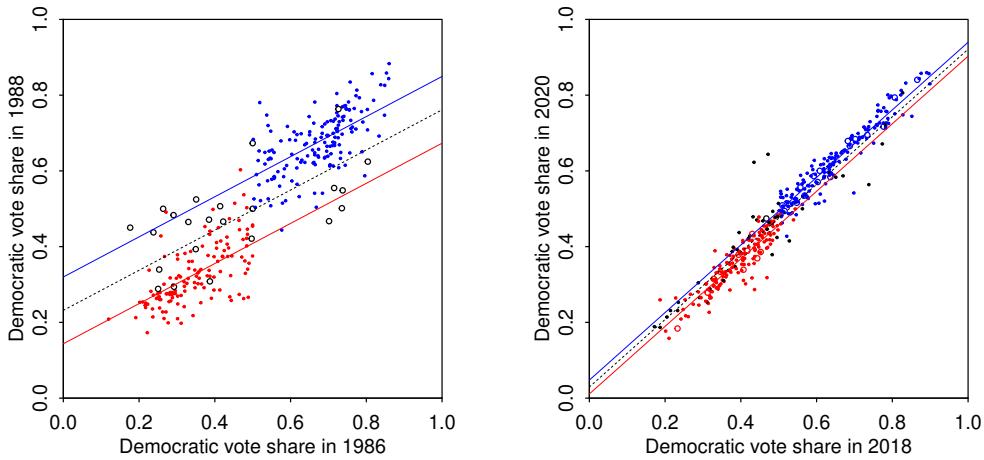
Median	MAD_SD	
(Intercept)	0.23	0.02
inc88	0.09	0.01
v86	0.53	0.04

Auxiliary parameter(s):

Median	MAD_SD	
sigma	0.07	0.00

The coefficient for incumbency has changed! Now it is just 0.09. Figure 45a shows the data and fitted regression. The regression has three lines corresponding to `inc88` equal to 1 (upper line on the graph), 0 (middle line), or -1 (lower line). Comparing two districts with identical voting patterns in the previous election, on average the model predicts the party with the incumbent to get 10 percentage points more of the vote. The estimated incumbency advantage is 10 percentage points.

How have things changed in the past few decades? We fit the same model to predict the district results in 2020:



**Figure 45** Contested elections for the U.S. House of Representatives in 1986 and 1988, and 2018 and 2020. Each graph shows the fitted regression predicting vote share in the second election from vote share in first election and incumbency status, a variable equal to +1 for Democratic incumbents (the dots on the graphs corresponding to points with  $x > 0.5$ ), -1 for Republican incumbents (the dots corresponding to points with  $x < 0.5$ ), and 0 for open seats (the open circles). The model in each case is a linear regression with no interactions, so the fit corresponds to parallel lines.

	Median	MAD_SD
(Intercept)	0.03	0.01
inc2020	0.02	0.00
v2018	0.89	0.02

Auxiliary parameter(s):  
 Median MAD\_SD  
 sigma 0.03 0.00

Now the estimated effect is only 2 percentage points. Figure 45b shows the results. The differences are striking. In the modern era of political polarization, there is little variation from election to election, and incumbency adds little predictive power.

This example relates to the week's reading as an example of building up from simple to multiple regression. It is relevant to the course as a whole in demonstrating a way in which we can learn by re-fitting a model to a new dataset. Statistics texts often focus on how to analyze a single dataset, but in practice we can make great progress by bringing in new data and looking at what has changed.

## 2. Predict teaching evaluations from beauty and other variables

These are data from teaching evaluations in economics classes at the University of Texas.<sup>65</sup>

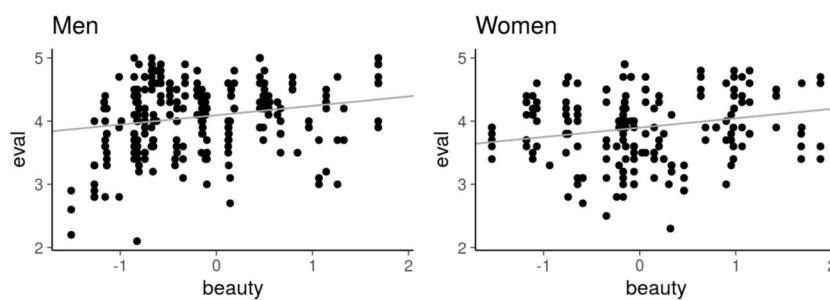
Here is a simple regression without interactions; the data and parallel fitted lines are displayed in Figure 46.

```
formula: eval ~ beauty + female
observations: 463
predictors: 3
-----
```

<sup>65</sup>The data come from Daniel Hamermesh and Amy Parker (2005), Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity, *Economics of Education Review* 24, 369–376; see Exercise 10.6 of *Regression and Other Stories*.

### 3.11. LINEAR REGRESSION WITH MULTIPLE PREDICTORS

137



**Figure 46** Data and fitted regression predicting teacher evaluations from instructors' attractiveness and sex, with no interactions. The fitted model is  $y = 4.09 + 0.15 * \text{beauty} - 0.20 * \text{female} + \text{error}$ .

	Median	MAD_SD
(Intercept)	4.09	0.03
beauty	0.15	0.03
female	-0.20	0.05

Auxiliary parameter(s):		
Median	MAD_SD	
sigma	0.54	0.02

Each data point is a class, and teaching evaluations are class averages on a 1 to 5 scale, with beauty of the instructor on a -2 to 2 scale, based on external ratings.

Next comes the model with an interaction:

	Median	MAD_SD
(Intercept)	4.11	0.03
beauty	0.20	0.04
female	-0.21	0.05
beauty:female	-0.11	0.06

Auxiliary parameter(s):		
Median	MAD_SD	
sigma	0.54	0.02

Is beauty a more important predictor for male or female instructors? The instructor can ask students to draw the regression lines and answer the question.

One can then add age as a predictor:

	Median	MAD_SD
(Intercept)	4.21	0.15
beauty	0.19	0.04
female	-0.22	0.05
age	0.00	0.00
beauty:female	-0.11	0.06

Auxiliary parameter(s):		
Median	MAD_SD	
sigma	0.54	0.02

What does it mean to have a coefficient of 0 with standard error 0? To aid in interpretability of the regression, we define a rescaled variable:

```
beauty$age10 <- beauty$age/10
```

Now here is the fit:

	Median	MAD_SD
(Intercept)	4.21	0.14
beauty	0.19	0.05
female	-0.22	0.05
age10	-0.02	0.03
beauty:female	-0.11	0.06

Auxiliary parameter(s):

Median	MAD_SD
sigma	0.54 0.02

The instructor can display this, then ask students to interpret this result and to draw the fitted regression lines on paper.

And here is the fit with one more predictor, an indicator for if English is not the instructor's native language:

	Median	MAD_SD
(Intercept)	4.23	0.14
beauty	0.19	0.04
female	-0.21	0.05
age10	-0.02	0.03
nonenglish	-0.34	0.10
beauty:female	-0.11	0.06

Auxiliary parameter(s):

Median	MAD_SD
sigma	0.53 0.02

This example is relevant to the week's reading in that it is about interpretation of coefficients in multiple regression. It fits into the course more generally as an example of the use of regression to explore data without there being a particular question of interest.

### Class-participation activities

#### 1. Memory quiz with pre-test, treatment, and outcome

This activity returns to a memory-quiz demonstration from earlier in the semester. Before class, we use the computer to simulate 20 random nouns.<sup>66</sup> Here's what came up for us:

friend verse cloth curtain metal attempt comparison size balloon match tiger  
form cable maid dress expansion worm goose mother liquid

We tell the students that we will be displaying 20 randomly chosen nouns on the screen for 30 seconds; they should try to memorize as many as they can. We then do this. When the 30 seconds are up, we spend three minutes discussing the rest of this activity (see below). When the three minutes have passed, we give the students a minute to individually write down as many words as they remember. We then display the words again and ask them to count how many they got correct.

The next part of the activity is another memory quiz with a new set of 20 nouns. But this time half the students will get 30 seconds to see the words and half will get 60 seconds. We do the randomization using the last digits of students' Social Security numbers: students with odd numbers get 30 seconds, students with even numbers get 60 seconds. Students without social

<sup>66</sup>For example, <https://wordcounter.net/random-word-generator>.

### 3.11. LINEAR REGRESSION WITH MULTIPLE PREDICTORS

139

Timestamp	x (pre-test)	z (0 if 30-second or 1 if 60-second)	y (test 2)
11/29/2021 9:25:41	11	1	13
11/29/2021 9:25:43	3	1	6
11/29/2021 9:25:45	6	0	7
11/29/2021 9:25:47	10	1	14
11/29/2021 9:25:48	8	1	9
11/29/2021 9:25:50	7	0	7
11/29/2021 9:25:50	8	0	6

Figure 47 Google form we created in class for students to enter data from the memory quiz experiment, along with the first few rows of the spreadsheet of responses.

security numbers can flip coins, with heads corresponding to odd and tails corresponding to even. We then display the next set of words:

power snail screw cake curve unit writing driving sister hair baby scarecrow  
cry discussion collar channel trousers sheep brick ocean

After 30 seconds, we tell the students with odd Social Security numbers to close their eyes. The students with even numbers get to look for another 30 seconds.

We then spend three minutes discussing the regression models we will fit to these data, first to predict outcome from pre-test and treatment, then including the interaction of these two predictors.

After three minutes, we give students a minute to individually write down as many words as they remember. We then display the words again and ask them to count how many they got correct. Next we ask students to enter their data  $x, y, z$  on a Google form that we create in front of them; see Figure 47.

Before analyzing the data, we ask students in pairs to sketch the regression lines they might expect to see and give possible coefficients. We then fit the regressions and discuss how these differed from students' guesses.

When we did this in class, students guessed that the coefficients for  $x$  and  $z$  would both be positive in the simple regression. They were less clear what would happen with interactions.

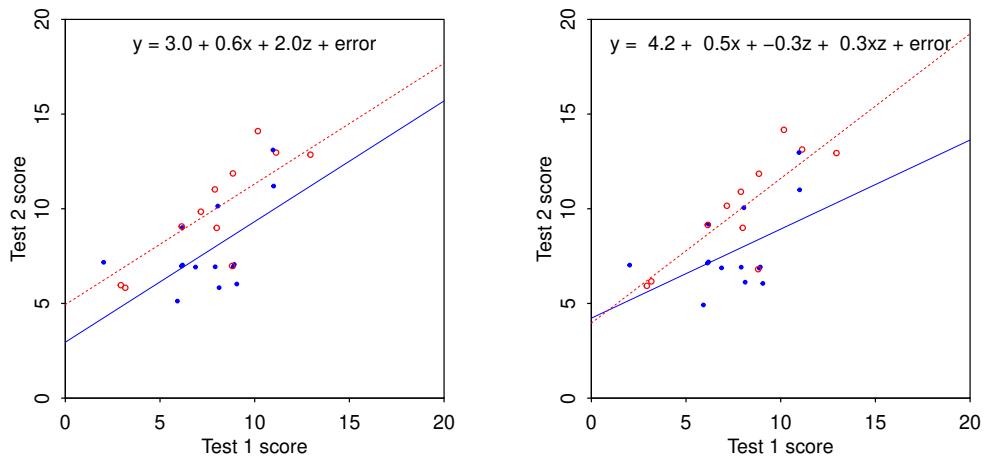
Here's the result from the data we gathered from our class:

```
formula:      y ~ x + z
observations: 25
predictors:   3
-----
              Median MAD_SD
(Intercept) 3.0    1.2
x           0.6    0.1
z           2.0    0.7

Auxiliary parameter(s):
Median MAD_SD
sigma 1.9    0.3
```

Whatever result we get, we want to go through the interpretation of each coefficient.

Here's our result for the model with the interaction:



**Figure 48** Data and fitted regressions predicting students' scores (number of words correctly identified) on a memory quiz, given score on an earlier test and a treatment: students with  $z = 0$  (open circles) were given 30 seconds to memorize the words, and students with  $z = 1$  (dots) were given 60 seconds. The left and right graphs show fitted regressions with and without interactions.

```

formula:      y ~ x + z + x:z
observations: 25
predictors:   4
-----
           Median MAD_SD
(Intercept) 4.2     1.7
x           0.5     0.2
z          -0.2     2.3
x:z        0.3     0.3

Auxiliary parameter(s):
Median MAD_SD
sigma 1.9     0.3

```

The uncertainties on the coefficients are large, but in our discussion we focus on the fitted lines without concerning ourselves with standard errors. To understand these coefficients, separately work out the lines for  $z = 0$  and  $z = 1$ . For example, the above fit can be written as,

$$\begin{aligned} \text{For } z = 0 : \quad & y = 4.2 + 0.5x + \text{error} \\ \text{For } z = 1 : \quad & y = 4.0 + 0.8x + \text{error}. \end{aligned}$$

Whatever lines we get from the fitted regression, we draw them over the range of  $x$  from 0 to 20. Recall that  $x$  and  $y$  are measures of the number of words guessed correct out of 20, so it is not possible to see data outside that range.

When we did this demonstration in class, we got an unexpected result. The coefficient for  $z$  in the model with interactions was negative! But did this really represent evidence that the treatment has a negative effect? (Again, we ignore the standard error for now, as our first challenge is to understand the fitted model.) Figure 48 shows the data and fitted models. We purposely delayed showing this graph to students because we want to train them to interpret the table of coefficients. Given the fitted model, they should be able to sketch the lines.

In the example of Figure 48, the graph on the right shows how tricky it can be to interpret

### 3.11. LINEAR REGRESSION WITH MULTIPLE PREDICTORS

141

coefficients in the presence of interactions. That  $-0.2$  coefficient for  $z$  corresponds to a lower prediction for  $y$  right at  $x = 0$ , but for almost the entire possible range of the data the red line is actually higher than the blue line, indicating that the prediction is higher under the treatment than the control.

This activity relates to the week's reading by demonstrating regression without and with interactions. It relates to the course as a whole as an example of causal inference adjusting for a pre-treatment variable.

#### 2. Design a study with regression in mind

Consider the memory quiz experiment from the design perspective. We start with all the options in the experiment: the number of words, the time given for memorization, the lag time before writing down the words, the option of other pre-treatment variables, and the treatments. We discuss with the class how the experiment would go with other settings. Issues that arise include cost (in class time) of data collection, floor and ceiling effects, and effect size. For example, if the treatment effect is too small, it will be difficult to detect it in the context of variation in the data.

This activity is relevant to the week's reading in that we are considering what would happen after fitting a multiple regression. It relates to the course as a whole in turning the focus from data analysis to design and data collection.

#### Computer demonstrations

##### 1. Regression with interactions

Students recorded their data from the memory quiz activity on a Google form (see Figure 47), and we saved the spreadsheet of responses in a file that we called `memory2.csv`. These data can be used to illustrate some important techniques for regression analysis.

The first step is to read in the data and fit the regression without interactions:

```
memory <- read.csv("memory2.csv", col.names=c("Time", "x", "z", "y"))
fit <- stan_glm(y ~ x + z, data=memory, refresh=0)
print(fit)
```

Here is the result:

	Median	MAD_SD
(Intercept)	3.9	1.3
x	0.6	0.2
z	-0.1	0.1

Auxiliary parameter(s):

	Median	MAD_SD
sigma	2.1	0.3

The coefficient for  $z$  is negative—that's surprising! OK, it has a high uncertainty, but still, one would have expected the treatment (an extra 30 seconds of memorization time) to improve students' scores on a memory quiz.

It's time to look more carefully at the data, first printing out the average values of  $x$ ,  $z$ , and  $y$ :

```
for (j in 2:4) print(mean(memory[,j]))
```

And this is what comes out:

```
[1] 7.72
[1] 1.64
[1] 8.76
```

The means of  $x$  and  $y$  seem reasonable enough, given that these are scores on a 0–20 scale. But how can the mean of  $z$  be 1.64? This is a variable that is only allowed to take on the values 0 and 1.

Examine the data more closely:

```
print(memory[,2:4])
```

Now you can see the problem. In the 8th row of data, a student entered the value 30 for  $z$ . Our instruction was to give  $z = 0$  if you were given 30 seconds to memorize the words or  $z = 1$  if you were given 60 seconds; see Figure 47. One of the students mistakenly entered 30 where it should have been 0.

Go in and fix this:

```
memory$z[memory$z==30] <- 0
```

When conducting this demonstration, we pause here to explain to the students that it is safest to do this sort of correction in the script, rather than to directly edit the raw data file, which can potentially introduce new errors. We then re-fit the model and also fit the model with interactions:

```
fit_2 <- stan_glm(y ~ x + z, data=memory, refresh=0)
print(fit_2)

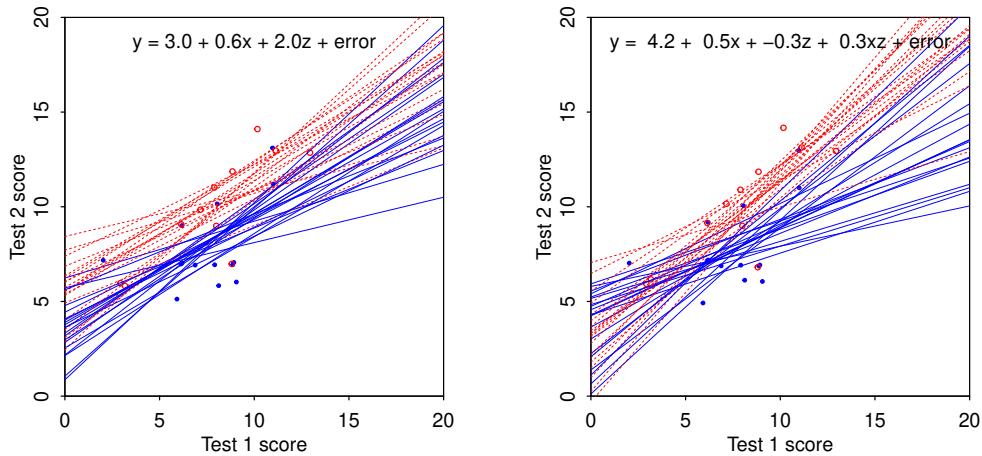
fit_3 <- stan_glm(y ~ x + z + x:z, data=memory, refresh=0)
print(fit_3)
```

Another thing that can be done is to plot the data and fitted regression lines, as shown in Figure 48. To keep things simple, we recommend you just go through one of these, the plot for the model with interactions:

```
jitt_x <- runif(nrow(memory), -0.2, 0.2)
jitt_y <- runif(nrow(memory), -0.2, 0.2)
par(mar=c(3,3,1,1), mgp=c(1.5,.5,0), tck=-.01, pty="s")
plot(memory$x + jitt_x, memory$y + jitt_y,
     col=ifelse(memory$z==0, "blue", "red"),
     pch=20, cex=.6, xlim=c(0,20), ylim=c(0,20), xaxs="i", yaxs="i",
     xlab="Test 1 score", ylab="Test 2 score")
abline(coef(fit_3)[1], coef(fit_3)[2], col="blue", lwd=.5)
abline(coef(fit_3)[1] + coef(fit_3)[3], coef(fit_3)[2] + coef(fit_3)[4],
       col="red", lwd=.5, lty=2)
beta <- fround(coef(fit_3), 1)
mtext(paste("y = ", beta[1], " + ", beta[2], "x + ", beta[3], "z + ", beta[4],
            "xz + error", sep=""), side=3, line=-1.5)
```

The next step is to use the posterior simulations to display uncertainty in the fitted regression lines. The result is shown in Figure 49. Again, we just show the code for the graph on the right, corresponding to the model with interactions.

```
sims_3 <- as.matrix(fit_3)
par(mar=c(3,3,1,1), mgp=c(1.5,.5,0), tck=-.01, pty="s")
plot(memory$x + jitt_x, memory$y + jitt_y, col=ifelse(memory$z==0, "blue", "red"),
     pch=20, cex=.6, xlim=c(0,20), ylim=c(0,20), xaxs="i", yaxs="i",
     xlab="Test 1 score", ylab="Test 2 score")
for (i in sample(nrow(sims_2), 20)){
  abline(sims_3[i,1], sims_3[i,2], col="blue", lwd=.5)
  abline(sims_3[i,1] + sims_3[i,3], sims_3[i,2] + sims_3[i,4],
         col="red", lwd=.5, lty=2)
}
```



**Figure 49** Uncertainties in the fitted lines from the models shown in Figure 48 predicting a memory quiz score from a pre-test and a treatment indicator. Each graph shows 20 random draws from the posterior distribution of the regression line.

```
beta <- fround(coef(fit_3), 1)
mtext(paste("y = ", beta[1], " + ", beta[2], "x + ", beta[3], "z + ", beta[4],
           "xz + error", sep=""), side=3, line=-1.5)
```

2. Adding predictors and interactions to a model; comparing regression with interactions with two separate regressions

Consider the earnings and height example from Sections 10.4 and 12.4 of *Regression and Other Stories*:

```
library("rstanarm")
earnings <- read.csv(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Earnings/data/",
  "earnings.csv"
))

# Regress earnings on height and sex
fit_1 <- stan_glm(earnk ~ height + male, data=earnings, refresh=0)
print(fit_1)

# Plot data and fitted lines
plot(earnings$height, earnings$earnk, pch=20, cex=.3,
     col=ifelse(earnings$male==1, "blue", "red"))
abline(coef(fit_1)[["Intercept"]], coef(fit_1)[["height"]], col="red")
abline(coef(fit_1)[["Intercept"]] + coef(fit_1)[["male"]], coef(fit_1)[["height"]],
      col="blue")

# Include interaction
fit_2 <- stan_glm(earnk ~ height + male + height*male, data=earnings, refresh=0)
print(fit_2)

# Plot data and fitted lines
plot(earnings$height, earnings$earnk, pch=20, cex=.3,
     col=ifelse(earnings$male==1, "blue", "red"))
abline(coef(fit_2)[["Intercept"]], coef(fit_2)[["height"]], col="red")
```

```
abline(coef(fit_2)["(Intercept)"] + coef(fit_2)[["male"]],
       coef(fit_2)[["height"]] + coef(fit_2)[["height:male"]], col="blue")
```

We go through interpretation of all four coefficients in this model, then fit separate regressions for each sex and compare these to the fitted regression with interaction:

```
# Separate regression for women
fit_3a <- stan_glm(earnk ~ height, data=earnings, refresh=0, subset=(male==0))
print(fit_3a)
# Separate regression for men
fit_3b <- stan_glm(earnk ~ height, data=earnings, refresh=0, subset=(male==1))
print(fit_3b)
# Plot data and fitted lines
plot(earnings$height, earnings$earnk, pch=20, cex=.3,
     col=ifelse(earnings$male==1, "blue", "red"))
abline(coef(fit_3a)["(Intercept)"], coef(fit_3a)[["height"]], col="red")
abline(coef(fit_3b)["(Intercept)"], coef(fit_3b)[["height"]], col="blue")
```

## Drills

### 1. Causal regressions

Set up each of these problems as a regression model.

(a) Effect of tutoring on test scores.

*Solution:* post\_test ~ treatment + pre\_test or  $y \sim z + x$

(b) Effect of a drug that is intended to lower blood pressure

(c) Effect of sanctions on a country's nuclear weapons program.

### 2. Interpret interaction coefficients

For each model, describe each coefficient in words.

(a)  $\text{hourly\_earnings} = -10.4 + 2.6*\text{educ} - 1.7*\text{female} - 0.3*\text{female}*\text{educ} + \text{error}$

*Solution:*

- For a man with zero years of education, we predict his hourly earnings to be -\$10.40, on average.
- Comparing two men who differ by one year in education, we expect the man with more education to have \$2.60 higher earnings, on average.
- Comparing a man and a woman who both have zero years of education, we expect the woman to have \$1.70 lower earnings, on average.
- The slope for education is 0.3 lower for women than for men. So, comparing two *women* who differ by one year in education, we expect the woman with more education to have  $\$2.60 - \$0.30 = \$2.30$  higher earnings, on average.

(b)  $\text{wage} = -0.9 + 0.5*\text{yrseduc} + 0.3*\text{married} + 0.1*\text{yrseduc}*\text{married} + \text{error}$

(c)  $y = 1.2 + 1.6x + 2.7z + 0.7xz + \text{error}$

## Discussion problems

### 1. Adjusting for pre-treatment variables

Consider an observational study on a topic of interest, comparing exposed to unexposed groups. Define the units of observation, the treatment and control (exposed and unexposed) conditions, an outcome of interest, and two pre-treatment predictors. For example, the units could be states

### 3.11. LINEAR REGRESSION WITH MULTIPLE PREDICTORS

145

in election years; the exposure could be a natural disaster in the state during the previous year; and the outcome could be the incumbent party's vote share in the election for governor, with pre-treatment predictors being the party's vote share in the previous elections for governor and president. For another example, the units are college students, exposure is taking a political science class in college, the outcome is a survey response regarding support for democratic institutions, and background variables are measures of parents' political attitudes.

For this problem, each pair of students should come up with an example on a topic of interest to them and write two regression models, first giving the simple comparison between the two groups and then adjusting for the pre-treatment predictors. The focus here is on how to write these models in mathematical notation and in R, not to be overly concerned with causal identification, which we return to in the next semester.

#### 2. What is gained by including a pre-test?

Consider an educational experiment with treatment,  $z$ , and outcome being a test score,  $y$ . Now consider including a pre-test,  $x$ , so that you estimate the treatment effect by fitting a regression of  $y$  on  $z$  and  $x$ . How will it help to include  $x$  in the model? How much will it help?

For this problem, you should address this question by designing a simulation study, where  $x$ ,  $z$ , and  $y$  are simulated under some assumptions and then the estimates can be compared with and without adjusting for  $x$ . Put that in a loop, and you can calculate how the estimates perform on average.

## 3.12 Assumptions, diagnostics, and model evaluation

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Actual vs. guessed exam scores	Sample size and statistical significance	Take difference or regress on an indicator	Assumptions and how they fail	Assumptions of regression
Model checking for baseball analytics	Assumptions of regression	Simulate and debug of regression	How to test regression assumptions	Patterns of residuals

### Reading

Chapter 11 of *Regression and Other Stories*: Assumptions, diagnostics, and model evaluation (Sections 11.1–11.7)

### Pre-class warmup assignments

1.  $R^2$  and explained variance
  - (a) Suppose the following regression is fit to the students in a class:  $\text{post\_test} = 30 + 0.7 * \text{pre\_test} + \text{error}$ . Explain the coefficient 0.7 in “comparative” or “descriptive” terms without using causal language.
  - (b) Suppose a regression is run, predicting students’ post-test from pre-test scores, and the  $R^2$  is 1. What does this tell us about the fitted model?
  - (c) Suppose a regression is run, predicting students’ post-test from pre-test scores, and  $R^2$  is 0. What does this tell us about the fitted model?
2. Linear regression fit to nonlinear data
  - (a) Simulate 50 data points  $y_i$  from the model  $y_i = a * \exp(-bx_i) * \exp(\text{error}_i)$ , with  $a = 0.1$ ,  $b = 0.2$ , and errors drawn from a normal distribution with mean 0 and standard deviation 0.2. In your simulation, draw  $x_i$  at random from the range (0, 20).
  - (b) Fit a linear regression to these data. Plot the data and the fitted regression line.
  - (c) Plot the residuals from the fitted model vs.  $x$ . Does the plot show any problems?

### Homework assignments

1. (a) Regression modeling and prediction (Exercise 10.5 of *Regression and Other Stories*)

The folder KidIQ contains a subset of the children and mother data discussed earlier in the chapter. You have access to children’s test scores at age 3, mother’s education, and the mother’s age at the time she gave birth for a sample of 400 children.

- i. Fit a regression of child test scores on mother’s age, display the data and fitted model, check assumptions, and interpret the slope coefficient. Based on this analysis, when do you recommend mothers should give birth? What are you assuming in making this recommendation?
- ii. Repeat this for a regression that further includes mother’s education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?
- iii. Now create an indicator variable of whether the mother has completed high school. Consider interactions between high school completion and mother’s age. Also create a plot that shows the separate regression lines for each high school completion status group.

- iv. Finally, fit a regression of child test scores on mother's age and education level for the first 200 children, and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

(b) Regression models with interactions (Exercise 10.6 of *Regression and Other Stories*)

The folder Beauty contains data (use file beauty.csv) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

- i. Run a regression using beauty (the variable beauty) to predict course evaluations (eval), adjusting for various other predictors. Graph the data and fitted model, and explain the meaning of each of the coefficients along with the residual standard deviation. Plot the residuals versus fitted values.
- ii. Fit some other models, including beauty and also other predictors. Consider at least one model with interactions. For each model, explain the meaning of each of its estimated coefficients.

2. (a) Assumptions of the regression model (Exercise 11.1 of *Regression and Other Stories*)

For the model in Section 7.1 of *Regression and Other Stories* predicting presidential vote share from the economy, discuss each of the assumptions in the numbered list in Section 11.1 of *Regression and Other Stories*. For each assumption, state where it is made (implicitly or explicitly) in the model, whether it seems reasonable, and how you might address violations of the assumptions.

(b) Descriptive and causal inference (Exercise 11.2 of *Regression and Other Stories*)

- i. For the model in Section 7.1 of *Regression and Other Stories* predicting presidential vote share from the economy, describe the coefficient for economic growth in purely descriptive, non-causal terms.
- ii. Explain the difficulties of interpreting that coefficient as the effect of economic growth on the incumbent party's vote share.

(c) *In pairs:* Working through your own example (Exercise 11.11 of *Regression and Other Stories*)

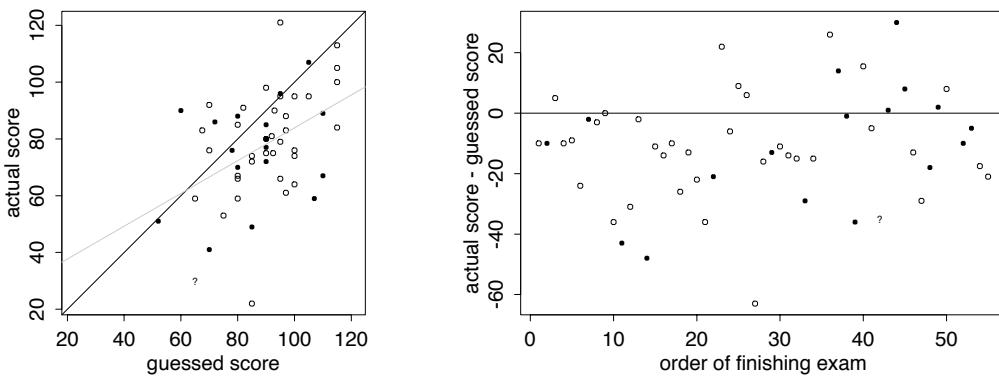
Continuing the example from the final exercises of the earlier chapters, graph and check the fit of the model you fit in Exercise 10.11 of *Regression and Other Stories* and discuss the assumptions needed to use it to make real-world inferences.

## Stories

### 1. Actual vs. guessed exam scores

We sometimes have included a question at the end of an exam asking students to guess their total scores on the other questions of the exam.<sup>67</sup> As an incentive, students received 5 points extra credit if the guess was within 10 points of the actual score. When the students completed their exams, we kept track of the order in which they were handed in, so that we could later check to see if students who finished the exam early were more or less accurate in their self-assessments than the students who took the full hour. When grading the exams, we did not look at the guessed score until all the other questions were graded. We then recorded the guessed grade, actual grade, and order of finish for each student. We had three reasons for including the self-evaluation question: it forced the students to check their work before turning in the exam; it taught them that subjective predictions can have systematic bias (in our experience, students tended to be overly optimistic about their scores); and the students' guesses provided us with an example of data and residuals, as we discuss next.

<sup>67</sup>From Section 3.5.1 of *Teaching Statistics: A Bag of Tricks*.



**Figure 50** (a) Actual vs. guessed midterm exam scores for a class of 53 students. Each symbol represents a student; empty circles are men, solid circles are women, and ? has unknown sex. The  $45^\circ$  line represents perfect guessing, and the dotted line is the linear regression of actual score on guessed score. Both men and women tended to perform worse than their guesses. (b) Difference between actual and guessed midterm exam scores, plotted against the order of finishing the exam. The exact order is only relevant for the first 20 or 25 students, who finished early; the others all finished within five minutes of each other at the end of the class period. The horizontal line represents perfect guessing. The students who finished early were highly optimistic, whereas the other students were less biased in their predictions.

Figure 50a displays the actual and guessed scores (out of a possible score of 125) for each student in a class of 53, with students indicated by solid circles (women), empty circles (men), and ? for a student with unknown sex. (This student had an indeterminate name, was not known by the teaching assistants, and dropped the course after the exam.) The points are mostly below the 45-degree line, indicating that most students guessed too high. Perhaps surprisingly, men do not appear appreciably more overconfident than women. The dotted line shows the linear regression of actual score on guessed score and displays the typical “regression to the mean” behavior discussed in Chapter 6 of *Regression and Other Stories*.

A class discussion should bring out the natural reasons for this effect. Figure 50b shows the difference between actual and guessed scores, plotted against the order of finish. Many of the first 20 or 25 students, who finished early, were highly overconfident; whereas the remaining students, who took basically the full hour to complete the exam, were close to unbiased in their predictions. Perhaps this suggests that students who finish early should take more time to check their results. (The students who finished early did, however, have higher than average scores on the exam.) Other studies have found that students’ test performances are overestimated by their teachers as well.

The data also have a detective-story aspect that can be fun to discuss. For example, why do the guesses scores max out at 115? Since students got extra credit for a guess within 10 points of their exam score, and the exam was only worth 125 points, it would not make sense to guess higher than 115. (In fact, however, all four students who guessed 115 were overconfident about their grades.) What about the student with uncertain sex who guessed 65 and scored 30? How could someone guess so poorly? There could be a logical motivation, based on the following reasoning: if he or she scored below 55 on the exam, he or she would drop the course anyway. The extra credit points would then only be useful with a score above 55, hence the guess of 65. In fact, the student did drop the course after the low exam score.

When teaching this course again, we varied the procedure by handing out Figure 50 a week before the midterm exam, discussing the overconfidence phenomenon, and warning the class that the

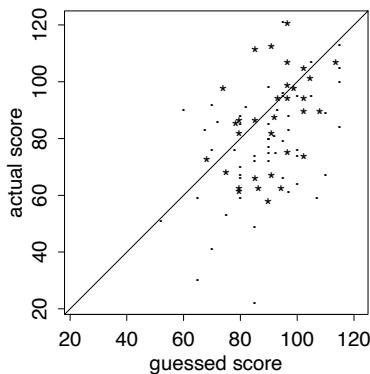


Figure 51 *Actual vs. guessed midterm exam scores for students in two terms of introductory statistics classes.* The dots represent students in the first term; the asterisks represent students in the second term, who were shown the data from the first term (Figure 50) a week before the exam. The students in the second term gave predictions that were less biased. A square scatterplot is used because the horizontal and vertical axes are on the same scale.

same question would appear on the exam. We were encouraged to find that, thus prepared, the students' guesses were less biased than those of the earlier class. Figure 51 displays the results for the unprepared class (indicated by dots, the same data as displayed in Figure 50a) and the prepared class (indicated by asterisks).

This story relates to the week's reading in that we are plotting observed vs. predicted values. It relates to the course as a whole as an example of learning from exploratory analysis of data.

## 2. Model checking for baseball analytics

Baseball analyst Bill James had this story criticizing a player evaluation that he had seen in a different book:<sup>68</sup>

*Total Baseball* has Glenn Hubbard rated as a better player than Pete Rose, Brooks Robinson, Dale Murphy, Ken Boyer, or Sandy Koufax, a conclusion which is every bit as preposterous as it seems to be at first blush.

To a large extent, this rating is caused by the failure to adjust Hubbard's fielding statistics for the ground-ball tendency of his pitching staff. Hubbard played second base or teams which had very high numbers of ground balls, as is reflected in their team assists totals. The Braves led the National League in team assists in 1985, 1986, and 1987, and were near the league lead in the other years that Hubbard was a regular. Total Baseball makes no adjustment for this, and thus concludes that Hubbard is reaching scores of baseballs every year that an average second baseman would not reach, hence that he has enormous value.”

Model checking! This idea is so fundamental to statistics—to science—and yet so many theories of statistics and theories of science have no place for it.

The alternative to the Jamesian, model-checking approach is exemplified by Pete Palmer's book, *Total Baseball*, mentioned in the above quote. Palmer did a lot of great stuff, and James is a fan of Palmer, but Palmer followed the all-too-common approach of just taking the results from his model and then . . . well, then, what can you do? You use what you've got.

What makes James special as a founder of modern sports analytics is that he's not just sifting the data; he's interested in getting it right, and he's interested in seeing where things went wrong.

<sup>68</sup>Bill James (2001), *The New Bill James Historical Baseball Abstract*, Free Press. The discussion here comes from Andrew Gelman (2016), Bill James does model checking, <https://statmodeling.stat.columbia.edu/2016/05/09/bill-james-does-model-checking/>.

A chicken is an egg’s way of making another egg.

To make the analogy explicit: the “egg” is the model and data, and the chicken is the inferences from the model. The chicken is implicit in the egg, but it needs some growing. The inferences are implicit in the model and the data, but it takes some computing.

All the effort that went into *Total Baseball* was useful for baseball analysis, in part for the direct relevance of the results (a delicious “chicken”) and in part because that book compiled so much information and made so many inferences that people such as James could come in and see which of these statements made no sense—and what this revealed about the problems with Palmer’s model.

Once you have the external counterexample—an implication that doesn’t make sense—you go find the internal flaw—the assumption in the model that went wrong, often an assumption that was so implicit in the construction of your procedure that you didn’t even realize it was an assumption at all. Or, conversely, if you first find an internal assumption that concerns you, you should follow the thread outward and figure out its external consequences: What does it imply that it does not make sense?

This story relates to the week’s readings as an example of considering the assumptions of a statistical model. It relates to the course as a whole in illustrating the connections between data, assumptions, and conclusions.

### Class-participation activities

#### 1. Sample size and statistical significance

We discuss with students the plan of fitting a regression to predict some outcome of interest from the General Social Survey, for example happiness or trust or behavior or attitude on some social or political issue. We then download the most recent General Social Survey file and then go through the codebook to look for relevant questions.<sup>69</sup> For example, we might fit a model to predict happiness given age and sex:

```
library("haven")
library("rstanarm")
gss2021 <- read_dta("2021_stata/gss2021.dta")
minidata <- gss2021[,c("marital","age","sex","happy")]
minidata$female <- minidata$sex - 1
minidata$age10 <- minidata$age/10
fit <- stan_glm(happy ~ age10 + female, data=minidata, refresh=0)
```

This recoding of the age and sex variables makes the coefficients more interpretable.

Before displaying the fitted model, we fit to a subset of the data, just the first 100 respondents:

```
fit_100 <- stan_glm(happy ~ age10 + female, data=minidata, refresh=0,
                      subset=1:100)
print(fit_100, digits=3)
```

Here is what we see:

```
observations: 85
predictors: 3
-----
              Median   MAD_SD
(Intercept) 2.09    0.23
age10        0.02    0.04
```

<sup>69</sup>See <https://gss.norc.org/> or simply search online for General Social Survey.

### 3.12. ASSUMPTIONS, DIAGNOSTICS, AND MODEL EVALUATION

151

```
female      -0.03   0.15
```

Auxiliary parameter(s):

```
Median MAD_SD  
sigma 0.65  0.05
```

The model was fit to just 85 data points because some of the respondents had missing data. Next we ask students what they expect to see when we fit the model to more data:

```
fit_1000 <- stan_glm(happy ~ age10 + female, data=minidata, refresh=0,  
subset=1:1000)  
print(fit_1000, digits=3)
```

They should expect the standard errors to decline. What about the coefficient estimates? The standard errors give a sense of how these estimates might change.

```
observations: 891  
predictors:  3  
-----  
           Median MAD_SD  
(Intercept) 2.226  0.077  
age10       -0.029  0.013  
female       0.023  0.044
```

Auxiliary parameter(s):

```
Median MAD_SD  
sigma 0.650  0.016
```

You can also do it the other way, first fitting the data to the first 2000 respondents and then considering a “replication study” on the next 1000.

This activity is relevant to the week’s readings in focusing on uncertainty in a fitted model. It relates to the course as a whole in connecting data analysis and expectations for future data.

## 2. Assumptions of regression

This activity starts by displaying Figure 52, which lists the assumptions of linear regression. Students can then discuss in pairs if they have any questions about these. After answering any questions that arise, we divide the students into seven groups and ask each group to construct an example of a regression problem violating one of these assumptions. Each group is assigned to one assumption, and they should do two things: first come up with a real-world example where the assumption is violated, then consider how to construct a simulation of this on the computer. We then go through one of these examples and code it live.

1. Validity
2. Representativeness
3. Additivity
4. Linearity
5. Independence of errors
6. Equal variance of errors
7. Normality of errors

Figure 52 *Assumptions of linear regression, in decreasing order of importance, as listed in Section 11.1 of Regression and Other Stories.*

This activity relates to the week's readings in its coverage of regression assumptions. It is relevant to the course as a whole in linking real-world concerns, statistical modeling, and simulation.

### Computer demonstrations

1. Take the difference or regress on an indicator variable

This demonstration is a bit of a review as it relates to Section 7.3 of *Regression and Other Stories*. In this example, you estimate the difference in mean tax rates in two countries (based on fake data).

```
# Create fake data
n <- 30
z <- sample(c(0,1), n, replace=TRUE, prob=c(0.5, 0.5))
y <- ifelse(z==1, rnorm(n, 30, 20), rnorm(n, 40, 20))

# Estimate difference by hand
y_1 <- y[z==1]
y_0 <- y[z==0]
diff <- mean(y_1) - mean(y_0)
se_1 <- sd(y_1) / sqrt(length(y_1))
se_0 <- sd(y_0) / sqrt(length(y_0))
se_diff <- sqrt(se_1^2 + se_0^2)
print(c(diff, se_diff))

# Estimate difference using regression on indicator variable
data <- data.frame(z, y)
fit <- stan_glm(y ~ z, data=data, refresh=0)
print(fit)
```

2. Simulate and debug

Simulate the process of experimentation and replication in the context of a hypothetical set of experiments, for example on political persuasion.

First create the underlying world. For this example, you could assume that a certain political attitude has 40% support among the general population, and then persuasion campaigns will be conducted to change opinions. Assume the true effect is 0.02 (shifting opinion, from 40% to 42% support).

You can simulate the experiment and a possible replication study.

First simulate an experiment on 2000 people: 1000 treated and 1000 controls.

```
n <- 1000
z <- rep(c(0,1), c(n,n))
y <- rbinom(n, 1, 0.40 + 0.02*z)
fake <- data.frame(y, z)
fit <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit, digits=2)
```

Here's what you'll see:

```
observations: 2000
predictors: 2
-----
              Median MAD_SD
(Intercept) 0.41    0.02
z           0.00    0.02
```

```
Auxiliary parameter(s):
  Median MAD_SD
  sigma 0.49  0.01
```

The estimate is 0, and the standard error is 0.02. Does this make sense? Sure, it's possible. The true effect is assumed to be 0.02, so this estimate is 1 standard error from the true value, which can easily happen.

Re-run the code to see what happens with a new simulation, and you'll see something like:

```
Median MAD_SD
(Intercept) 0.39  0.02
z           0.00  0.02
```

```
Auxiliary parameter(s):
  Median MAD_SD
  sigma 0.49  0.01
```

The estimated coefficient is still zero! Let's try again. Same thing. Something is going wrong. Go back in and look at the simulated data:

```
print(z)
print(y)
```

The vector  $z$  has 2000 elements—that's right—but  $y$  is only of length 1000. The problem was this line of code:

```
y <- rbinom(n, 1, 0.40 + 0.02*z)
```

Go in and fix it:

```
y <- rbinom(2*n, 1, 0.40 + 0.02*z)
fake <- data.frame(y, z)
fit <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit, digits=2)
```

Now the results make sense.

This example relates to the week's readings as an example of checking that inferences make sense. It relates to the course as a whole in demonstrating debugging of simulation code.

## Drills

### 1. Assumptions of regression and how they can fail

Consider a regression fit to a set of different countries, predicting the rate of some illegal behavior (for example, tax evasion or speeding) given country-level predictors (per-capita income, average education level, etc.). For each assumption, give an example of how it can fail:

#### (a) Validity

*Solution:* You won't know the actual rate of the illegal behavior; you only have an estimate of this outcome measure.

#### (b) Representativeness

#### (c) Additivity and linearity

#### (d) Independence of errors

#### (e) Equal variance of errors

#### (f) Normality of errors

2. How to test the assumptions of regression

For each assumption, give an example of how it can be tested:

(a) Validity

*Solution:* Taking some approximately measured quantity and measuring it more carefully. For example, voter turnout from a survey response and validated vote from public records.

(b) Representativeness

(c) Additivity and linearity

(d) Independence of errors

(e) Equal variance of errors

(f) Normality of errors

### Discussion problems

1. Assumptions of regression

Consider the implications of regression assumptions for a real-world study. The class should first pick a topic of interest that could be studied using regression—for example, this could be a problem of causal inference, or survey sampling and adjustment, or measurement and estimation of trends—and then go through the list of assumptions in Figure 52 and discuss how they apply to the example.

2. Patterns of residuals

Anna takes continuous data  $x_1$  and binary data  $x_2$  and creates fake data  $y$  from the model,  $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + \text{error}$ . She gives these data to Barb, who, not knowing how the data were constructed, fits a linear regression predicting  $y$  from  $x_1$  and  $x_2$  and makes a plot of  $y$  vs.  $x_1$ , using dots and circles to display points with  $x_2 = 0$  and  $x_2 = 1$ , respectively. The residual plot indicates to Barb that she should fit the interaction model. Sketch the residual plot that Barb could have seen when she fit the regression without the interaction.

## 3.13 Regression with linear and log transformations

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Logarithm of world population	Predictive uncertainties	Centered and standardized predictors	Log and antilog	When to use the log scale
Price elasticity of demand	Combining predictors to create a score	Regressions with logged variables	Exponential growth and decline	Straight line fit to nonlinear data

### Reading

Chapter 12 of *Regression and Other Stories*: Transformations and regression

### Pre-class warmup assignments

#### 1. Linear transformations and interactions

Consider the following model predicting earnings given height and sex:

```
Median MAD_SD
(Intercept) -9.3 15.2
height       0.4  0.2
male        -29.3 24.3
height:male  0.6  0.4
```

Auxiliary parameter(s):

```
Median MAD_SD
sigma 21.4   0.3
```

- Define the scaled variable `height_cm <- 2.54 * height`. Give the estimated coefficients from regressing earnings on `height_cm`, `male`, and their interaction.
- Define the centered and scaled variable `height_c_cm <- 2.54 * (height - 66)`. Give the estimated coefficients from regressing earnings on `c_height_c_cm`, `male`, and their interaction.
- Defined the scaled variable `earnings_dollars <- 1000 * earnings`. Give the estimated coefficients from regressing `earnings_dollars` on `height`, `male`, and their interaction.

#### 2. Correlation and regression

- Consider a regression model,  $y = a + bx + \text{error}$ , where  $x$  has standard deviation 20 and  $y$  has standard deviation 0.2. Suppose the correlation between  $x$  and  $y$  is  $-0.4$ . What can you say about  $a$  and  $b$ ?
- Consider a regression model,  $y = 0.2 + 0.3x + \text{error}$ , where  $x$  has standard deviation 2 and  $y$  has standard deviation 0.4. What is the correlation between  $x$  and  $y$ ?

### Homework assignments

#### 1. (a) Plotting linear and quadratic regressions (Exercise 12.1 of *Regression and Other Stories*)

The folder `Earnings` has data on weight (in pounds), age (in years), and other information from a sample of American adults. We create a new variable, `age10 = age/10`, and fit the following regression predicting weight:

```
    Median  MAD_SD
(Intercept) 148.7   2.2
age10       1.8     0.5
```

Auxiliary parameter(s):

```
    Median  MAD_SD
sigma     34.5   0.6
```

- i. With pen on paper, sketch a scatterplot of weights versus age (that is, weight on  $y$ -axis, age on  $x$ -axis) that is consistent with this information, also drawing the fitted regression line. Do this just given the information here and your general knowledge about adult heights and weights; do not download the data.
- ii. Next, we define  $age10\_sq = (age/10)^2$  and predict weight as a quadratic function of age:

```
    Median  MAD_SD
(Intercept) 108.0   5.7
age10       21.3   2.6
age10sq     -2.0   0.3
```

Auxiliary parameter(s):

```
    Median  MAD_SD
sigma     33.9   0.6
```

Draw this fitted curve on the graph you already sketched above.

- (b) Plotting regression with a continuous variable broken into categories (Exercise 12.2 of *Regression and Other Stories*)

Continuing Exercise 12.1 of *Regression and Other Stories*, we divide age into 4 categories and create corresponding indicator variables,  $age18\_29$ ,  $age30\_44$ ,  $age45\_64$ , and  $age65\_up$ . We then fit the following regression:

```
stan_glm(weight ~ age30_44 + age45_64 + age65_up, data=earnings)
```

```
    Median  MAD_SD
(Intercept) 147.8   1.6
age30_44TRUE  9.6   2.1
age45_64TRUE 16.6   2.3
age65_upTRUE  7.5   2.7
```

Auxiliary parameter(s):

```
    Median  MAD_SD
sigma     34.1   0.6
```

- i. Why did we not include an indicator for the youngest group,  $age18\_29$ ?
  - ii. Using the same axes and scale as in your graph for Exercise 12.1 of *Regression and Other Stories*, sketch with pen on paper the scatterplot, along with the above regression function, which will be discontinuous.
2. (a) *In pairs*: Working through your own example  
Summarize what you have learned about your example from all the analysis you have done during the semester.

## Stories

### 1. Logarithm of world population

We tell this story to students in stages: first, presenting the time series of world population since the year 0 in tabular form (the first two columns of Figure 53, which we write on the board before

### 3.13. REGRESSION WITH LINEAR AND LOG TRANSFORMATIONS

157

Year	Population	log (population)	Residual	exp(residual)
1	170 million	18.95	0.67	1.96
400	190	19.06	0.11	1.12
800	220	19.21	-0.42	0.66
1200	360	19.70	-0.60	0.55
1600	545	20.12	-0.86	0.42
1800	900	20.62	-0.69	0.50
1850	1200	20.91	-0.49	0.61
1900	1625	21.21	-0.27	0.76
1950	2500	21.64	0.08	1.08
1975	3900	22.08	0.48	1.61
2000	6080	22.53	0.88	2.41
2020	7795	22.78	1.09	2.99

Figure 53 Time series of estimated world population and residuals from a logarithmic fit, which reveal that the population was growing even faster than exponentially. After these data are graphed (see Figure 54), students should try to guess the population in the year 1400; it was 350 million.

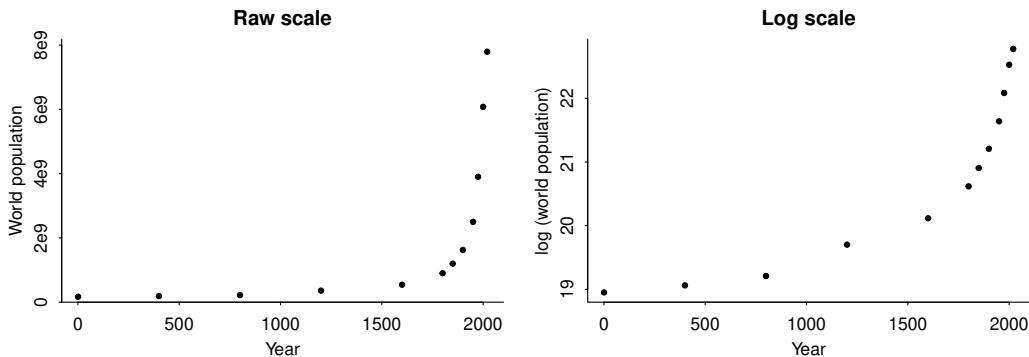


Figure 54 World population over time, graphed on the original and logarithmic scales. This is a subtle example since the growth is faster than linear, even on the logarithmic scale. We project the log-scale graph on the board where we then later draw the fitted regression line,  $y = 18.3 + 1.7 * \text{year}/1000$ . The data in the original and logarithmic scales, and the residuals from the regression, are given in Figure 53.

the class begins), then graphing the raw data (Figure 54a) and on the logarithmic scale (Figure 54b). On the raw scale, all that can be clearly seen is that the population has increased fast in recent centuries. On the log scale, it is clear that the rate of increase itself increased.<sup>70</sup>

Next we ask students to guess the population in the year 1400. After collecting some guesses, we reveal that the world population in that year was 350 million—actually lower than the year 1200 population, because of plague and other factors—just to illustrate that even interpolation can sometimes go awry.

Another possible discussion topic is the source of the population numbers: for example, how did they estimate the population of the world for the year 1?

A good way to understand the logarithmic model is with the regression line:

```
fit <- stan_glm(log_pop ~ year, data=population, refresh=0)
print(fit)
```

This yields,

<sup>70</sup>From Sections 3.8.2 and 5.1.3 of *Teaching Statistics: A Bag of Tricks*.

```
Median  MAD_SD
(Intercept) 18.3    0.5
year         0.0     0.0
```

```
Auxiliary parameter(s):
Median  MAD_SD
sigma   0.7     0.2
```

You can interpret each number in the output: the model predicts a log population of 18.3 in year 0 (which actually is in the range of the data this time, unlike when fitting models to data from 1900 to 2000), an increase in log population of 0.0 per year, and a residual standard deviation of 0.7 on the log scale . . . Wait! What about that coefficient of 0.0?

You can include more digits in the print output:

```
Median  MAD_SD
(Intercept) 18.276  0.489
year        0.002   0.000
```

```
Auxiliary parameter(s):
Median  MAD_SD
sigma   0.728   0.163
```

But really the right thing to do is rescale:

```
population$year_1000 <- population$year/1000
fit_2 <- stan_glm(log_pop ~ year_1000, data=population, refresh=0)
print(fit_2)
print(fit)
```

And we get this:

```
Median  MAD_SD
(Intercept) 18.3    0.5
year_1000   1.7     0.3
```

```
Auxiliary parameter(s):
Median  MAD_SD
sigma   0.7     0.2
```

We draw this fitted line on the board where the log-scale data (Figure 54b) have been plotted. The slope is 1.7: according to the line, every 1000 years the log population increases by 1.7, or every year the log population increases by  $1.7 \times 10^{-3}$ , an increase of a factor of 0.17%. The residual standard deviation implies that you can expect the predictions to be off by about 0.7, and a range of  $\pm 0.7$  on the log scale corresponds to a range of  $[\exp(-0.7), \exp(0.7)] = [0.5, 2.0]$  on the scale of population. So the data are typically within a factor of 2 of the model's prediction.

## 2. Price elasticity of demand

How much does the demand for a product go up or down if its price is changed? For most items, we would expect an increase in price to lead to a decrease in demand: this is called a negative *elasticity*. More generally, the elasticity is defined as the derivative of logarithm of demand with respect to logarithm of price or as the percentage change in demand resulting from a 1% increase in price. The definition in terms of a 1% change is only approximate because  $\log(1.01)$  is not exactly 0.01.

To say this more slowly: consider the model,

$$\log y = a + b \log x,$$

### 3.13. REGRESSION WITH LINEAR AND LOG TRANSFORMATIONS

159

where  $x$  is price and  $y$  is demand. Here  $b$  is the elasticity, and we if we exponentiate both sides of the above model we get  $y = Ax^b$ . A price increase of 1% corresponds to multiplying  $x$  by 1.01, hence multiplying  $y$  by  $1.01^b$  or, equivalently, to adding  $b \log(1.01)$  to  $\log y$ .

For example, suppose the price elasticity of demand for a particular product is  $-0.8$ . Then a 1% increase in price would result in demand changing by a factor of  $1.01^{-0.8} = 0.992$  (more precisely, 0.09921, but we could never measure any of these quantities to that level of accuracy), hence a decrease in demand of 0.8%. From the other direction, a 1% *decrease* in price would result in demand changing by a factor of  $1.01^{0.8} = 1.008$ , an increase of 0.8%.

On the logarithmic scale, the corresponding calculation is that increasing  $\log x$  by 0.01 would result in  $\log y$  decreasing by 0.008, and decreasing  $\log x$  by 0.01 would lead to  $\log y$  increasing by 0.008.

We can better understand the concept of elasticity by considering various possible values:

- Suppose the elasticity  $b$  in the above expression equals 0. Then demand is completely inelastic: changes in price do not affect demand for the product. A stylized example here is refrigerators in wealthy countries. Everyone wants exactly one refrigerator. If the price goes up, you'll still want a refrigerator, and if the price goes down, you won't suddenly want a second fridge. This is a simplification—if the price of refrigerators changes enough, people will indeed consider owning zero or two of them—but it gives a sense of what it takes for demand to be inelastic.
- Suppose the elasticity is  $-1$ . Then demand decreases as price increases (and vice-versa) just enough so that the product, demand \* price, is a constant—for example, demand decreasing by 1% when price increases by 1%. This implies that the dollar value of the demand does not change with price.
- Suppose the elasticity is  $-1.5$ . Then demand decreases *faster* than the price increases: for example, an increase in price of 1% leads to a decrease of 1.5% in demand.
- What about a positive elasticity? For example, suppose elasticity is  $+0.5$ . Then a 1% increase in price leads, paradoxically, to *more* demand. Such behavior can be imagined for some status-related products where exclusivity itself represents part of the appeal, but it is not something we would usually expect to see.

With all this in mind, we can consider how elasticity could influence a pricing decision. Suppose you are selling a product with elasticity of zero. What should you do? You should raise its price! The demand won't change, so you should be able to sell the same amount, and you'll just make more money.

What if you are selling a product with an elasticity of  $-0.5$ ? Then if you raise the price by 1%, the demand will drop by only 0.5%, and you'll make more money. More generally, whenever the elasticity is larger than  $-1$ , it makes sense to raise the price. From the other direction, if the elasticity of a product you are selling is more negative than  $-1$ , it makes sense to *lower* your price, as the demand will increase by a higher percentage than your price is dropping, so demand \* price will increase. The elasticity will change as you move along the demand curve, so at some point there will be diminishing returns from changing the price. Here we are just considering elasticity in an instantaneous sense.

These statements are idealizations: in the real world, “demand” does not equate to sales, changing the price can involve costs of its own, there will be competing vendors who can alter their prices as well, and price elasticity of demand is itself variable and not always so easy to estimate. The point of this story is not to get into all these qualifications but rather to demonstrate how a linear model on the log-log scale can be interpreted.

This example relates to the week's reading as an example of logarithmic transformations for linear

models. It relates to the course as a whole as an example of interpreting a coefficient in such a model. Unlike most of our stories, this one does not directly involve data or model fitting; instead we are focusing on the deterministic part of the model in order to better understand how it works when both  $x$  and  $y$  are log transformed.

### Class-participation activities

#### 1. Predictive uncertainties

Suppose that, using a survey with 500 respondents, a regression is fit predicting a “feeling thermometer” response on some celebrity on a 0–100 scale,<sup>71</sup> given a party identification predictor (on a -3 to 3 scale). Just describe this problem; do not collect, download, or simulate data. Suppose the data were to exist, the regression model were fit, and `posterior_linpred` and `posterior_predict` were performed for someone who is strongly Republican ( $x = 3$ ). What are plausible values for the point prediction and predictive uncertainties?

Students should work in pairs to come up with reasonable values here, and then the instructor can write the different guesses on the board as a starting point for class discussion. This relates to the week’s reading in that to understand transformations you need to understand the scale of a problem: you need a sense of whether a coefficient is expected to be of order 100 or 10 or 1 or 0.1 or 0.01 or whatever. The activity relates to the course as a whole as an example of thinking through an understanding a model quantitatively.

#### 2. Combining predictors to create a total score

Though class discussion, the instructor creates a Google form with several questions on some topic of interest. For example, if the goal is to measure left-right ideology, the questions can include party identification, political ideology, and a few issue attitudes. Put these together to create a combined score. When doing this you have to be aware of the scale of the different measurements. To make the problem interesting, include measurements on different scales, for example an ideology scale from -3 to 3 and a 0–100 feeling-thermometer scale.

This activity relates to the week’s reading as an example of a custom transformation; it relates to the course as a whole by connecting ideas of measurement and modeling.

### Computer demonstrations

#### 1. Centered and standardized predictors

Here is a series of regressions fit to the same data with different linearly-transformed predictors.<sup>72</sup> For each, type in the code and consider what the fitted regression will be, then hit Return and look at the output. Keep doing this, changing the model in various ways, until you can successfully predict what the results will be.

```
# Setup
library("rstanarm")
earnings <- read.csv(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Earnings/data/",
  "earnings.csv"))
```

<sup>71</sup>For example, the American National Election Study asks, “I’d like to get your feelings toward some of our political leaders and other people who are in the news these days. I’ll read the name of a person and I’d like you to rate that person using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don’t feel favorable toward the person and that you don’t care too much for that person. You would rate the person at the 50 degree mark if you don’t feel particularly warm or cold toward the person.”

<sup>72</sup>From Section 12.2 of *Regression and Other Stories*.

```
)  
earnings <- na.omit(earnings[, c("earn", "height", "male", "education")])  
  
# Fit regression  
fit <- stan_glm(earn ~ height + male + education, data=earnings, refresh=0)  
print(fit)  
  
# Re-run regression with centered predictors  
earnings$height_c <- earnings$height - mean(earnings$height)  
earnings$male_c <- earnings$male - mean(earnings$male)  
earnings$education_c <- earnings$education - mean(earnings$education)  
fit_2 <- stan_glm(earn ~ height_c + male_c + education_c,  
    data=earnings, refresh=0)  
print(fit_2)  
  
# Re-run regression with standardized predictors (except sex)  
earnings$z_height <- earnings$height_c / (2 * sd(earnings$height))  
earnings$z_education <- earnings$education_c / (2 * sd(earnings$education))  
fit_3 <- stan_glm(earn ~ z_height + male_c + z_education,  
    data=earnings, refresh=0)  
print(fit_3)
```

## 2. Regressions with logged variables

Take the familiar dataset `earnings`. Regress log earnings on height, and plot the model on both the log and the linear scale. In addition, you could run, plot, and interpret the log-log model.<sup>73</sup>

```
# Setup  
library("rstanarm")  
earnings <- read.csv(paste0(  
    "https://raw.githubusercontent.com/avehtari/",  
    "ROS-Examples/master/Earnings/data/",  
    "earnings.csv"))  
earnings <- earnings[earnings$earn > 0, ]  
height_jitter_add <- runif(nrow(earnings), -0.4, 0.4)  
  
# Regression of log earnings on height  
logmodel <- stan_glm(log(earn) ~ height, data=earnings, refresh=0)  
print(logmodel)  
  
# Plot model on log scale  
plot(earnings$height + height_jitter_add, log(earnings$earn),  
    xlab="height", ylab="log (earnings)", pch=20, col="gray", bty="l", cex=.3)  
sims <- as.matrix(logmodel)  
curve(coef(logmodel)[1] + coef(logmodel)[2]*x, add=TRUE)  
  
# Plot model on linear scale  
plot(earnings$height + height_jitter_add, earnings$earn, xlab="height",  
    ylab="earnings", pch=20, col="gray", bty="l", cex=.3, ylim=c(0,250e3))  
curve(exp(coef(logmodel)[1] + coef(logmodel)[2]*x), add=TRUE)
```

## Drills

### 1. Log and antilog

<sup>73</sup>From Section 12.4 of *Regression and Other Stories*.

Follow the instructions to either log or exponentiate the following expressions:

(a) Log:  $y = e^a e^{3b}$

*Solution:*  $\log y = a + 3b$

(b) Log:  $y = e^{a+3b}$

(c) Log:  $y = e^{0.5-3x+\text{error}}$

(d) Log:  $y = 0.3 e^{2x} e^{\text{error}}$

(e) Log:  $y = a * 1.01^t$

(f) Log:  $y = 700 x^p$

(g) Antilog:  $\log(y) = a + b \log x$

(h) Antilog:  $\log(y) = a + x \log b$

(i) Antilog:  $\log(y) = 4.0 + 17.2 x_1 + 3 x_2 + \text{error}$

## 2. Examples of exponential growth and decline

(a) How do you express  $y = 15 * 3^x$  on the log scale?

*Solution:*  $\log y = \log(15) + \log(3) * x$

(b) How do you express  $y = a * b^x$  on the log scale?

(c) How do you express  $\log(z) = \log(10) + \log(0.7) * x$  on the linear scale?

(d) If something increases every year by 4%, how can you express this process on the linear scale?  
How about on the log scale?

(e) Describe in words the process  $y = 1000 * 1.03^x$ .

(f) Describe in words the process  $\log(u) = \log(2) + \log(3) * t$ .

## Discussion problems

### 1. General rules of when to use the log scale

Would Figure 2.4 in *Regression and Other Stories* be improved by plotting either or both axes on the log scale? Can you come up with general rules of when to use the log scale in graphing and regression? To get started, consider the example of a country-level regression predicting carbon emissions from GDP, population, and land area; or predicting carbon emissions per capita from GDP per capita and land area per capita; or predicting  $\log(\text{carbon emissions})$  from  $\log(\text{GDP})$ ,  $\log(\text{population})$ , and  $\log(\text{land area})$ ; or predicting  $\log(\text{carbon emissions per capita})$  from  $\log(\text{GDP per capita})$  and  $\log(\text{land area per capita})$ .

### 2. Straight line fit to an exponential or power-law pattern

Consider one of the examples from the class discussion of exponential growth or decline, or power-law growth or decline. What would happen if you try to fit a linear model to data from such an example? Sketch data showing the exponential or power-law pattern along with your best guess of the straight line that would result from fitting a linear regression. Then use R to simulate some points, graph them, fit the regression, and graph the fitted line. Where does the line fit well and where does it have problems?

## Chapter 4

# Week by week: the second semester

We continue the outline of the course, now going through the 13 weeks of the second semester. Again, we have two 75-minute classes a week, hence the description for each week includes two sets of readings and homework, two stories, two class-participation activities, two computer demonstrations, and so forth.

### 4.14 Review of basic statistics and regression modeling

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Biased samples and coverage of intervals	Self-selected treatment assignment	Causal inference adjusting for pre-treatment	Regression coefficients as comparisons	Sampling and adjustment
The problem of too much talent?	Design a study to explore nonlinearity	Simulating patterns of bias	Log transformations	Causal inference, adjustment

#### Reading

1. Chapters 1–12 and Appendix A of *Regression and Other Stories* (review)
2. Appendix B of *Regression and Other Stories*: 10 quick tips to improve your regression modeling

#### Pre-class warmup assignments

1. *No assignment before the first class.*
2. Estimates and standard errors
  - (a) A survey is conducted of 700 people to estimate the proportion who support a law banning abortion. Assuming the respondents are a simple random sample from the population, what is the standard error of the estimated proportion?
  - (b) A survey is conducted including 500 Republicans, 600 Democrats, and 700 Independents to ask their opinions on a proposed tax plan. Assuming the respondents are a simple random sample from the population, what is the standard error of the estimated difference in the proportions of Republicans and Democrats who support this plan?
  - (c) A company performs an experiment in which 50 work teams are given a special training, and these are compared to 100 teams that are not given this training. Each team's efficiency is measured before and after the experiment, and then a regression is performed predicting post-test from pre-test and a treatment indicator. The estimated treatment effect is 0.35 with a

standard error of 0.30. Suppose the experiment were replicated with twice the sample size. What would you expect the standard error to be?

### Homework assignments

1. *No homework assignment due for the first class.*
2. Reading and commenting on a published article

The instructor should choose a research paper on a topic of interest to students. The assignment is then to read the article and discuss the following issues:

- (a) To what extent is the research generalizing from sample to population, from treatment to control group, and from observed data to underlying quantities of interest?
- (b) How are the data and inferences displayed graphically? How could the display be improved?
- (c) What statistical methods are used in the paper? How do the methods and statistical conclusions line up with the substantive goals and conclusions that are reported?
- (d) What questions do you have about the paper? How do you think it could be improved?

### Stories

1. Biased samples and coverage of confidence intervals

Real-world surveys have bias. Samples are not completely representative of the populations they are studying, nor can a sample truly be considered to be a random draw of balls from an urn.

For example, a few years ago we estimated nonsampling error in political polls by collecting a large number of state polls taken in the final weeks of presidential, senatorial, and gubernatorial elections and comparing the published poll estimates with the actual election outcomes.<sup>1</sup> We found that “average survey error as measured by root mean square error is approximately 3.5 percentage points, about twice as large as that implied by most reported margins of error.”

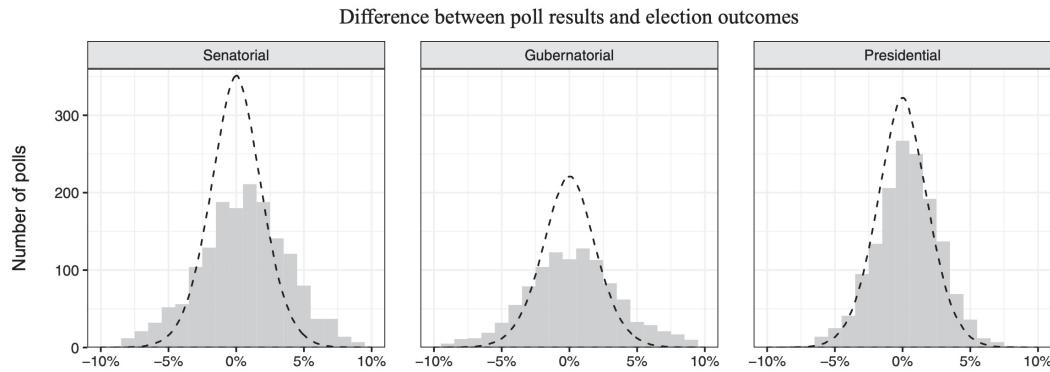
Figure 55 shows the results, separately displaying polling error for elections for senator, governor, and president. For each graph, the histogram displays the distribution of polling errors in the data: all are within 10% of the election outcome and the vast majority are within 5%, but that is not so impressive given that the standard deviation of a proportion from a simple random sample of size 1000 is approximately  $\sqrt{0.5^2/1000} = 0.016$ , or 1.6%. Each plot also shows a dotted line indicating the variation that would be expected if the polling were ideal. For each survey, we took the estimated Republican share of the two-party support and then computed the simple sampling error using the formula  $\sqrt{p(1-p)/n}$ . For each plot, the dotted line shows the mixture of the normal distributions for all those surveys.

Real-world polls typically include some adjustments for nonrandom sampling and nonresponse, with different approaches used by different survey organizations.<sup>2</sup> Reported standard errors can be higher than  $\sqrt{p(1-p)/n}$ . Even after these published adjustments, though, polling errors are a bit more variable than would be expected from published uncertainties. This makes sense; three sources of nonsampling error are differential nonresponse (Democrats or Republicans being more or less likely to answer pollsters during a particular election campaign), uncertainty about voter turnout, and shifts in opinion between the poll and election day.

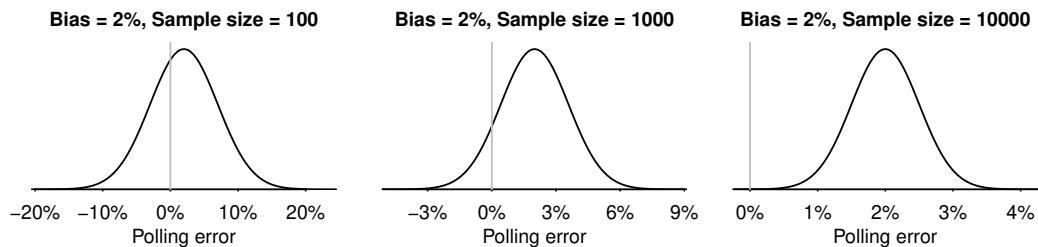
Just about all real-world surveys have bias. The larger the sample size, the more *relatively* important the bias is. Consider a set of surveys, all of which are estimating a proportion that is

<sup>1</sup>Houshmand Shirani-Mehr, David Rothschild, Sharad Goel, and Andrew Gelman (2018), Disentangling bias and variance in election polls, *Journal of the American Statistical Association* 113, 607–614.

<sup>2</sup>D. Stephen Voss, Andrew Gelman, and Gary King (1995), Pre-election survey methodology: Details from nine polling organizations, 1988 and 1992, *Public Opinion Quarterly* 59, 98–132.



**Figure 55** Distribution of polling errors (Republican share of two-party support in the poll minus Republican share of the two-party vote in the election) for state-level presidential, senatorial, and gubernatorial election polls between 1998 and 2014. Positive values indicate the Republican candidate received more support in the poll than in the election. For comparison, the dashed lines show the theoretical distribution of polling errors assuming each poll is generated via simple random sampling.



**Figure 56** Distribution of survey errors for three different hypothetical scenarios, each with bias of 2 percentage points and with sample size 100, 1000, or 10 000. The three graphs are on different scales. As sample size increases, the range of possible errors narrows, but the bias does not go away.

near 50% with a bias of 2 percentage points. Figure 56 shows sampling distributions for surveys of size  $n = 100, 1000$ , and  $10\,000$ . Multiplying the sample size by 10 has the effect of dividing the standard error by  $\sqrt{10} \approx 3.1$ , and the three sample sizes correspond to standard errors of 0.05, 0.016, and 0.005.

The three graphs in Figure 56 are on different scales, and they show how a fixed bias becomes more consistent as the sampling error decreases.

Another way to see the importance of bias is to consider coverage of confidence intervals. For each of the three hypothetical surveys in Figure 56, we simulated a large number of replications. For each replication, we computed the 95% interval—the estimate  $\pm 2$  standard errors. And then we checked the coverage of these intervals: the percentage of 95% intervals that contained the true population proportion being estimated. The coverage of these 95% intervals is 94% for the survey with sample size 100, 77% for the survey with sample size 1000, and only 2% for the survey with sample size 10000. The problem is that with the large survey the standard errors are very small, yielding confidence intervals that are way too narrow.

Figure 57 shows confidence intervals resulting from 100 simulated datasets from each of the three survey designs. For each graph, the gray vertical line indicates the true parameter value. For the

<sup>3</sup>Valerie Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman (2021), Unrepresentative big surveys significantly overestimated US vaccine uptake, *Nature* 600, 695–700, with discussion at Andrew Gelman (2021), “Unrepresentative big surveys significantly overestimated US vaccine uptake,” and the problem of generalizing from sample to population, <https://statmodeling.stat.columbia.edu/2021/12/26/unrepresentative-big-surveys/>.

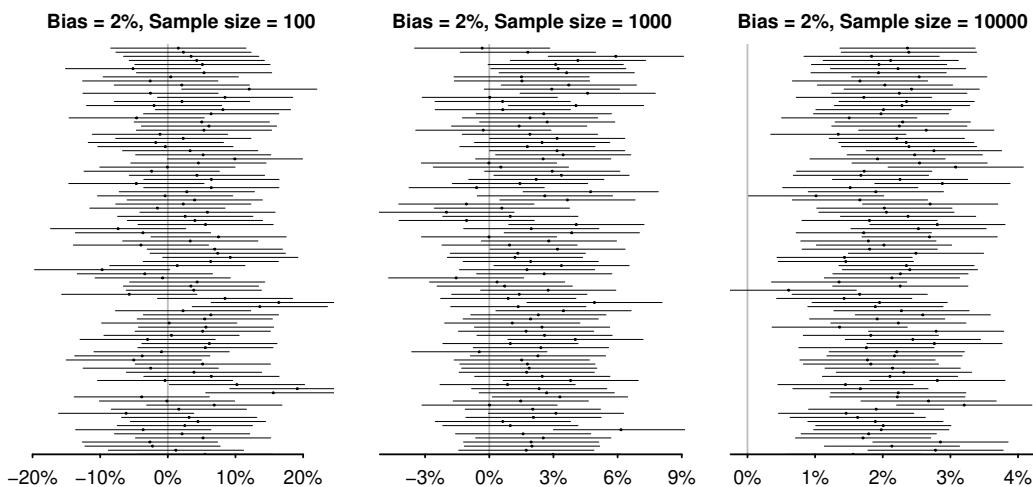


Figure 57 100 simulated 95% intervals relative to the true population proportion (indicated by a vertical line), for each of three hypothetical sampling designs. The three graphs are on different scales. The larger surveys give smaller standard errors and thus narrower confidence intervals.

larger surveys, confidence intervals are narrower (note the different scales on the three graphs), and the result is that fewer of the nominal 95% intervals contain the true value.

For another example, a recent study reported the following:<sup>3</sup>

“Two large online surveys of U.S. adults consistently overestimated first-dose COVID-19 vaccine uptake relative to counts released by the CDC throughout the spring of 2021. By May 2021, Delphi-Facebook’s COVID-19 Trends and Impact Survey (collecting about 250,000 responses per week) and the Census Bureau’s Household Pulse Survey (about 75,000 responses every 2 weeks) overestimated first dose uptake by 17 and 14 percentage points, respectively, relative to the CDC’s historically-adjusted estimates released May 26, 2021. These errors are orders of magnitude larger than the uncertainty intervals reported by each survey, which are minuscule due to the large sample sizes. Meanwhile, a more traditional survey run by Axios-Ipsos, with only about 1000 responses per week, provides more reliable estimates with reasonable uncertainty.”

This story relates to the week’s readings as an example of the real-world challenges of statistical inference. It relates to the course as a whole in connecting data collection to data analysis.

## 2. The “problem of too much talent”?

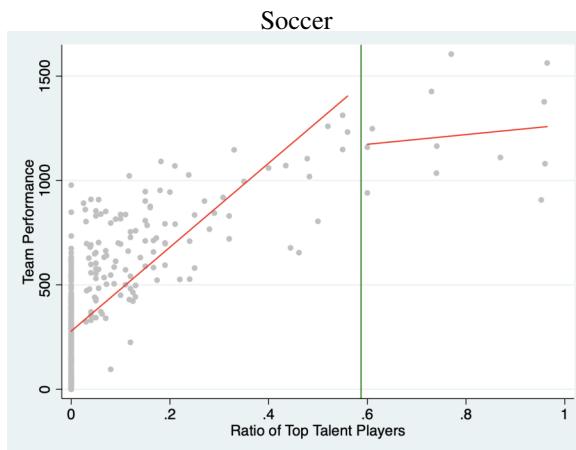
In 2014, an article was published in a psychology journal claiming that, in the sports of soccer and basketball, “talent facilitates performance—but only up to a point, after which the benefits of more talent decrease and eventually become detrimental as intrateam coordination suffers.”<sup>4</sup> This was summarized in a news report as, “Basketball and soccer teams with the greatest proportion of elite athletes performed worse than those with more moderate proportions of top level players.”<sup>5</sup>

<sup>4</sup>Roderick Swaab, Michael Schaefer, Eric Anicich, Richard Ronay, and Adam Galinsky (2014), The too-much-talent effect: Team interdependence determines when more talent is too much or not enough, *Psychological Science* 25, 1581–1591. The discussion here is taken from Andrew Gelman (2016), Should this paper in Psychological Science be retracted? The data do not conclusively demonstrate the claim, nor do they provide strong evidence in favor. The data are, however, consistent with the claim (as well as being consistent with no effect), <https://statmodeling.stat.columbia.edu/2016/06/28/khkhkj/>.

<sup>5</sup>Cindi May (2014), The surprising problem of too much talent: A new finding from sports could have implications in business and elsewhere, *Scientific American*, <https://www.scientificamerican.com/article/the-surprising-problem-of-too-much-talent/>.

#### 4.14. REVIEW OF BASIC STATISTICS AND REGRESSION MODELING

#### 167



**Figure 58** Data, summarizing the relationship between average performance of soccer teams and the percentage of elite players on the team, that were used to support the statement, “teams with the greatest proportion of elite athletes performed worse,” a claim that arose from a misinterpretation of a quadratic curve fit to these points.

In this study, “Top talent was indexed by the percentage of players within each national team who had contracts with one of the world’s elite club teams.”

This finding could indeed have implications elsewhere, but first we should check exactly what the authors are claiming. The paper displays a curve, based on data from several years of international soccer competitions, of average team performance as a function of percentage of top talent, showing a quadratic relationship, with team performance as  $x$  increases from 0 to 60%, then being flat until about 80%, then declining slightly as the percentage of top talent increases to 100%. The instructor can find this graph as Figure 2 in the published article and project it onto the screen.<sup>6</sup> But is it really true that “teams with the greatest proportion of elite athletes performed worse”? As one observer asked, “what soccer teams are comprised entirely of elite players?” Table 1 of the linked paper reports that the mean percentage of top talent was 7% and the standard deviation was 16%, so there can’t be many teams in the 70–100% range where the curve is in its declining phase.

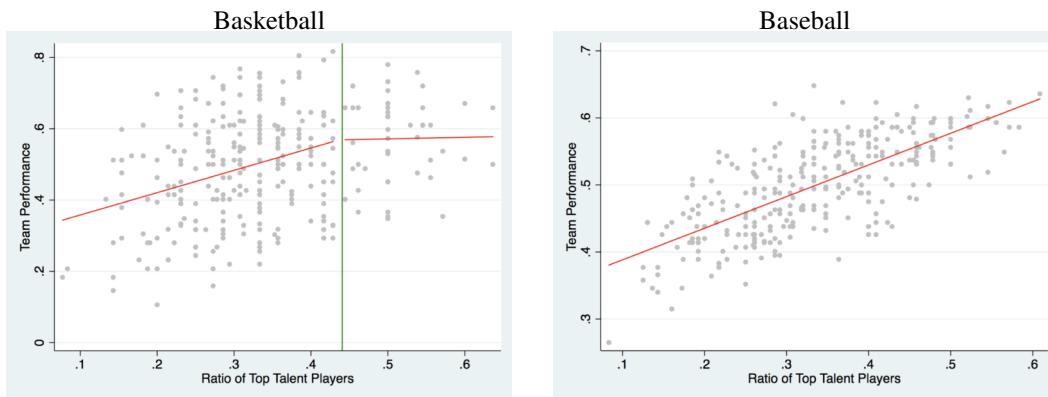
We have discussed the fitted curve. But what about the original data? In a later correspondence, the authors of the paper prepared some graphs showing the data underlying their quadratic fit.<sup>7</sup> Figure 58 shows these data, along with a piecewise linear function estimated by the authors. Ignoring all the fitted lines, the data are consistent with a monotonic pattern—more talent is always better, on average—but with diminishing returns. There are not enough teams on the right half of the graph to say more than that. Also, the fitted broken-line regression does not seem helpful; there’s no evidence of any sharp break at 0.58 or anywhere else in the range of the data. The data do not support the claim in the paper’s abstract that, after some point, having more talent “eventually become[s] detrimental.”

What happened? Three things. First, when you fit a quadratic curve to a function that asymptotes, you can get this artifact where it appears that the function decreases. Second, there’s very little information in the data about the high end of the curve, so any apparent pattern there can come just from noise. Third, we suspect the authors of this paper and the editors of the journal where it was published fell in love with their hypotheses.

They also looked at data from soccer and baseball, and again their storytelling went beyond what

<sup>6</sup>[https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6157&context=lkcsb\\_research](https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6157&context=lkcsb_research).

<sup>7</sup>Roderick Swaab, Michael Schaefer, Eric Anicich, Richard Ronay, and Adam Galinsky, Response to post, “Thirty-somethings are shrinking and other challenges for U-shaped inferences,” <http://datacolada.org/wp-content/uploads/2014/09/AuthorsResponse3.pdf>.



**Figure 59** Data that were used in fitting a model suggesting a decline in performance for top-talent teams in basketball but not baseball. As with Figure 58, we do not think the data support this claim.

could be learned from their data. Figure 59 shows the data for those two sports, which again are consistent with diminishing returns but no absolute decline in performance for the most talented teams. In the article, though, the authors claimed to find a “too-much-talent effect” in basketball but not baseball, because their fitted quadratic curves had a decline for one sport and not the other; see Figures 4 and 6 in the linked article. Again, though, the data do not support this conclusion.

This story is relevant to the week’s readings as it is an example of bad interpretation of a regression model that does not fit the data. In this case, a quadratic regression was fit to data showing diminishing returns, leading to a mistaken conclusion that the data showed a non-monotonic pattern. The story is relevant to the course as a whole in pointing toward the value of flexible nonlinear models, at topic that we discuss in Chapter 22 of *Regression and Other Stories*.

### Class-participation activities

#### 1. Experiment with self-selected treatment assignment

We begin the second semester’s activities with an in-class experiment that demonstrates the challenge of causal inference, a topic that will occupy much of the later part of the course. We start by giving students the link to a Google form, which has questions shown in Figure 60, and we give them a couple minutes to fill it out.

We then display the questions on the screen and explain that the goal of this experiment is to estimate how people react differently to the scenario, depending on whether it involves an athlete or a theater student. The simple thing to do is compare the average responses of students who received the theater vignette to the average responses to the sports vignette. But there is a potential bias here because the participants in the experiment (that is, the students in the class) selected their own treatments.

We ask students to use hypothetical scatterplots to sketch possible patterns of bias. Each scatterplot should show one dot per student, plotting the outcome (the response to the final survey question, on a 0 to 100 scale) vs. a measure of relative interest in sports and theater (the difference between the “How interested are you in sports” and “How interested are you in theater” questions, thus on a -4 to 4 scale), using two different colors corresponding to the two treatments.

We then discuss how to analyze the data to estimate the average treatment effect. Start with a regression of the outcome on the treatment indicator, coded as 1 for the sports vignette or 0 for the theater vignette. Then we add the “sports minus theater” predictor. (Why not “theater minus sports”? Because it’s easier to interpret the model if we use a consistent coding, in this case

- What is your age?
- What is your sex (male / female / other or prefer not to answer)?
- How interested are you in sports (on a 1–5 scale, with 1 = not at all and 5 = very much)?
- How interested are you in theater (on a 1–5 scale, with 1 = not at all and 5 = very much)?
- How interested are you in cooking (on a 1–5 scale, with 1 = not at all and 5 = very much)?
- How interested are you in politics (on a 1–5 scale, with 1 = not at all and 5 = very much)?

You will next read a vignette. Would you like it to be about sports or theater?

*Respondent receives one of two vignettes*

:

*Sports vignette:*

William is on the varsity soccer team. The night before an important final exam, the coach calls up to remind him of an upcoming practice. The next day, William goes to the practice and misses the exam without notifying the instructor. With a zero on the final exam, William would fail the class. The instructor allows him to take a makeup exam but will only give partial credit. If you were the instructor, how much credit (between 0 and 100%) would you give for the makeup exam?

*Theater vignette:*

William is in the university theater program. The night before an important final exam, the director calls up to remind him of an upcoming rehearsal. The next day, William goes to the rehearsal and misses the exam without notifying the instructor. With a zero on the final exam, William would fail the class. The instructor allows him to take a makeup exam but will only give partial credit. If you were the instructor, how much credit (between 0 and 100%) would you give for the makeup exam?

Figure 60 *Questions for a survey experiment designed to estimate differences in attitude in a vignette involving an athlete or theater student. Responses to the vignette question are compared, but the treatment is self-selected, so we need to adjust for differences between the two groups.*

with theater being the baseline and sports being positive.) Adding this predictor allows for the possibility that sports fans are more or less lenient than theater fans.

Then we need to add the interaction of the treatment indicator and the “sports minus theater” predictor. A positive coefficient for this interaction implies that sports fans are more lenient in the sports condition, and theater fans are more lenient in the theater condition.

All the above steps are necessary to adjust for obvious potential biases. We can include the responses to the other questions as predictors as well.

This activity relates to the week’s readings as an example of using regression to adjust for differences between treatment and control groups, and using an interaction to model varying treatment effects. It relates to the semester as a whole by giving a challenging problem of causal inference.

## 2. Design a study to explore a nonlinear relation

We remind students of the failed study that purported to show a negative effect of having too much top talent in a sports team; see Figure 58. Suppose you wanted to do this right; how would you design such a study?

We first separate the question into two parts: causal and descriptive. The causal question: if you add (or subtract) top-talent players from your team, what would you expect to happen to the team’s performance? The descriptive question is: How much better or worse are teams with more top talent, on average? We can briefly discuss the causal challenge, but the main purpose of this

activity is the descriptive problem: how to redo that “too much talent” study to better address the questions it was purporting to answer.

Issues to discuss here include measurement (how to define “top talent”), performance (use score differential rather than win-loss record) and, most importantly, getting enough data of top-talent teams to estimate a potential decline in performance at the high end. After considering data collection, we can discuss analysis: what is a flexible family of curves that would allow possible non-monotonic effects without that being driven by diminishing returns? The curve  $y = a * (1 - e^{-bx})$  asymptotes but has no decline at all, so it seems that we might want to include a sum of such functions. It’s not so clear how best to do this. Another option would be piecewise constant. The broken lines in Figure 58b don’t look right.

This activity relates to the week’s readings as an example of the connections between data collection, modeling, and inference. It relates to the course as a whole by introducing nonlinear modeling.

### Discuss reading and homework

1. For the first day of class, the instructor should discuss the course plan (see Section 1.2):
  - (a) Goals of the course
  - (b) Components of the course; discuss Figure 61
  - (c) Structure of each class period; discuss Figure 62
  - (d) Students’ responsibilities
  - (e) Roles of mathematics, computing, and applications
  - (f) Also discuss questions from final exam

1. Reading
2. Homework
3. Feedback sheet
4. Classes
5. Final exam

Figure 61 *Components of the course. The instructor should project these onto the screen or write them on the board on the first day of class.*

1. Story
2. Activity
3. Discuss reading and homework
4. Computer demo
5. Drill
6. Discussion problem

Figure 62 *Structure of each class period. The instructor should project these onto the screen or write them on the board on the first day of class.*

2. For later classes, this is a time to discuss issues that students raised in the shared document for that class; see Section 1.3.

### Computer demonstrations

#### 1. Causal inference using regression adjusting for pre-treatment variables

We analyze the data from the sports-theater survey:

```
library("rstanarm")
survey <- read.csv("Sports and theater survey.csv")
colnames(survey) <- c("time", "age", "sex", "interest_sports",
  "interest_theater", "interest_cooking", "interest_politics",
  "topic", "response_sports", "response_theater")
print(survey)
survey$male <- ifelse(survey$sex=="Male", 1,
  ifelse(survey$sex=="Female", 0, 0.5))
survey$choose_sports <- ifelse(survey$topic=="Sports", 1, 0)
survey$response <- ifelse(survey$topic=="Sports",
  survey$response_sports, survey$response_theater)
print(survey)

fit_1 <- stan_glm(response ~ choose_sports, data=survey, refresh=0)
print(fit_1)

survey$diff <- survey$interest_sports - survey$interest_theater
plot(survey$diff, survey$response, col=ifelse(survey$choose_sports==1,
  "red", "blue"))

fit_2 <- stan_glm(response ~ choose_sports + diff, data=survey, refresh=0)
print(fit_2)
fit_3 <- stan_glm(response ~ choose_sports + diff + choose_sports:diff,
  data=survey, refresh=0)
print(fit_3)
fit_4 <- stan_glm(response ~ choose_sports + diff + choose_sports:diff +
  interest_cooking + interest_politics + male, data=survey, refresh=0)
print(fit_4)
```

Finally we can estimate an average treatment effect for the students in the class:

```
coef(fit_3)[["choose_sports"]] +
  coef(fit_3)[["choose_sports:diff"]]*mean(survey$diff)
```

#### 2. Simulating potential patterns of bias

The sports-theater demonstration demonstrated the challenge of causal inference for a self-selected treatment. Here we simulate such a problem in a slightly simpler setting.

Consider an education experiment with pre-test and post-test, where the treatment is more likely to be given to students who performed poorly on the pre-test. A simple comparison will give a biased estimate of treatment effect; we must adjust for pre-test. First we simulate the data:

```
n <- 100
x <- rnorm(n, 50, 20)
x[x < 0] <- 0
x[x > 100] <- 100
z <- rbinom(n, 1, 1 - x/100)
a <- 10
b <- 0.7
theta <- 10
y <- a + b*x+ theta*z + rnorm(n, 0, 10)
fake <- data.frame(x, y, z)
```

Then we perform the simple comparison:

```
library("rstanarm")
diff <- mean(y[z==1]) - mean(y[z==0])
print(diff)
fit_1 <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit_1)
```

Next, the comparison adjusting for pre-test:

```
fit_2 <- stan_glm(y ~ z + x, data=fake, refresh=0)
print(fit_2)
```

This works!

And we can and should plot the data and fitted model:

```
plot(fake$x, fake$y, col=ifelse(fake$z==1, "red", "blue"),
     pch=20, xlab="pre-test", ylab="post-test", bty="l")
abline(coef(fit_2)[["Intercept"]], coef(fit_2)[["x"]], col="blue")
abline(coef(fit_2)[["Intercept"]] + coef(fit_2)[["z"]], coef(fit_2)[["x"]],
      col="red")
```

Now we consider a more difficult setting in which selection depends on an unmeasured variable:

```
x_obs <- x + rnorm(n, 0, 20)
fake_new <- data.frame(x_obs, y, z)
fit_3 <- stan_glm(y ~ z + x_obs, data=fake_new, refresh=0)
print(fit_3)
```

In this case, adjusting for  $x_{obs}$  does not correct for all of the bias, but we would still recommend it as a step in the right direction.

## Drills

### 1. Regression coefficients as comparisons

For each of the following examples, express the underlined coefficient as a comparison.

(a) incumbent party vote share = 46.3% + 3.0% \* (percentage economic growth) + error

*Solution:* Comparing two election years for which economic growth differs by 1%, on average we would expect incumbent party vote share to be 3 percentage points higher in the year when economic growth is higher.

(b) post-test = 20 + 9.3 \* treatment + 0.75 \* pre-test + error

(c) post-test = 20 + 9.3 \* treatment + 0.75 \* pre-test + 0.35 \* treatment \* pre-test + error

(d) post-test = 20 + 9.3 \* treatment + 0.75 \* pre-test + 0.35 \* treatment \* pre-test + error

### 2. Log transformations

For each of the following examples, express the underlined coefficient as a comparison, first on the transformed scale, then on the untransformed scale.

(a)  $\log(\text{metabolic rate}) = 1.2 + \underline{0.74} * \log(\text{body mass}) + \text{error}$

*Solution:*

i. Comparing two species of animals that differ by 1 in  $\log(\text{body mass})$ , on average we would expect  $\log(\text{metabolic rate})$  to be 0.74 higher for the heavier species.

ii. Comparing two species of animals that differ by 1% in body mass, on average we would expect the heavier species to have a body mass that is 0.74% higher.

#### 4.14. REVIEW OF BASIC STATISTICS AND REGRESSION MODELING

173

- (b)  $\log(\text{population}) = 2.1 + \underline{0.014} * (\text{year} - 1900) + \text{error}$
- (c)  $\log(\text{demand}) = 8.5 - \underline{0.8} * \log(\text{price}) + \text{error}$

#### Discussion problems

##### 1. Sampling and adjustment

Students should work in pairs or small groups and consider a sampling problem of interest to them, for example a public opinion survey or a sample of trees or wild animals or a sample of business records. For this example, consider ways in which the sample would not be representative of the population, because of undercoverage, unequal sampling probabilities, nonresponse, and other issues. Then discuss what sort of statistical adjustments could be done to account for these problems of non-representativeness. We return to this topic in Chapter 17 of *Regression and Other Stories*; the point of this discussion problem is to get students thinking about this important topic.

##### 2. Causal inference and adjustment for context

Students should work in pairs or small groups and consider a causal inference problem of interest to them that could be studied with a randomized experiment, for example estimating the effect of a new persuasion technique in political marketing or the health effects of some environmental exposure. For this problem, consider how the effect of interest could vary across the population, over time, or based on context, and from there discuss challenges of using data from the specific conditions of an experiment to generalize to a population of interest. What statistical adjustments could be done to bridge from sample to population? We return to this topic in Chapter 19 of *Regression and Other Stories*; the point of this discussion problem is to get students thinking about this important topic.

## 4.15 Logistic regression

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Item-response analysis of final exams	“Two truths and a lie” game	Displaying a logistic curve	Divide-by-4 rule	Real-world logistic regression
Survey nonresponse	Predict the views of others	Logistic regression probabilities	Interpret logistic coefficients	Where logistic regression makes no sense

### Reading

Chapter 13 of *Regression and Other Stories*: Logistic regression

### Pre-class warmup assignments

1. Logistic function
  - (a) What are the values of  $\text{logit}(x)$ , for  $x = 0.01, 0.25, 0.5, 0.75, 0.99$ ?
  - (b) What are the values of  $\text{logit}(x)$ , for  $x = 1, 2, 3$ ?
  - (c) For what value  $x$  is  $\text{logit}(x)$  equal to 0? 1? 2? -1? -2?
2. Logistic regression in R
  - (a) Simulate 100 data points  $y_i$  from the model  $\Pr(y_i = 1) = \text{logit}^{-1}(a + bx_i)$ , where  $a = -0.5$ ,  $b = 0.1$ , and the data points  $x_i$  are sampled randomly from the interval  $(0, 20)$ . Fit a logistic regression to these data.
  - (b) Plot the data and the fitted curve on a single graph.

### Homework assignments

1. (a) Fitting logistic regression to data (Exercise 13.1 of *Regression and Other Stories*)  
The folder NES contains the survey data of presidential preference and income for the 1992 election analyzed in Section 13.1 of *Regression and Other Stories*, along with other variables including sex, ethnicity, education, party identification, and political ideology.
  - i. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.
  - ii. Evaluate and compare the different models you have fit.
  - iii. For your chosen model, discuss and compare the importance of each input variable in the prediction.(b) Sketching the logistic curve (Exercise 13.2 of *Regression and Other Stories*)  
Sketch the following logistic regression curves with pen on paper:
  - i.  $\Pr(y = 1) = \text{logit}^{-1}(x)$
  - ii.  $\Pr(y = 1) = \text{logit}^{-1}(2 + x)$
  - iii.  $\Pr(y = 1) = \text{logit}^{-1}(2x)$
  - iv.  $\Pr(y = 1) = \text{logit}^{-1}(2 + 2x)$
  - v.  $\Pr(y = 1) = \text{logit}^{-1}(-2x)$
2. (a) Understanding logistic regression coefficients (Exercise 13.3 of *Regression and Other Stories*)

Chapter 7 of *Regression and Other Stories* discusses a model predicting incumbent party's two-party vote percentage given economic growth:  $\text{vote} = 46.2 + 3.1 * \text{growth} + \text{error}$ , where growth ranges from  $-0.5$  to  $4.5$  in the data, and errors are approximately normally distributed with mean  $0$  and standard deviation  $3.8$ . Suppose instead we were to fit a logistic regression,  $\Pr(\text{vote} > 50) = \text{logit}^{-1}(a + b * \text{growth})$ . Approximately what are the estimates of  $(a, b)$ ?

Figure this out in four steps: (i) use the fitted linear regression model to estimate  $\Pr(\text{vote} > 50)$  for different values of growth; (ii) plot these probabilities and draw a logistic curve through them; (iii) use the divide-by-4 rule to estimate the slope of the logistic regression model; (iv) use the point where the probability goes through  $0.5$  to deduce the intercept. Do all this using the above information, without downloading the data and fitting the model.

- (b) Graphing a fitted logistic regression (Exercise 13.3 of *Regression and Other Stories*)

We downloaded data with weight (in pounds) and age (in years) from a random sample of American adults. We then defined a new variable,

```
heavy <- weight > 200
```

and fit a logistic regression, predicting heavy from height (in inches):

```
stan_glm(formula = heavy ~ height, family=binomial(link="logit"),
          data=health)
      Median MAD_SD
(Intercept) -21.51   1.60
height       0.28   0.02
```

- i. Graph the logistic regression curve (the probability that someone is heavy) over the approximate range of the data. Be clear where the line goes through the 50% probability point.
- ii. Fill in the blank: near the 50% point, comparing two people who differ by 1 inch in height, you'll expect a difference of \_\_\_\_ in the probability of being heavy.

## Stories

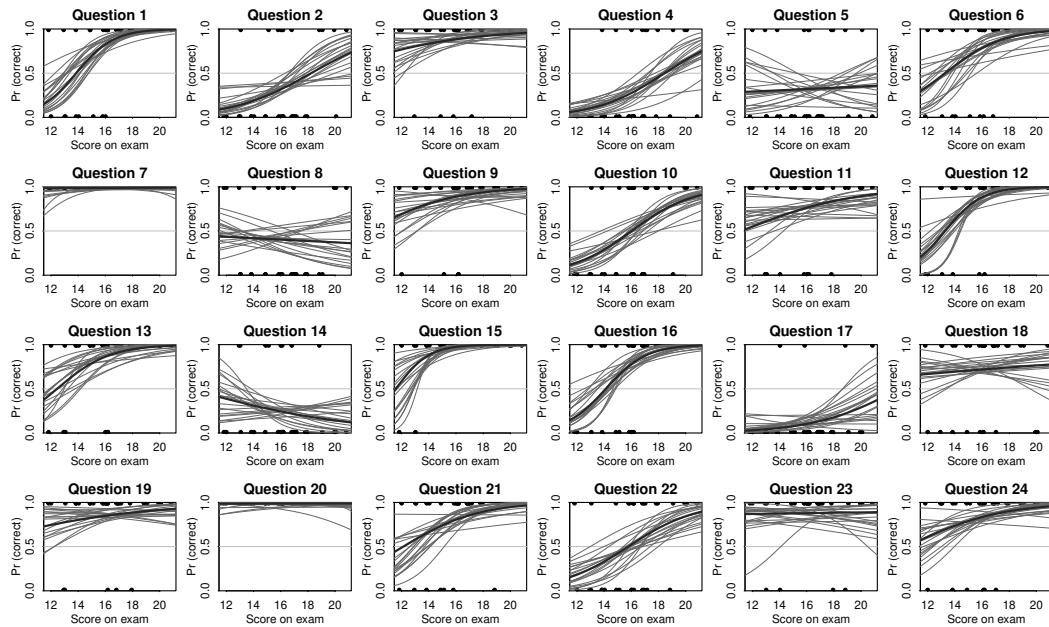
### 1. Item-response analysis of final exams

Our final exam for the first semester of the course was a multiple-choice test with 24 questions: two questions corresponding to each of the first 12 chapters of the textbook; see Appendix B. To get a sense of how the students did on the exam, we plot in Figure 63 the performance on each question vs. total exam score, so that, on each graph, each dot corresponds to a student. Each graph also shows a fitted logistic regression giving the estimated probability of a correct answer given total score, along with several gray lines representing uncertainty in the fitted curve as indicated by draws from the posterior distribution of the estimated coefficients.

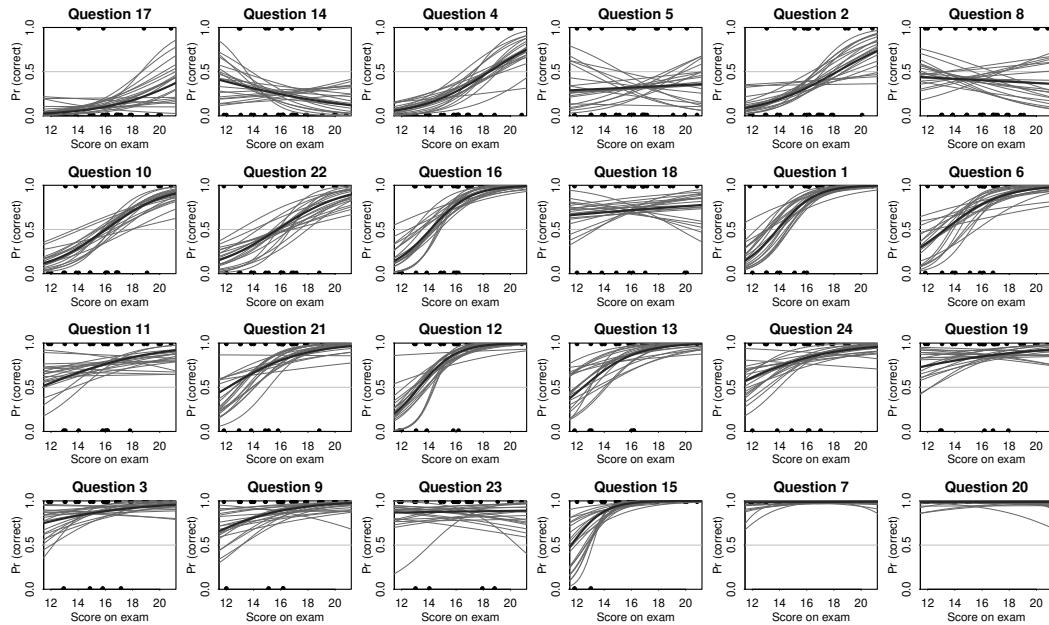
We can make the display in Figure 63 more readable by putting the questions in logical order. Figure 64 orders the questions in increasing proportion of correct answers. Now we can see patterns in the data more clearly.

Hey—what happened with question 17, that almost all the students got wrong? We went back and checked, and it turned out we messed up on the answer key for questions 17, 14, and 5, so we recalculated the students' scores. Figure 65 shows the new plot, reordering the questions in order of increasing percentage of correct answers.

This story is relevant to the week's reading in demonstrating a simple logistic regression with a single predictor. The example of final exam grades is something that all students can relate to. The story relates to the course as a whole by showing how we can learn from displaying data and a fitted model, even to the extent of discovering a problem with our exam grading. It also illustrates the benefits of fitting a simple model many times and then comparing the results.



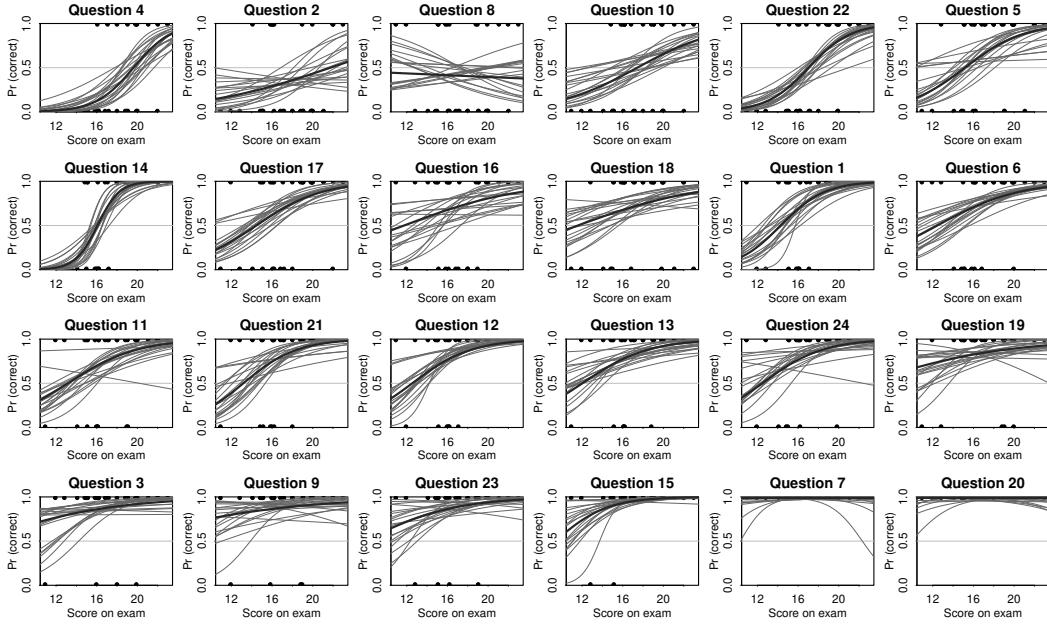
**Figure 63** First of three grids of graphs showing correct or incorrect answers on first-semester final exam questions plotted vs. exam score, for 32 students in a class. For each question we fit a logistic regression and display the fitted logistic curve and its uncertainty.



**Figure 64** Reordering Figure 63: Final exam results plotted vs. students' total scores, with questions ordered in increasing percentage of correct answers. Looking carefully at this graph reveals data errors.

#### 4.15. LOGISTIC REGRESSION

177



**Figure 65** After correcting data errors, final exam results plotted vs. students' total scores, with questions ordered in increasing percentage of correct answers.

#### 2. Survey nonresponse

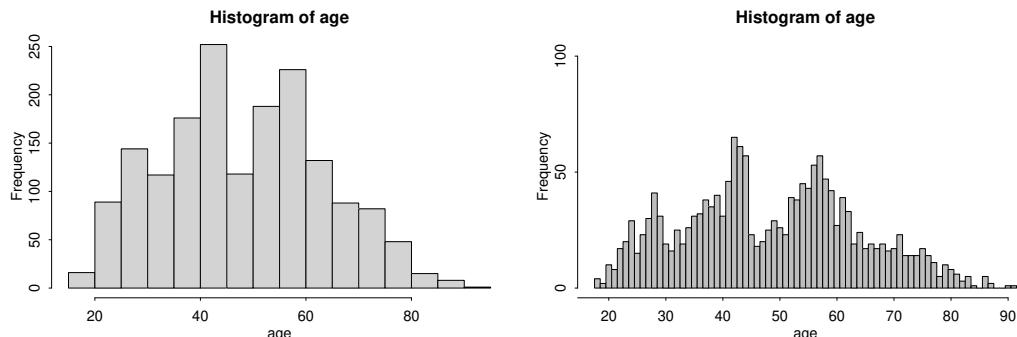
Figure 66 shows the distribution of ages from the respondents to a survey that we analyzed as part of a research project. The project was on “social penumbras”—asking respondents how many people they knew in various social groups, such as gay people, Muslims, or active-duty military service members—and the survey also included various demographic variables such as age, sex, and ethnicity. We first used the default settings for the `hist()` function in R, then we specified the histogram bars to display more detail, as shown in Figure 66b. The distribution of ages looks wrong. Beyond the expected pattern of fewer respondents at the youngest and oldest ages (lower response rates among the youngest and fewer people alive at the oldest ages), there is a strange pattern with peaks in the late twenties, early forties, and mid-fifties. What’s going on?

The data were collected from an internet panel. The survey organization, YouGov, “aims for a sample of American adults using quota sampling on age, sex, and other demographics.” We received data from 3333 people who were interviewed in the first wave of the survey. The respondents were contacted again a year later, and 2106 people responded; after data cleaning, these were reduced to a sample of 1700 respondents to both waves. The wave 1 sample was unweighted, but weights were supplied for the final sample to help deal with dropout.<sup>8</sup> The histograms shown in Figure 66 show the ages of the final 1700 respondents.

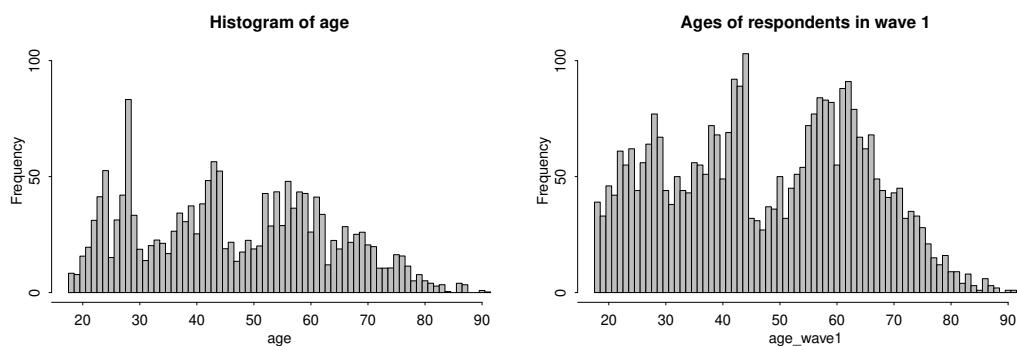
Perhaps the weird age patterns in Figure 66 came from using unweighted data? Figure 67a shows the result using weighted data. The weights increase the representation of the younger ages (no surprise, given that younger people are less likely to respond to the second wave of the survey), but the strange patterns are still there.

As noted above, we have been working with a sample of people who responded to both waves of the survey. Perhaps the unusual age distribution came from some combination of nonresponse and adjustments that were made in the second wave? To check, we plot in Figure 67b the age

<sup>8</sup> Andrew Gelman and Yotam Margalit (2020), Social penumbras predict political attitudes, *Proceedings of the National Academy of Sciences* 118 (6), e2019375118.



**Figure 66** Distribution of ages from a two-wave internet panel survey designed by us and administered by YouGov: (a) Initial try at a histogram, (b) more precise version with a bar for every age. There is a strange pattern with peaks in the late twenties, early forties, and mid-fifties.



**Figure 67** (a) Histogram of ages from the weighted sample of the second wave of our internet panel survey; (b) ages from the larger sample in the first wave. The distribution of ages still has mysterious peaks in the late twenties, early forties, and mid-fifties through early sixties.

distribution of the 3333 first-wave respondents. They also show that strange pattern of ages. How did it happen? We suspect it's a combination of nonresponse and survey design. To start with, younger people are less likely to respond to surveys. To correct for this, YouGov uses a quota sampling design, with one of the factors being age. In a quota sample, you keep adding more respondents until you get your pre-specified "quota" in each category. In this case, we think the age quotas are 18–29 years old, 30–44, 45–64, and 65+. So we end up with about the right number of people in each age category, but because of differential nonresponse by age, we see not enough people at the low end of each category and too many at the high end. The result is a sawtooth pattern with big drops at 18, 30, and 45. The drops are not completely sharp because we do not actually have age in our survey; we just have year of birth, so our estimated ages can be off by 1.

All of this motivates us to look more carefully at the pattern of nonresponse by age. The original sample of 3333 people was selected using quota sampling so we have no clear sense of response patterns there, but we can look at who among them responded when contacted for the second wave.

Figure 68a shows, for each age, the proportion of people in the initial panel who responded to the request a year later. The pattern is striking. The response rate is below 10% for the youngest group and then rapidly rises from there, stabilizing at about 75% for people 50 and older. There is some variation at the highest ages but this can be attributed to tiny sample sizes: when there is only one person in an age category, the observed response rate can only be 0 or 100%.

Figure 68b shows a logistic regression fit to these data. The curve captures the general overall

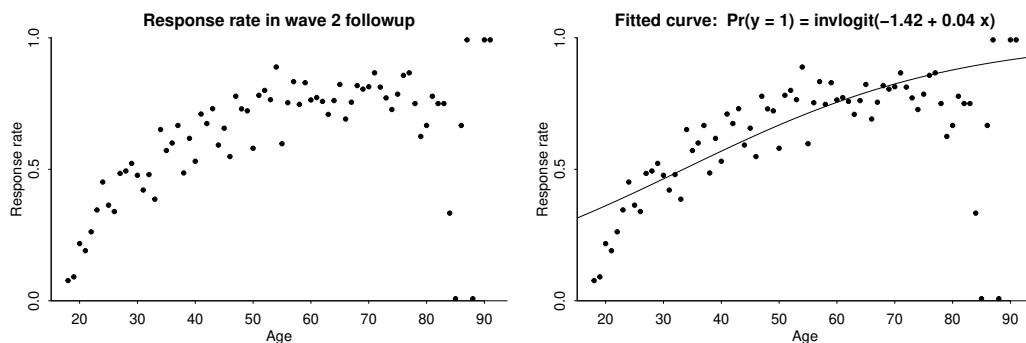


Figure 68 (a) For each age, the proportion of the first-wave panel who responded to the second wave of our survey; (b) a logistic regression fit to these data.

trend but has some obvious problems too. Nonetheless, we pursue this further by predicting response given age, sex, and ethnicity:

```
family:      binomial [logit]
formula:     respond ~ age + female + ethnicity
observations: 3333
predictors:   6
-----
                           Median MAD_SD
(Intercept)      -1.32    0.14
age              0.04    0.00
female           0.28    0.08
ethnicityBlack   -0.55    0.12
ethnicityHispanic -0.56    0.13
ethnicityOther    -0.21    0.16
```

We ask students to discuss these results in pairs.

This story relates to the week’s reading as an example of logistic regression applied in a case where it does not fit the data well. It relates to the course more generally by demonstrating the real-world challenge of survey nonresponse and how this can show up in data.

### Class-participation activities

#### 1. “Two truths and a lie” activity, calibrating probability estimates

In this activity,<sup>9</sup> students divide into groups of four—it’s OK if some groups have three or five students in them—to play “two truths and a lie.” We display the instructions in Figure 69 onto the screen and explain the procedure. In this game, one person makes three personal statements; two of these statements should be true and one should be false. The other students in the group then briefly confer and together guess which statement is the lie. They should jointly construct a numerical statement of their certainty about their guess, on a 0–10 scale, where 0 represents pure guessing and 10 corresponds to complete certainty. The true statement is then revealed so that the students know if they guessed correctly. Each group of students then rotates through, with each student playing the role of storyteller, so that when the activity is over, each group of four students has four certainty scores, each corresponding to a success or a failure. Figure 70 shows an example.

<sup>9</sup>Andrew Gelman (2023), “Two truths and a lie” as a class-participation activity, *American Statistician* 77, 97–101.

Within your group:

1. One person tells three personal statements, one of which is a lie.
2. Others discuss and guess which statement is the lie, and they jointly construct a numerical statement of their certainty in the guess (on a 0–10 scale).
3. The storyteller reveals which was the lie.
4. Enter the certainty estimate and the outcome (success or failure) and submit in the Google form.

Rotate through everyone in your group so that each person plays the storyteller role once.

Figure 69: Instructions for the “two truths and a lie” activity, to project onto the screen for students.

Certainty	Outcome
8	Success
4	Success
7	Failure
5	Success

The Google form interface shows the title 'Two truths and a lie'. Below it is a note about saving progress. The 'Certainty in your guess' section contains a scale from 0 to 10 with radio buttons. The 'Outcome' section has two radio button options: 'Guessed right' and 'Guessed wrong'. At the bottom are 'Submit' and 'Clear form' buttons.

Figure 70 (a) Example of data produced by a group of four students playing the “two truths and a lie” game. Each number is a group consensus: here, the group of students B, C, D assigned a certainty of 8 to their guess of student A’s lie, and they were correct; students A, C, D assigned a certainty of 4 to their guess of student B’s lie, and they happened to be correct, and so on. (b) Students play the game in groups and then enter the data one at a time into a Google form, so that the instructor can analyze all the results together.

We then give students the URL of a Google form where they can enter their data using their phones or laptops. The form is set up to take one response at a time, so each group should enter four responses corresponding to their four guesses. Alternatively we could set up a longer Google form allowing a group to enter all four responses together, but that would require additional data processing on the analysis end, so we go with this simpler approach that does not keep track of the clustering of the responses.

We download the data from the Google form as a csv file, read it into our statistical software, and announce that we will display the data (a scatterplot of the success/failure outcome vs. certainty score) along with a fitted curve showing the estimated probability of a choice being correct as a function of certainty score. If the class is sufficiently advanced, we explain that the fitted curve will be a logistic regression; otherwise we simply say we will fit a curve.

Before making the plot and displaying the data and fit, we ask students in their groups to sketch what they think the scatterplot and fitted curve for the class will look like, and then we lead the class in discussion. Some possible prompts include: What do you think the range of certainty scores will look like: will there be any 0’s or 10’s? Will there be a positive relation between  $x$  and  $y$ : are guesses with higher certainty be more accurate, on average? How strong will the relation be

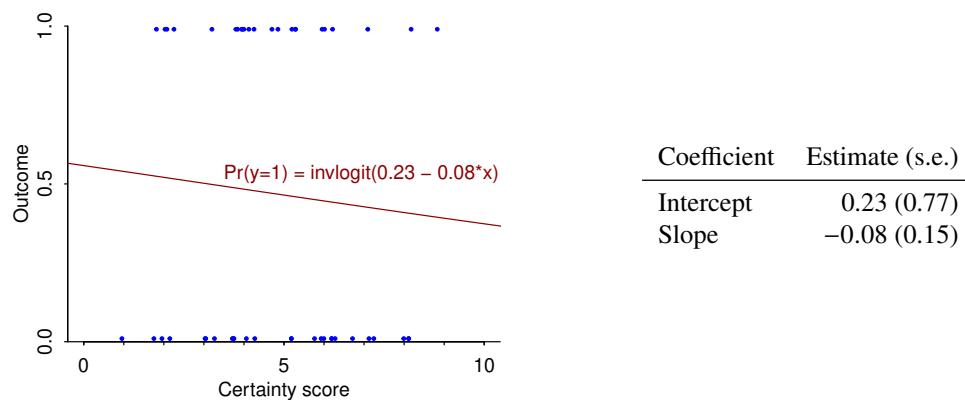


Figure 71 (a) Scatterplot from the “two truths and a lie” activity performed in a class of 49 students, along with a curve showing a fitted model predicting correctness given the certainty scores (which have been jittered to avoid points overlapping on the graph); (b) Coefficient estimates and standard errors from the fitted logistic regression. The instructor can perform these steps in real time on the data that students have entered in their Google form.

between  $x$  and  $y$ : what will the curve look like? If students have seen logistic regression, we ask them to give approximate numerical values for the intercept and slope coefficients corresponding to their sketched curves.

After this discussion, we display the data and fitted curve and conduct a followup discussion of what has been learned. Figure 71 shows an example of real data from our class. In this case, there was essentially no relation between the certainty score and the outcome (coded as 1 for a successful guess and 0 for an error). In fact, the estimated logistic regression coefficient is negative: higher certainty scores correspond to slightly lower rates of accuracy! It’s hard to see this in the plot of raw responses; in general, scatterplots are not so helpful for displaying discrete data. The standard error from the regression gives a sense of the uncertainty in the fit; in this case, the right panel of Figure 71 shows an estimated slope of  $-0.08$  with standard error 0.15, indicating that the sign of the underlying relationship is unclear from the data. For this class, the students’ stated certainty scores do not form a useful predictor of their actual knowledge.

In the discussion that followed in our class, students conjectured why their guesses were so bad (23 of 49 correct, not much better than the  $\frac{1}{3}$  success rate that would come from random guessing) and why their certainty judgments were not predictive of their accuracy. In many ways, the class discussion before seeing the data was better than the post-data followup, which illustrates a general point that concepts can sometimes be clearer in theory, with real data providing a useful check on speculation.

For any particular class, the interpretation of the “two truths and a lie” experiment will depend on the data that come in, and the instructor should be prepared for anything. In the class discussion before the scatterplot and fitted model are revealed, it is natural for students to expect the certainty score to be a strong predictor of empirical accuracy. If this occurs, great; if not, this is an excellent opportunity to discuss the challenges of measurement and the value of statistical evaluation of a measurement protocol.

This activity is relevant to the week’s reading by demonstrating logistic regression based on data directly collected from students. It relates to the course as a whole as an example of calibration of measurements. This sort of calibration problem arises in many areas of science and policy. For example, consider a hiring setting where interviewers give numerical ratings for the candidates, and then later when there are data on job performance, the ratings can be retrospectively calibrated, given an empirical measure of the information provided by the ratings.

2. Predict the views of others

In preparation for conducting a survey in class, we ask students in pairs to come up with some questions to ask, questions for which they are interested in learning the average opinions of their fellow students. Topic areas could include course policies (Should there be a midterm exam in the course? Should there be a final project?), public health (for example, something on vaccine or mask requirements), economic or foreign policy, or personal questions (for example, questions about leisure activities). We then have a brief class discussion and come up with five questions for students to privately consider. For each we will request a response on a 1–4 scale (“strongly disagree,” “disagree,” “agree,” “strongly agree”), using an even number of categories so that respondents can’t hide in the middle category to express no opinion. Each student should separately write responses (each on the 1–4 scale) on a sheet of paper.

We then type the questions into a Google form, and then, for each, an item asking whether the majority of students in the class will have responded positively or negatively. We announce that will give a prize to any student who guesses all five correct. We add a couple of demographic questions such as age and sex to the Google form and then share the URL with the students and give them a couple minutes to fill in the questions individually.

We then open up R and read in the data from the Google form. We label the five survey responses as  $Q_1, \dots, Q_5$  and the guesses as  $G_1, \dots, G_5$ . For each of the five questions  $j$ , we will fit a logistic regression predicting  $G_j$  from  $Q_j$ . Here,  $G_j$  is a binary response (1 for students who guess that the majority of students responded with 3’s or 4’s to  $Q_j$ , 0 for those who guess that the majority responded with 1’s or 2’s). We set up the code to fit the regressions, but before running, we ask students in pairs to guess what the fits will look like. We then fit the model on the computer and discuss the results. If there is time, we can also add the demographics as predictors in the model and discuss how the coefficient for  $Q_j$  changes.

It is helpful in this activity to have five different questions. For example, one might guess that  $G_j$  should have a positive coefficient in the logistic regression: students who agree with a particular position would be more likely to think that others would agree with them. Having data from five different questions gives us a chance to see if such a pattern holds consistently; results from just one question could be a fluke.

This activity relates to the week’s reading by asking students to think hard about the results of a simple logistic regression, with the small sample size motivating a discussion of inferential uncertainty (that is, the standard error of the coefficient of interest). It is relevant to the course as a whole as an example of the “secret weapon”: learning from multiple parallel datasets.

### Computer demonstrations

1. Display fitted logistic curve and uncertainties

Here is the code we used to fit a logistic regression and then graph the data, fitted curve, and uncertainties. First we read in and organize the data:

```
responses <- read.csv("final_exam_responses.csv")
answers <- read.csv("final_exam_answers.csv")
n <- nrow(responses)
score <- rep(NA, n)
for (i in 1:n){
  score[i] <- sum(as.character(responses[i,]) == as.character(answers))
}
```

Next we write a function for fitting the logistic regression and plotting it along with the data. But first we need some setup:

```
jitt <- runif(n, -0.2, 0.2) # avoid overplotting of data
```

Then the plotting function:

```
fit_and_plot <- function(j, responses, answers){  
  y <- as.character(responses[,j]) == as.character(answers[j])  
  data <- data.frame(y, score)  
  fit <- stan_glm(y ~ score, family=binomial(link="logit"), data=data,  
  refresh=0, mean_PPD=FALSE)  
  plot(score + jitt, y, xlab="Score on exam", ylab="Pr (correct)",  
  main=paste("Question", j))  
  curve(invlogit(coef(fit)[1] + coef(fit)[2]*x), add=TRUE)  
}
```

We try it out on question 1:

```
fit_and_plot(1, responses, answers)
```

It works, as can be seen in the graphics window of our RStudio session.

Next we want to display uncertainty. We add the following lines within the `fit_and_plot()` function to display curves corresponding to 20 random draws from the posterior distribution of the fitted model:

```
sims <- as.matrix(fit)  
n_sims <- nrow(sims)  
for (s in sample(n_sims, 20)) {  
  curve(invlogit(sims[s,1] + sims[s,2]*x), col="gray40", lwd=.5, add=TRUE)  
}
```

We put these at the end within the `fit_and_plot()` function, save it, and then again run it for the first question:

```
fit_and_plot(1, responses, answers)
```

It works!

Next we can plot all 24 questions:

```
pdf("exams_grid_1.pdf", height=6, width=10)  
par(mfrow=c(4,6))  
par(mar=c(3,3,2,.1), mgp=c(1.5,.5,0), tck=-.01)  
for (j in 1:24){  
  fit_and_plot(j, responses, answers)  
}  
dev.off()
```

This basically works, but there are some weird things going on with questions 7 and 20, where all the students got the correct answer, so the scales of these plots are off. We can fix this by adding `ylim=c(0,1)` to the `plot()` call inside the function.

This demonstration relates to the week's readings by showing how to fit and display logistic regressions. It relates to the course as a whole by showing data analysis that solves an applied problem.

2. Compute the difference in logistic regression probabilities and compare to the divide-by-4 rule

The divide-by-4 rule, explained in Section 13.2 of *Regression and Other Stories*, is a quick approximate calculation of the predicted difference in probabilities associated with a logistic regression coefficient. The rule works well when the predicted probabilities are not far from  $\frac{1}{2}$  but

gives a large overestimate of the probability difference when probabilities are close to 0 and 1. We explore this with a simulation.

```
# Define function
invlogit <- plogis # invlogit of linear predictor gives you p

# Graph inverse logit function
curve(invlogit(4 + 2*x), from=-12, to=8)
abline(v=-2) # Halfway point = -2; steepest slope = 0.5

# Calculate difference in probabilities around halfway point
x_left <- -2
x_right <- -1
p_left <- invlogit(4 + 2 * (x_left))
p_right <- invlogit(4 + 2 * (x_right))
print(p_right - p_left) # exact difference in p
abline(v = c(x_left, x_right), h = c(p_left, p_right), lty="dashed")

# Repeat with a much more spread-out function
curve(invlogit(10 - 0.1*x), from=0, to=200)
abline(v=100) # Halfway point = 100; steepest slope = ?
x_left <- 100
x_right <- 101
p_left <- invlogit(10 - 0.1 * (x_left))
p_right <- invlogit(10 - 0.1 * (x_right))
print(p_right - p_left) # exact difference in p
abline(v = c(x_left, x_right), h = c(p_left, p_right), lty="dashed")
```

## Drills

### 1. Divide-by-4 rule

For each of the following models, calculate the halfway point (where the predicted probability is 0.5) as well as the curve's steepest slope (using the divide-by-4 rule). When you are done, check your results visually by graphing the curves in R.

- (a)  $\Pr(y = 1) = \text{logit}^{-1}(x)$   
*Solution:* The halfway point is at  $x = 0$  and the steepest slope is  $1/4$ .
- (b)  $\Pr(y = 1) = \text{logit}^{-1}(1 + x)$
- (c)  $\Pr(y = 1) = \text{logit}^{-1}(4x)$
- (d)  $\Pr(y = 1) = \text{logit}^{-1}(1 + 4x)$
- (e)  $\Pr(y = 1) = \text{logit}^{-1}(-x)$
- (f)  $\Pr(y = 1) = \text{logit}^{-1}(1 - 4x)$
- (g)  $\Pr(y = 1) = \text{logit}^{-1}(-20 + 0.4x)$

### 2. Interpret logistic regression coefficients

For each of the following examples (all except the first based on Pew Research surveys), interpret the constant and the slope estimate (where appropriate, you can use the divide-by-4 rule). Plot the curves on an adequate range.

- (a) Probability of buying a new laptop in a given year, given age of current laptop:

$$\Pr(\text{purchase}) = \text{logit}^{-1}(-3 + 0.8 * \text{laptop\_age})$$

*Solution:*

- Intercept: For someone with a brand-new laptop, this model predicts the probability of buying a new laptop this year as  $\text{logit}^{-1}(-3.0) = 0.05$ , or 5%.

- Slope: Comparing people who have owned a laptop 0, 1, or 2 years, the estimated probability of buying a new one is  $\text{logit}^{-1}(-3.0) = 0.05$ ,  $\text{logit}^{-1}(-2.2) = 0.10$ , or  $\text{logit}^{-1}(-1.4) = 0.20$ , that is, from 5%, 10%, or 20%.
- (b) Probability of being married, given age (centered on mean age  $\approx 50$  years):  
 $\Pr(\text{married}) = \text{logit}^{-1}(0.24 + 0.01 * \text{age})$
- (c) Probability of identifying as a Democrat, given sex:  
 $\Pr(\text{Democrat}) = \text{logit}^{-1}(-0.89 + 0.50 * \text{female})$
- (d) Probability of owning your home (rather than renting), given marital status and age (centered):  
 $\Pr(\text{own}) = \text{logit}^{-1}(0.57 + 1.71 * \text{married} + 0.04 * \text{age})$

### Discussion problems

#### 1. Real-world example of logistic regression

Students should work in pairs to come up with examples where logistic regression could make sense, other than the cases we've considered in class or the readings. Any example should include a binary outcome  $y$ , a continuous predictor  $x$ , a range of values for  $x$ , and a sample size.

The class should then discuss and settle on one of these examples to work on together. For this example, students should sketch in pairs what they think the data could like, along with the logistic curve showing  $\Pr(y = 1)$  as a function of  $x$ , and from that sketched curve determine the intercept and slope of the logistic regression. Once this has been done, the instructor should call on some groups and ask them to go to the board, write their intercept and slope, and graph their guessed curve. Different pairs can have much different curves, which can spark some discussion about what curves are reasonable for the example.

#### 2. Real-world example where logistic regression does not make sense

One way to understand logistic regression is to ask what it would take for the model to *not* make sense for binary data. Some possibilities are if  $\Pr(y = 1)$  is not a smooth function or is not a monotonic function of  $x$ , or if it asymptotes out at a value other than 0 or 1. Students should work in pairs to come up with examples for which logistic regression would not make sense, with any example including a binary outcome  $y$ , a continuous predictor  $x$ , a range of values for  $x$ , and a sample size.

For their example, each pair of students should sketch what they think the data could like, showing  $\Pr(y = 1)$  as a function of  $x$ , along with the best-fit logistic curve—that is, what would happen if logistic regression were to be fit to those data, despite their not fitting the model. Once this has been done, the instructor should call on some groups and ask them to go to the board and sketch their data and curves. The point of this example is not to come up with alternative models but rather to explore different ways that logistic regression can fail, thus clarifying the assumptions of the model.

## 4.16 Working with logistic regression

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
“Keys to the White House”	Job training and predictive comparisons	Predictions from logistic regression	Probabilities in logistic regression	Experimental design
Opiate of the masses	Logistic regression with interactions	Linear or logistic regression	Understand the logit function	Design with pre-test

### Reading

Chapter 14 of *Regression and Other Stories*: Working with logistic regression

### Pre-class warmup assignments

1. Plot logistic regression in R
  - (a) Plot the curve,  $\Pr(y = 1) = \text{logit}^{-1}(2 + 3x)$ , for  $x$  in the range  $(0, 5)$ .
  - (b) Consider the model,  $\Pr(y = 1) = \text{logit}^{-1}(2 + 3x - 3z)$ , for  $x$  in the range  $(0, 5)$  and  $z$  taking on the values 0, 1, or 2. Plot the curves as a function of  $x$  for the values  $z = 0, 1$ , and 2, all on the same graph.
2. Plot logistic regression with interactions
  - (a) Consider the model,  $\Pr(y = 1) = \text{logit}^{-1}(-0.2 + 0.3x - 1.5z - 0.2 * x * z)$ , for  $x$  in the range  $(0, 10)$  and  $z$  taking on the values 0, 1, or 2. Write the model for  $y$  given  $x$  for each of  $z = 0$ ,  $z = 1$ , and  $z = 2$ .
  - (b) Make a single plot of  $\Pr(y = 1)$  as a function of  $x$ , showing all three curves of  $z = 0, 1$ , and 2. Label all three curves on the graph.
  - (c) At what value of  $x$  do the three curves cross? Explain why they cross at this point.

### Homework assignments

1. (a) Logistic regression with two predictors (Exercise 13.4 of *Regression and Other Stories*)

The following logistic regression has been fit:

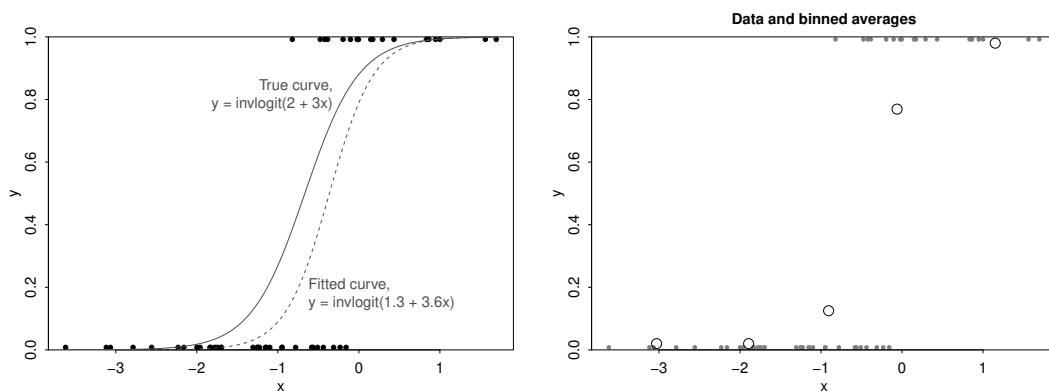
	Median	MAD_SD
(Intercept)	-1.9	0.6
x	0.7	0.8
z	0.7	0.5

Here,  $x$  is a continuous predictor ranging from 0 to 10, and  $z$  is a binary predictor taking on the values 0 and 1. Display the fitted model as two curves on a graph of  $\Pr(y = 1)$  vs.  $x$ .

- (b) Interpreting logistic regression coefficients (Exercise 13.5 of *Regression and Other Stories*)

Here is a fitted model from the Bangladesh analysis predicting whether a person with high-arsenic drinking water will switch wells, given the arsenic level in their existing well and the distance to the nearest safe well:

```
stan_glm(formula = switch ~ dist100 + arsenic,  
family=binomial(link="logit"), data=wells)
```



**Figure 72** (a) Data simulated from a logistic regression model, along with the logistic regression fit to these data; (b) Binned averages, plotting  $\bar{y}$  vs.  $\bar{x}$  for the data divided into five bins based on the values of  $x$ .

	Median	MAD_SD
(Intercept)	0.00	0.08
dist100	-0.90	0.10
arsenic	0.46	0.04

Compare two people who live the same distance from the nearest well but whose arsenic levels differ, with one person having an arsenic level of 0.5 and the other person having a level of 1.0. You will estimate how much more likely this second person is to switch wells. Give an approximate estimate, standard error, 50% interval, and 95% interval, using two different methods:

- i. Use the divide-by-4 rule, based on the information from this regression output.
  - ii. Use predictive simulation from the fitted model in R, under the assumption that these two people each live 50 meters from the nearest safe well.
2. (a) Graphing binary data and logistic regression (Exercise 14.1 of *Regression and Other Stories*)  
 Reproduce Figure 72 with the model,  $\Pr(y = 1) = \text{logit}^{-1}(0.4 - 0.3x)$ , with 50 data points  $x$  sampled uniformly in the range  $[A, B]$ . (In Figure 72 the  $x$ 's were drawn from a normal distribution.) Choose the values  $A$  and  $B$  so that the plot includes a zone where values of  $y$  are all 1, a zone where they are all 0, and a band of overlap in the middle.
- (b) Limitations of logistic regression (Exercise 14.6 of *Regression and Other Stories*)  
 Consider a dataset with  $n = 20$  points, a single predictor  $x$  that takes on the values  $1, \dots, 20$ , and binary data  $y$ . Construct data values  $y_1, \dots, y_{20}$  that are inconsistent with any logistic regression on  $x$ . Fit a logistic regression to these data, plot the data and fitted curve, and explain why you can say that the model does not fit the data.
- (c) *In pairs:* Working through your own example (Exercise 13.13 of *Regression and Other Stories*)  
 Continuing the example from the final exercises of the earlier chapters, fit a logistic regression, graph the data and fitted model, and interpret the estimated parameters and their uncertainties.

## Stories

1. “Keys to the White House” and why it’s better to model continuous outcomes when possible

A book came out in several years ago called “The Keys to the White House,” predicting the presidential election winner by tallying 13 factors and following the rule that “When five or fewer statements are false, the incumbent party wins.” We have been told that “Retrospectively, the Keys

<sup>10</sup>Allan Lichtman (2008), *The Keys to the White House, 2008 Edition*, Rowman & Littlefield.

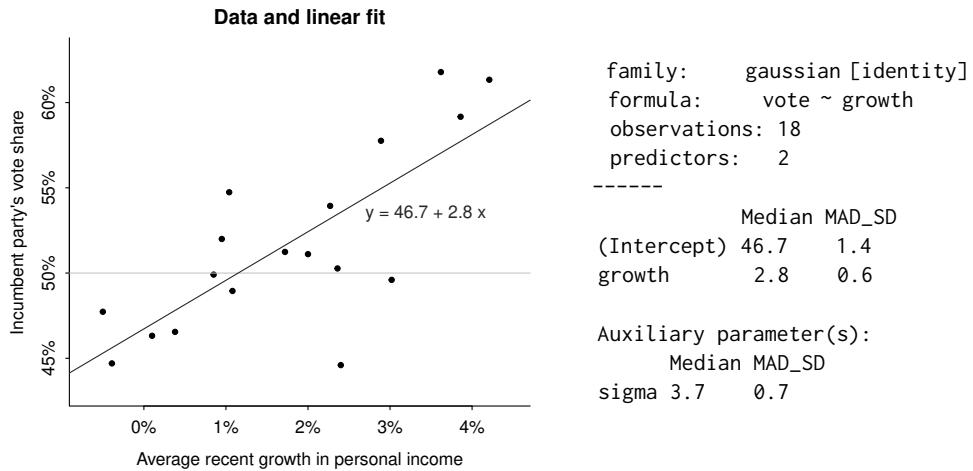


Figure 73: Forecasting U.S. presidential elections from the economy using a linear regression.

account for presidential election outcomes from 1860 to 1980” and that the model has correctly predicted the winner of each presidential race since 1984.<sup>10</sup>

The idea of a perfectly accurate deterministic prediction is ridiculous, yet the “Keys to the White House” continue to get publicity every four years.<sup>11</sup> We explore some problems with this idea as a way of introducing some important general themes of statistical prediction.<sup>12</sup>

To start with, there’s a problem with attempting to predict the winner of every election. Since the Second World War, there have been three U.S. presidential elections that have been essentially tied in the final vote: 1960, 1968, and 2000. It’s meaningless to say that a forecasting method predicts the winner correctly (or incorrectly) in these cases. And from a statistical point of view, you don’t want to adapt your model to fit these tossups—it’s just an invitation to overfitting.

To put it another way: suppose the method mispredicted 1960, 1968, and 2000. Would we think any less of it? No. No credit should come from predicting a coin flip. At the other extreme, just about any method should predict the winner in a landslide election such as Ronald Reagan’s victory in 1984, so not much is learned from a correct forecast in that year. A method that predicts vote share, rather than just the winner, could get credit from these close elections and these landslides by predicting the vote share with high accuracy.

Beyond all this, the model didn’t actually correctly predict all elections since 1984. The model predicted a Gore win in 2000, which was counted as a correct forecast because Al Gore won the popular vote that year—but then it was also counted as a win that the model predicted Trump’s victory in 2016. You can’t have it both ways!

As discussed in Chapter 7 of *Regression and Other Stories*, our preferred approach is to predict vote share rather than win/loss. Figure 73 shows the result for a linear regression predicting

<sup>10</sup>For example, Allison Gordon (2020), History professor who has accurately predicted every election since 1984 says Trump will lose, CNN, <https://www.cnn.com/2020/08/07/us/allan-lichtman-trump-biden-2020-trnd/index.html>.

<sup>11</sup>See Andrew Gelman (2007), Some thoughts on “the keys to the White House,” [https://statmodeling.stat.columbia.edu/2007/02/05/some\\_thoughts\\_o\\_1/](https://statmodeling.stat.columbia.edu/2007/02/05/some_thoughts_o_1/); Andrew Gelman (2007), Treating discrete variables as if they were continuous, [https://statmodeling.stat.columbia.edu/2007/05/25/treating\\_discrete/](https://statmodeling.stat.columbia.edu/2007/05/25/treating_discrete/); Andrew Gelman (2014), Basketball stats: Don’t model the probability of win, model the expected score differential, <https://statmodeling.stat.columbia.edu/2014/02/25/differential/>; and Andrew Gelman (2020), Rodman, <https://statmodeling.stat.columbia.edu/2020/04/26/rodman/>.

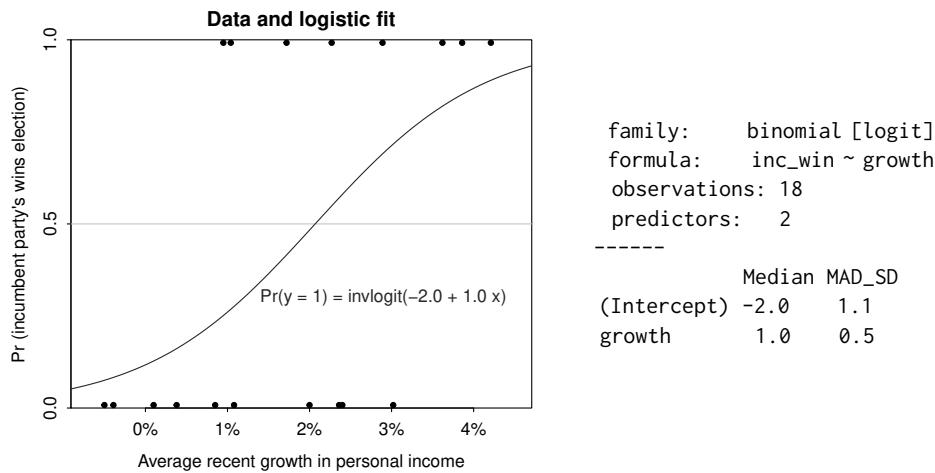


Figure 74: Forecasting winners of U.S. presidential elections from the economy using a logistic regression.

incumbent party share of the two-party vote from economic growth, fit to all the presidential elections between 1948 and 2020.

What happens if we just predict the winner of the election? Figure 74 shows the fitted logistic regression. Just including wins and losses throws away a lot of information.

A similar issue arises in sports: if you're predicting football or basketball or baseball or soccer games, don't model the probability of a win, model the expected score differential. Yeah, we know, we know, what you really want to know is who wins. But the most efficient way to get there is to model the score differential and then map that back to win probabilities.

## 2. Opiate of the masses and post-materialism: Logistic regression with interaction

Surveys have found that, in recent decades, people who regularly attend religious services are more likely to vote for Republicans. Traditionally we think of voters as being divided by income or social class, and the correlation between religious attendance and partisan voting came as a surprise.

There have been two ways of thinking about religion and voting. The first is the “opiate of the masses” model, in which lower-income voters are distracted from their economic interests and vote based on social issues instead. The second model is “postmaterialism,” in which voting on social issues is a sort of luxury. We summarize in Figure 75 and write this on the board for students to see during the discussion.

We ask students to discuss (in pairs) these two models. Neither is true in any absolute sense; there are people of all income levels who make voting decisions based on various mixes of economic and social concerns. We will interpret these models as posited statistical relationships. We ask each pair of students to make a graph with probability of Republican vote on the  $y$ -axis and income on the  $x$ -axis, with three lines corresponding to voters of low, medium, and high religious attendance.

We then show Figure 76, which displays actual data from 2004.<sup>13</sup> This graph is consistent with the “postmaterialism” model, *not* with the “opiate of the masses” model.

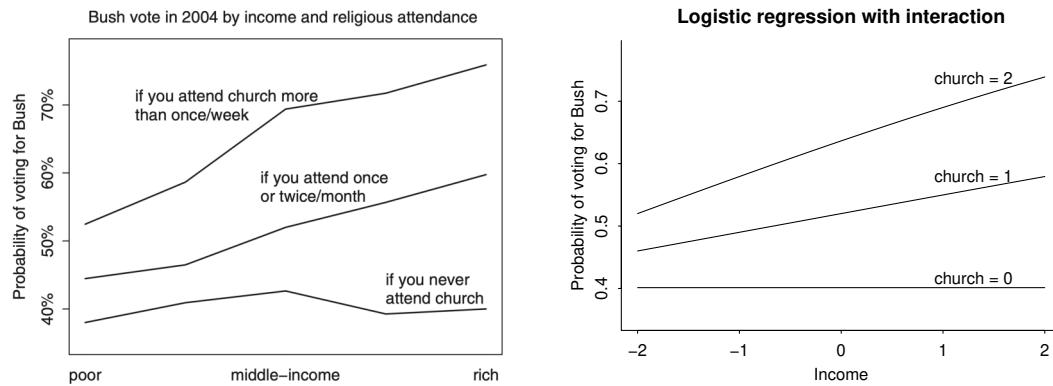
Figure 76a can be approximated by a logistic regression. What would that regression look like?

<sup>13</sup>From Andrew Gelman, David Park, Boris Shor, and Jeronimo Cortina (2009), *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*, Princeton University Press.

Two models of voting on social issues:

- *Opiate of the masses*: Rich people vote their interests; poor people vote “Gods, guns, and gays.”
- *Postmaterialism*: Poor people vote based on economics; rich people have the luxury to vote on social issues.

**Figure 75** Two conceptual models of voting on social issues. We write these on the board and then ask students to express them as logistic regressions of vote preference given income, religious attendance, and their interaction.



**Figure 76** (a) From survey data: probability of supporting the Republican presidential candidate in 2004 as a function of income, for three different levels of religious attendance. (b) Logistic regression model with interaction,  $\text{Pr}(\text{Republican support}) = \text{logit}^{-1}(b_0 + b_1 * \text{income} + b_2 * \text{church} + b_3 * \text{income} * \text{church})$ . We ask students to come up with coefficients  $b_0, b_1, b_2, b_3$  to reproduce this graph.

We'll work it out. But first, in a logistic regression predicting vote for Bush from income, church attendance, and their interaction, is the coefficient of the interaction positive, negative, or zero? We ask the students to consider this in pairs, and then we give the answer, which is that the interaction is positive, because the slope of the regression predicting Bush vote from income is higher for higher values of church attendance. Equivalently, the slope of the regression predicting Bush vote from church attendance is higher for higher values of income.

To put numbers on this story, our first step is to define the predictors. Let the income variable range from  $-2$  (poor) to  $2$  (rich) and the church attendance variable range from  $0$  (never attend) to  $2$  (attend more than once per week). Now we want to assign parameters to the following model:

$$\text{Pr}(\text{Republican support}) = \text{logit}^{-1}(b_0 + b_1 * \text{income} + b_2 * \text{church} + b_3 * \text{income} * \text{church}).$$

We work out the coefficients in steps:

- (a)  $\text{logit}^{-1}(b_0)$  is the probability of supporting the Republican when the two predictors both equal zero. From the graph, this is approximately 40%. So  $b_0$  is about  $-0.4$ . (Recall the divide-by-4 rule: if we wanted a probability of 50%, then  $b_0$  would be 0; going down from 50% to 40% is a drop of 0.10, or approximately 0.40 on the logistic scale.)
- (b)  $b_1$  is the difference in logit probability of supporting the Republican, comparing two people who differ by 1 in income when their value of church is 0. In this graph, the line when church is 0 is flat, so  $b_1$  is approximately zero.
- (c)  $b_2$  is the difference in logit probability of supporting the Republican, comparing two people who differ by 1 in church when their value of income is 0. From the graph, the difference between the different lines in the middle-income category (which is 0 according to our coding)

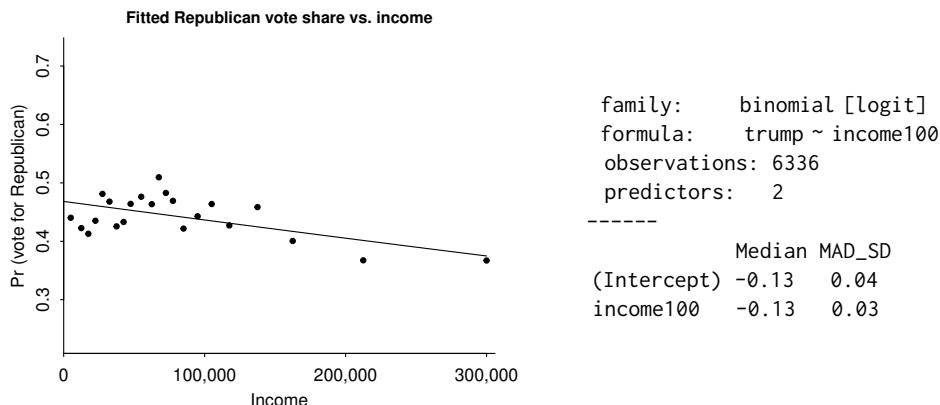


Figure 77 (a) From the 2020 American National Election Study, support for Donald Trump within each of 22 family income categories; (b) Fitted logistic regression predicting Trump support given income (in hundreds of thousands of dollars).

is approximately 12%, which corresponds to approximately 0.48 on the logistic scale. And the difference is positive, not negative, so  $b_1$  is about 0.6.

- (d) Finally,  $b_2$  corresponds to the difference in slopes for income, comparing values of church that differ by 1. The three lines in the graph have approximate slopes of 0, 0.03 (the middle line has a “rise” of about 0.12 divided by a “run” of 4), and 0.06 (the top line’s slope is approximately 0.24/4), so the interaction is approximately 0.03 on the probability scale, or 0.12 on the logistic scale.

Hence, the lines in Figure 76a roughly correspond to the logistic regression,

$$\text{Pr}(\text{Republican vote}) = \text{logit}^{-1}(-0.4 + 0 * \text{income} + 0.48 * \text{church} + 0.12 * \text{income} * \text{church}).$$

We graph this model in Figure 76b. It’s a pretty good match to the data!

How have things changed since 2004? We went to the American National Election Study and downloaded data from the 2020 survey.<sup>14</sup> This survey had five categories of religious attendance and 22 categories of family income, from \$0–\$10 000 (which we coded as \$5000) to \$250 000+ (which we coded as \$300 000).

We start by fitting a logistic regression of support for Donald Trump (among respondents who supported one of the two major-party candidates) given income. Figure 77 shows the raw data and the fitted model. Things have changed since 2004: now the richer voters are less likely to support the Republican candidate. However, this is just one data source. Other polls have found Trump doing slightly better among higher-income voters. About all we can say for sure is that the relation between income and voting in 2020 was fairly weak.

Next we include religion in the model. Figure 78 shows the fitted logistic regression as a table of estimates and standard errors and as a graph of curves. As in 2004, there is a positive interaction, indicating that religious attendance is a stronger predictor of Republican vote among richer voters.

Finally, we loop back to the raw data by making a separate plot for each category of religious attendance, as shown in Figure 79. We point out two things to the class. First, these lines are actually logistic curves; they appear to be straight lines just because their ranges are narrow. Second, the lines do not appear to go through the points for each graph. That is because the fitted model is linear in the two predictors, and there is no requirement that the data be linear. There is a

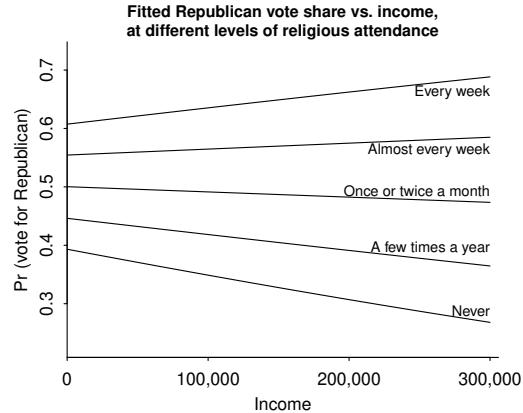
<sup>14</sup> American National Election Studies, <https://electionstudies.org/>.

#### 4. WEEK BY WEEK: THE SECOND SEMESTER

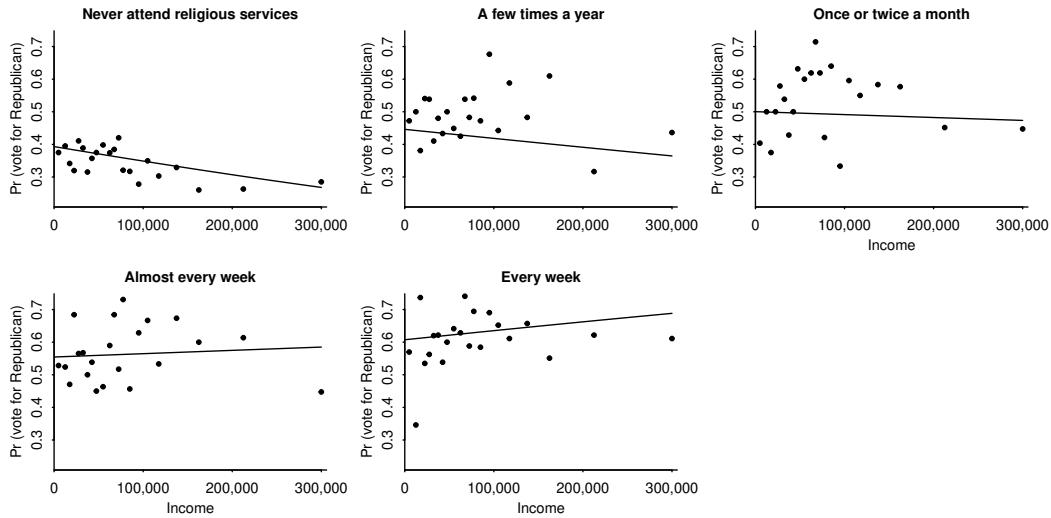
```

family:      binomial [logit]
formula:     trump ~ church +
             income100 + church:income100
observations: 6313
predictors:  4
-----
              Median  MAD_SD
(Intercept) -0.65   0.07
church       0.22   0.03
income100   -0.27   0.06
church:income100  0.08   0.02

```



**Figure 78** (a) Logistic regression predicting vote intention in 2020 given church attendance (on a 1–5 scale), family income (in hundreds of thousands of dollars), and their interaction; (b) Curves showing the fitted model.



**Figure 79** Fitted logistic regression from Figure 78 predicting vote intention in 2020 given church attendance, along with raw data, with separate graphs for each category of religious attendance.

big jump from the “never” category to the “few times a year” category which is not captured by the model that labels these as 1, 2, 3, 4, 5.

Given that we can plot the raw data, why fit the logistic regression at all? We can give three reasons:

- The fitted model with its four coefficients is a convenient summary of a complex pattern, and we can interpret each of the coefficients, indeed even more so if we center the predictors before fitting the model, as is shown in Figure 80.
- The standard errors give us a quantification of uncertainty about the fitted coefficients that would be difficult to get from the plot of the data.
- If we plan to compare estimates across states or countries or demographic groups or over time, it is helpful to have numerical summaries that can be plotted.

To demonstrate this last point, we display logistic regressions of Republican vote preference given income, religious attendance, and their interaction, fit separately to each election from 1952

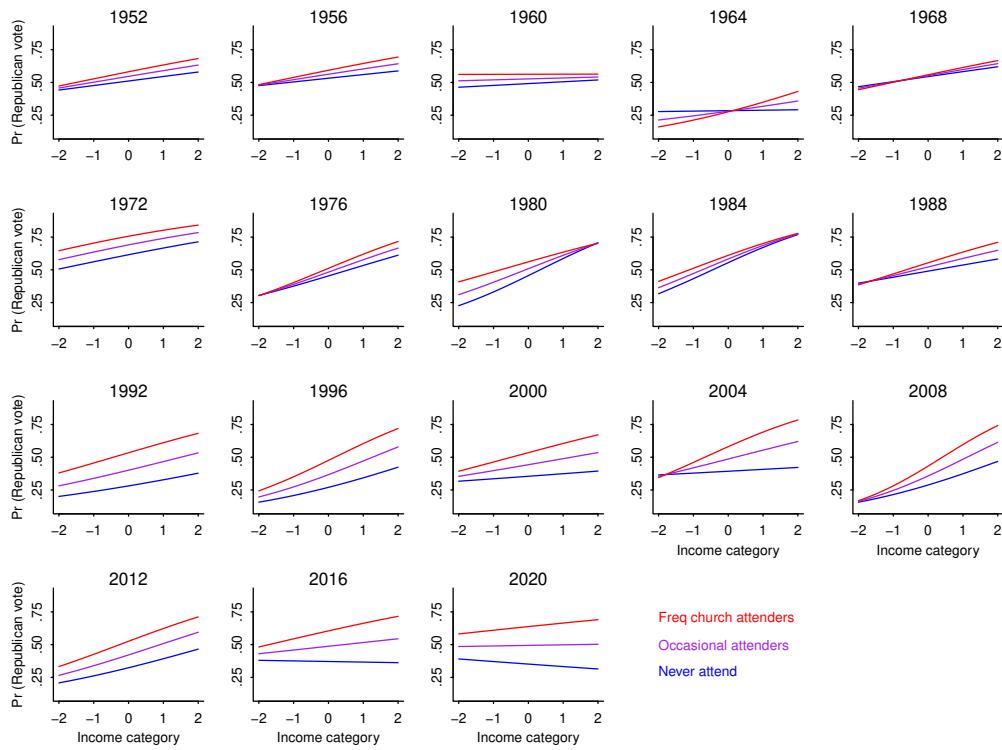
```

family: binomial [logit]
formula: trump ~ church +
          income100 + church:income100
observations: 6313
predictors: 4
-----
              Median MAD_SD
(Intercept) -0.65  0.07
church       0.22  0.03
income100   -0.27  0.06
church:income100  0.08  0.02
-----
```

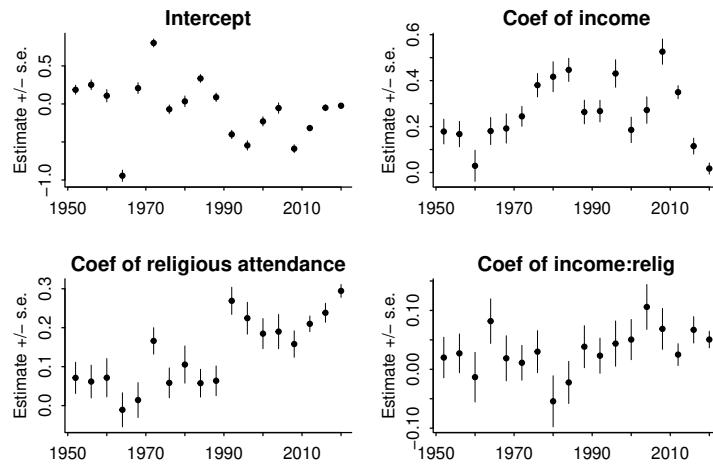
```

family: binomial [logit]
formula: trump ~ c_church +
          c_income100 + c_church:c_income100
observations: 6313
predictors: 4
-----
              Median MAD_SD
(Intercept) -0.25  0.03
c_church     0.28  0.02
c_income100 -0.10  0.04
c_church:c_income100  0.08  0.02
-----
```

**Figure 80** (a) Repeated from Figure 78a, logistic regression predicting Trump vote intention in 2020 given church attendance (on a 1–5 scale), family income (in hundreds of thousands of dollars), and their interaction; (b) Fit using centered predictors, so that the intercept and main effects are much easier to interpret.



**Figure 81** Logistic regressions fit separately to survey data from each U.S. presidential election, predicting Republican vote preference given income and religious attendance. Each predictor was coded on a  $-2$  to  $2$  scale, and the three lines on each plot correspond to religious attendance scores of  $2$ ,  $0$ , and  $-2$ . This display demonstrates the way in which a fitted regression model can be used as a data reduction tool that allows us to step back and see the bigger picture.



**Figure 82** Estimates and standard errors of the logistic regressions shown in Figure 81. Again, this shows the value of fitted models as data reduction.

through 2020: all the presidential elections for which the American National Election Study asked these questions. For these models, income is recorded in five categories (based on percentiles of family income) and is coded on a scale from  $-2$  to  $2$ , and religious attendance is also recorded in five categories, which we code from  $-2$  (never attend religious services except at weddings and funerals) to  $2$  (attend at least once per week). We show the fitted curves in Figure 81 and the time series of estimated coefficients and standard errors in Figure 82.

This story is relevant to the week's reading because it involves interpreting logistic regression fits and comparing them to data. It relates to the course as a whole as an example of statistical modeling for data reduction.

### Class-participation activities

#### 1. Job training programs and average predictive comparisons

Suppose a job training program is studied in an experimental context with pre-treatment employment history  $x$  (a continuous variable ranging from  $0$  for people who have never held a regular job to  $10$  for people who have had continuous employment during their working years), a treatment indicator  $z$ , and an outcome  $y$  that equals  $1$  for people who are employed at the end of the study or  $0$  otherwise.

Further suppose that the true pattern of the data can be fit using a logistic regression of  $y$  on  $x$  and  $z$  with no interactions, and that the coefficient of  $z$  is . . . what would it take for the effect to be about  $10\%$ , we ask the class? From the divide-by-4 rule, the coefficient will be  $0.4$ . Further suppose that, in the absence of the treatment, the probability of being employed at the end of the study ranges from  $5\%$  (when  $x = 0$ ) to  $50\%$  (when  $x = 5$ ) to  $95\%$  (when  $x = 10$ ). From that we can figure out the intercept coefficient for  $x$ ; students should do this in pairs, using R to figure it out. Because the probability is  $50\%$  when  $x = 5$ , we can write the model (in the absence of treatment) as  $\text{Pr}(y = 1) = \text{logit}^{-1}(b * (x - 5))$ . Then we just solve for  $b$  by working out the probability when  $x = 10$ : we must have  $0.95 = \text{logit}^{-1}(b * 5)$ , so  $b = \text{logit}(0.95)/5 = 0.59$ . So the model in the absence of treatment is  $\text{Pr}(y = 1) = \text{logit}^{-1}(0.59 * (x - 5)) = \text{logit}^{-1}(-2.95 + 0.59x)$ . Throwing in the coefficient for the treatment, we get  $\text{Pr}(y = 1) = \text{logit}^{-1}(-2.95 + 0.59x + 0.4z)$ .

How to think about this model? We ask students to sketch (with pen on paper) the two curves of  $\text{Pr}(y = 1)$  vs.  $x$  corresponding to  $z = 0$  and  $z = 1$ . One way to think about these two coefficients is that a difference of  $1$  in  $z$  represents the same predictive difference as a difference in  $x$  of . . . [we

pause while students work this out] . . .  $0.4/0.59 = 0.68$ . So getting the treatment is comparable to moving up by 0.68 on the  $x$  scale.

What, then, is the treatment effect on the probability scale? How much does the treatment increase your probability of getting a job? The answer is that it depends on your pre-treatment value of  $x$ . That's the predictive comparison discussed in Section 14.4 of *Regression and Other Stories*. We ask students to consider average predictive comparisons for different hypothetical groups.

Now suppose an experiment was being designed to study the efficacy of the job training program. What values of  $x$  would you like to include in the study? We ask students to discuss in pairs; the answer is that, under the model we've specified, you're most likely to find a clear effect if you choose for your experiment people with values of  $x$  near 0.5.

If the study is successful, what would this imply about the generalization of the experimental findings to the real world? We again have students discuss in pairs. The answer to the question is that, first, if the logistic model is correct, you'd expect to see smaller average effects in the real world as the treatment gets applied to people outside the "sweet spot" of maximum effect; second, you wouldn't really know how large the effect is outside that central region, as your inference would be entirely subject to extrapolation. If there is time, students can discuss the implications of these findings for other problems such experiments on medical treatments.

This activity is relevant to the week's reading on average predictive comparisons and relates to the course as a whole by considering the questions of data collection, experimental design, inference from data, and real-world extrapolation.

## 2. Logistic regressions with interactions

Figures 76, 78, and 79 show positive interactions of income and religion predicting Republican vote preference. We display Figure 78 on the screen and task the students to form small groups and in each group come up with a different real-world example of an interaction. We then discuss a few of these with the class as a whole, sketching hypothesized regressions and discussing how one could gather data to fit these models. The key to getting this activity to work is to get dirty with the numbers: for well-defined examples, students should hypothesize curves of the form  $\Pr(y = 1) = \text{logit}^{-1}(b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2)$  with specified values of  $b_0, b_1, b_2, b_3$  and then draw these curves on the board, with each model displayed as two curves (for example,  $\Pr(y = 1)$  vs.  $x_1$  for two different values of  $x_2$ ). Other students can then disagree with the details—or, if not, the instructor can ask the students what sort of data would be required for the students to decide that their guessed curves are wrong? The idea is to understand, not just the logistic model in general, but particular instantiations of this model, here with the additional challenge of an interaction.

This activity relates to the week's reading by exploring logistic regression with interactions, and it relates to the course as a whole by focusing on understanding models quantitatively, being able to connect a general model with reasonable specific numbers.

## Computer demonstrations

### 1. Predictive distributions from a fitted logistic regression

We fit a logistic regression of voter registration status on sex and age, using survey data from Pew Research:

```
# Read data from here: https://github.com/avehtari/ROS-Examples
library("foreign")
library("dplyr")
library("rstanarm")
pew <- read.dta(paste0(
```

```
"https://raw.githubusercontent.com/avehtari/",
"ROS-Examples/master/Pew/data/",
"pew_research_center_june_elect_wknd_data.dta"
))
pew <- pew %>% select(c("age", "regicert", "sex")) %>%
  na.omit() %>% filter(age != 99)
pew$registered <- ifelse(pew$regicert=="absolutely certain", 1, 0)

# Estimate logistic model (takes some time)
fit <- stan_glm(registered ~ age + sex, family=binomial(link="logit"),
  data=pew, refresh=0)
print(fit, digits=4)

# Get estimated probability (using function, and by hand)
new <- data.frame(age=50, sex="male")
predict(fit, type="response", newdata=new) # function
invlogit(coef(fit)[1] + coef(fit)[2]*50) # by hand

# Linear predictor (X*beta) with uncertainty
linpred <- posterior_linpred(fit, newdata=new)
hist(linpred)

# Expected outcome (p) with uncertainty
hist(invlogit(linpred)) # One way
epred <- posterior_epred(fit, newdata=new) # Other way
hist(epred)
abline(v = c(mean(epred), mean(epred) + c(-2, 2)*sd(epred)))

# Predictive distribution (0 or 1) for a new observation
postpred <- posterior_predict(fit, newdata=new)
mean(postpred)
```

This can be extended by making comparisons between predictions for new observations (for example, male/female, different ages).

## 2. Linear or logistic regression with binary outcomes

When probabilities are far from 0 or 1, logistic and linear regressions should yield similar results for coefficient estimates and standard errors (after dividing by 4), with the big difference coming in predictions. The linear model starts to fall apart at the boundaries.

We demonstrate with some experimentation, first setting up a well-behaved example for which linear and logistic regressions give similar results, then pushing toward the boundary to see what happens.

```
library("rstanarm")
n <- 100
x <- runif(n, 0, 1)
a <- 0.5
b <- -1
y <- rbinom(n, 1, invlogit(a + b*x))
fake <- data.frame(x, y)
fit_logistic <- stan_glm(y ~ x, family=binomial(link="logit"),
  data=fake, refresh=0)
print(fit_logistic, digits=2)
fit_linear <- stan_glm(y ~ x, data=fake, refresh=0)
print(fit_linear, digits=2)
```

We ask students in pairs to interpret the intercept and slope. The slope of the linear fit is about  $\frac{1}{4}$  that of the logistic regression, for reasons discussed in Section 13.2 of *Regression and Other Stories*. The intercepts can be understood by considering the values of the functions at  $x = 0$ .

Next we plot the data and fitted lines:

```
plot(x, y)
curve(invlogit(coef(fit_logistic)[1] + coef(fit_logistic)[2]*x),
      add=TRUE, col="blue")
curve(coef(fit_linear)[1] + coef(fit_linear)[2]*x, add=TRUE, col="red")
```

But what happens when we extend the range of the model? Change the coefficients and rerun the above code. Let's first put it into a function:

```
compare_linear_logistic <- function(n, a, b){
  x <- runif(n, 0, 1)
  y <- rbinom(n, 1, invlogit(a + b*x))
  fake <- data.frame(x, y)
  fit_logistic <- stan_glm(y ~ x, family=binomial(link="logit"),
    data=fake, refresh=0)
  print(fit_logistic, digits=2)
  fit_linear <- stan_glm(y ~ x, data=fake, refresh=0)
  print(fit_linear, digits=2)
  plot(x, y, main=paste("y = invlogit(", a, " + ", b, "x)", sep=""))
  curve(invlogit(coef(fit_logistic)[1] + coef(fit_logistic)[2]*x),
        add=TRUE, col="blue")
  curve(coef(fit_linear)[1] + coef(fit_linear)[2]*x, add=TRUE, col="red")
}
```

And now we can look at some examples, starting with what we did before:

```
compare_linear_logistic(100, 0.5, -1)
```

And then try some others, for example:

```
compare_linear_logistic(100, 2, -1)
compare_linear_logistic(100, 5, -10)
```

## Drills

### 1. Expected probabilities in logistic regression

Consider the well-switching model in Section 14.2 of *Regression and Other Stories*:

	Median	MAD_SD
(Intercept)	-0.22	0.09
dist100	-0.90	0.11
arsenic	0.47	0.04
educ4	0.17	0.04

For the following cases, calculate the difference in expected probabilities of well-switching and interpret your findings:

- (a) Two people both of whom are 50 meters away from well (`dist100 = 0.5`) and whose arsenic levels both are 2 (`arsenic = 2`), but who differ by education, with 4 and 8 years of education, respectively (`educ_4 = 1` and `educ_4 = 2`)

*Solution:*

- Using the divide-by-4 rule:  $0.17 * (2 - 1)/4 = 0.04$ . The model predicts that the higher-educated person is 4 percentage points more likely to switch.

- Exact computation using the logistic function:  $\text{logit}^{-1}(-0.22 - 0.90 * 0.5 + 0.47 * 2 + 0.17 * 2) - \text{logit}^{-1}(-0.22 - 0.90 * 0.5 + 0.47 * 2 + 0.17 * 1) = 0.65 - 0.61 = 0.04$ .

In this case, the predicted probabilities are close enough to 0.5 that the divide-by-4 rule works well for approximating the difference in probabilities between the two cases. When there is doubt, you can always do the exact computation.

- (b) Shared: 50 meters away from well ( $\text{dist100} = 0.5$ ), arsenic level 2 ( $\text{arsenic} = 2$ ); Different: education levels of 12 and 16 years, respectively ( $\text{educ\_4} = 3$  and  $\text{educ\_4} = 4$ )
- (c) Shared: 5 (median) years of education ( $\text{educ\_4} = 1.25$ ), 1.3 (median) level of arsenic ( $\text{arsenic} = 1.3$ ); Different: distance from well of 0 and 500 meters, respectively ( $\text{dist100} = 0$  and  $\text{dist100} = 5$ )

## 2. Understand the logit function

For models with the following parameters, sketch and write the formula for the corresponding  $\text{logit}^{-1}(a + bx)$  function. When you are done, check your results visually in R.

- (a) Halfway point = 0; steepest slope = 1  
*Solution:*  $\text{logit}^{-1}(4 * (x - 0)) = \text{logit}^{-1}(4x)$ ; that is,  $a = 0, b = 4$ .
- (b) Halfway point = 0; steepest slope = -2
- (c) Halfway point = 2; steepest slope = -2
- (d) Halfway point = -50; steepest slope = 0.34

## Discussion problems

### 1. Experimental design for logistic regression

Suppose a certain disease has a 20% mortality rate, and a new drug is hypothesized to reduce the mortality rate to 10%. Frame this as a logistic regression, and suppose a randomized experiment is performed with  $n/2$  people getting the treatment and  $n/2$  getting the control. How large must  $n$  be so that the uncertainty in the estimated treatment effect is low enough that we can be nearly certain of correctly identifying its beneficial effect? To figure this out, start by guessing  $n$ , then figure out the standard error of the corresponding experiment, then adjust  $n$  up or down as necessary to get a design with the desired standard error.

### 2. Experimental design for logistic regression with pre-treatment predictor

Consider estimating the effect of a job-training program ( $z = 1$  if a person gets the program and  $z = 0$  otherwise) on a binary outcome  $y$  corresponding to whether the person is employed within six months after the program, and a continuous pre-treatment predictor  $x$  on a 1–5 scale representing employment history. Suppose that, in the absence of the program, the probability of employment in six months ranges from 20% for people on the low end of the scale to 70% for people on the high end, and further suppose that the treatment increases these probabilities to 35% and 70%. Express this hypothesized pattern as a logistic regression model with an interaction.

Next suppose this experiment is conducted with  $n/2$  people getting the treatment and  $n/2$  getting the control, with random assignment to a group of people whose values of  $x$  are uniformly distributed between 1 and 5. How large must  $n$  need to be so that the uncertainty in the estimated treatment effect is low enough that we can be nearly certain of correctly identifying its beneficial effect? To figure this out, start by guessing  $n$ , then figure out the standard error of the corresponding experiment, then adjust  $n$  up or down as necessary to get a design with the desired standard error.

With this value of  $n$ , how accurately can the interaction be estimated?

## 4.17 Other generalized linear models

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Patterns of gun ownership	How similar are you to your friends?	Simulating overdispersed data	Coefficients in negative binomial regression	Identification in linear models
Structure in social networks	Alternative models for discrete data	Generalized linear model with offset	Parameters in ordered logistics regression	Functional forms for nonlinear models

### Reading

Chapter 15 of *Regression and Other Stories*: Other generalized linear models

### Pre-class warmup assignments

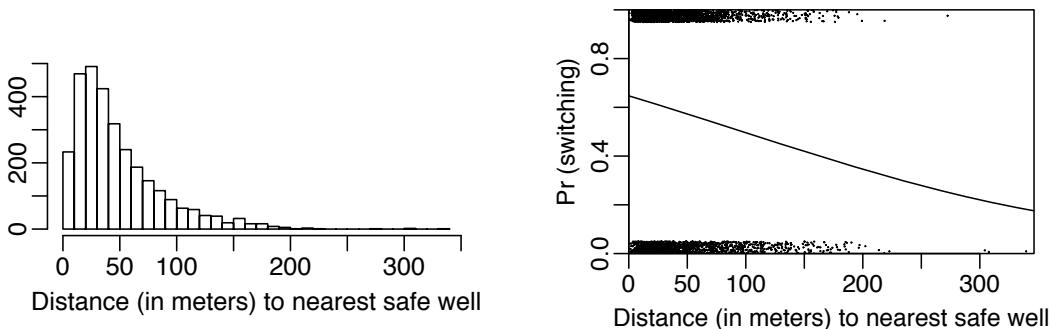
1. Negative binomial regression in R
  - (a) Simulate data from a negative binomial regression model with  $E(y|x) = \exp(a + bx)$  and reciprocal dispersion parameter 2, with 100 data points, an intercept of 0.3 and a slope of 0.4, and values of  $x$  randomly drawn from the range (0, 5).
  - (b) Fit a negative binomial regression to these data. Compare the estimates and standard errors of the three parameters to the true values of these parameters.
  - (c) Make a graph displaying the simulated data and fitted regression curve.
2. Ordered logistic regression in R
  - (a) Simulate data from an ordered logistic regression model where  $y$  can take on the values 1, 2, 3, 4, or 5, and where  $\Pr(y > k | x) = \text{logit}^{-1}(a + bx - c_k)$ , for  $k = 1, \dots, 4$  and with  $0 = c_1 < c_2 < c_3 < c_4$ . The model thus has five parameters:  $a, b, c_2, c_3$ , and  $c_4$ . Simulate 50 data points from the model with  $a = -5, b = 2, c_2 = 1, c_3 = 2, c_4 = 5$  and values of  $x$  drawn at random from the interval (0, 5).
  - (b) Fit an ordered logistic regression to these data. Compare the estimates and standard errors of the three parameters to the true values of these parameters.
  - (c) Make a graph displaying the simulated data and fitted regression curves.
  - (d) Label where on this plot is the predicted probability that  $y = 4$ , given  $x$ . Explain.

### Homework assignments

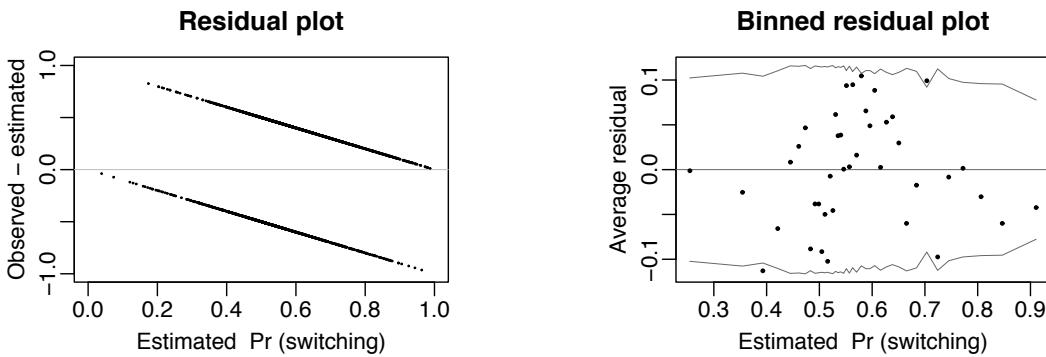
1. (a) Graphing logistic regression (Exercise 14.3 of *Regression and Other Stories*)

The well-switching data described in Section 13.7 of *Regression and Other Stories* are in the folder Arsenic.

- i. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.
- ii. Make a graph similar to Figure 83b displaying  $\Pr(\text{switch})$  as a function of distance to nearest safe well, along with the data.
- iii. Make a residual plot and binned residual plot as in Figure 84.
- iv. Compute the error rate of the fitted model and compare to the error rate of the null model.
- v. Create indicator variables corresponding to  $\text{dist} < 100$ ;  $\text{dist}$  between 100 and 200; and



**Figure 83** (a) Histogram of distance to the nearest safe well, for each of the unsafe wells in our data from Araihazar, Bangladesh. (b) Graphical expression of the fitted logistic regression,  $\text{Pr}(\text{switching wells}) = \text{logit}^{-1}(0.61 - 0.62 * \text{dist100})$ , with (jittered) data overlaid. The predictor  $\text{dist100}$  is  $\text{dist}/100$ : distance to the nearest safe well in 100-meter units.



**Figure 84** (a) Residual plot and (b) binned residual plot for the well-switching model. The strong patterns in the raw residual plot arise from the discreteness of the data and inspire us to use the binned residual plot instead. The bins are not equally spaced; rather, each bin has an equal number of data points. The light lines in the binned residual plot indicate theoretical 95% error bounds.

`dist > 200`. Fit a logistic regression for  $\text{Pr}(\text{switch})$  using these indicators. With this new model, repeat the computations and graphs for part (a) of this exercise.

2. (a) Poisson and negative binomial regression (Exercise 15.1 of *Regression and Other Stories*)

The folder `RiskyBehavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was “number of unprotected sex acts.”

- Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?
- Extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?
- Fit a negative binomial (overdispersed Poisson) model. What do you conclude regarding effectiveness of the intervention?

- iv. These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions?
- (b) Offset in a Poisson or negative binomial regression (Exercise 15.2 of *Regression and Other Stories*)  
Explain why putting the logarithm of the exposure into a Poisson or negative binomial model as an offset, is equivalent to including it as a regression predictor, but with its coefficient fixed at the value 1.

## Stories

### 1. Understanding gun ownership using generalized linear models

From a recently published article in the criminal justice literature:

“The gun ownership literature is vast, with dozens of studies seeking to explain who owns guns and why. We build on this literature in two key ways. First, we introduce a new variable into the fold: moral concern about harming others. We theorize that this concern actively inhibits gun ownership. Second, we direct theoretical and empirical attention to a predictor of gun ownership that has frequently been overlooked in the contemporary gun literature: childhood socialization. Using data from a national sample of 1100 adults, we find that moral concerns about harm represent a barrier to gun ownership and limit the number of guns people own . . .”

To measure Moral Harm, we used four items: ‘If we saw a mother slapping her child, we would be outraged’; ‘Compassion for those who are suffering is the most crucial virtue’; ‘It can never be right to kill a human being’; ‘The government must first and foremost protect all people from harm.’ These items are drawn from previous measures . . . We averaged the four items to create a Moral Harm index . . .

We measure Childhood Socialization using a five-item additive index that gauges both gun presence and gun socialization experiences. Specifically, the respondents were asked if when they were growing up, their family members or guardians did any of the following: ‘Keep a firearm in the house’; ‘Teach you how to shoot a firearm’; ‘Teach you how to clean a firearm’; ‘Take you hunting’; and ‘Take you to a gun show.’”<sup>15</sup>

The questions being asked are all on a 1–5 scale (from strongly agree to strongly disagree), so the averages are on a 1–5 scale as well. The researchers measure gun ownership by asking people how many guns (“none,” “one,” “two,” “three,” or “four or more”) they own of each of four types (shotgun, rifle, assault rifle, and handgun), and they record “four or more” as 4, so that the total is somewhere between 0 and 16. They report:

“Around 36% of the sample own a gun. Of these gun owners, the majority own multiple guns: around 70% own two or more guns and the average number of guns owned is greater than four.”

They fit logistic regressions predicting whether a respondent owns a gun and negative binomial regressions predicting the number of guns owned. The results are summarized in Figure 85, which the instructor can project onto the screen to start a class discussion.

Start with the logistic regression. Students can work in pairs going through the two model fits and explaining to each other the coefficients in each model. They can give numerical interpretations using the usual divide-by-4 rule. For example, in the first regression, comparing two people who differ by 1 on the Moral Harm scale but are the same in ethnicity, sex, age, education, income,

<sup>15</sup>Nathaniel Schutten, Justin Pickett, Alexander Burton, Francis Cullen, Cheryl Lero Jonson, and Velmer Burton (2021), Understanding gun ownership in the twenty-first century: Why some Americans own guns, but most do not, *Justice Quarterly*, <https://osf.io/preprints/socarxiv/t4cf7/>

Variables	Gun Ownership				Gun Quantity			
	Model 1		Model 2		Model 3		Model 4	
	b	SE	b	SE	b	SE	b	SE
Moral Harm	-.543***	.104	-.400**	.146	-.565***	.090	-.325**	.103
Childhood Socialization	—	—	.666***	.057	—	—	.544***	.040
Fear of Crime	—	—	.003	.098	—	—	.035	.062
Victimization	—	—	.354	.223	—	—	.344*	.165
Dangerous World Beliefs	—	—	.193	.113	—	—	.129	.087
Confidence in the Police	—	—	-.149	.116	—	—	-.117	.082
Confidence in the Government	—	—	.112	.101	—	—	.157*	.068
Racial Resentment	—	—	-.416**	.124	—	—	-.371***	.090
Southerner	—	—	.760***	.182	—	—	.173	.140
Rightward Political Views	—	—	.486**	.181	—	—	.572***	.137
Born-Again Protestant	—	—	.125	.257	—	—	-.160	.161
Religiosity	—	—	.004	.132	—	—	-.007	.104
White	.254	.206	.208	.225	.346*	.161	.354*	.165
Male	.374*	.168	-.039	.196	.493**	.143	-.008	.142
Age	.004	.006	.002	.006	-.002	.005	.004	.004
Education	-.207***	.059	-.234**	.078	-.106	.055	-.134*	.062
Income	.047	.031	.075*	.037	.036	.029	.066*	.029
Married	.563**	.195	.456*	.203	.374*	.156	.201	.153
Child in Household	.057	.193	-.094	.227	.239	.184	.158	.183
N		1,065		1,051		1,057		1,044

**Figure 85** Regressions predicting gun ownership (logistic) and gun quantity (negative binomial). The instructor jump go through these fitted models and discuss some of the coefficients with the class. One, two, and three asterisks correspond to coefficients with p-values less than 0.05, 0.01, and 0.001, that is, where the 95%, 99%, and 99.9% intervals exclude zero. We do not recommend this sort of significance-level reporting, but we keep it here just so students can see how regressions are often summarized in practice.

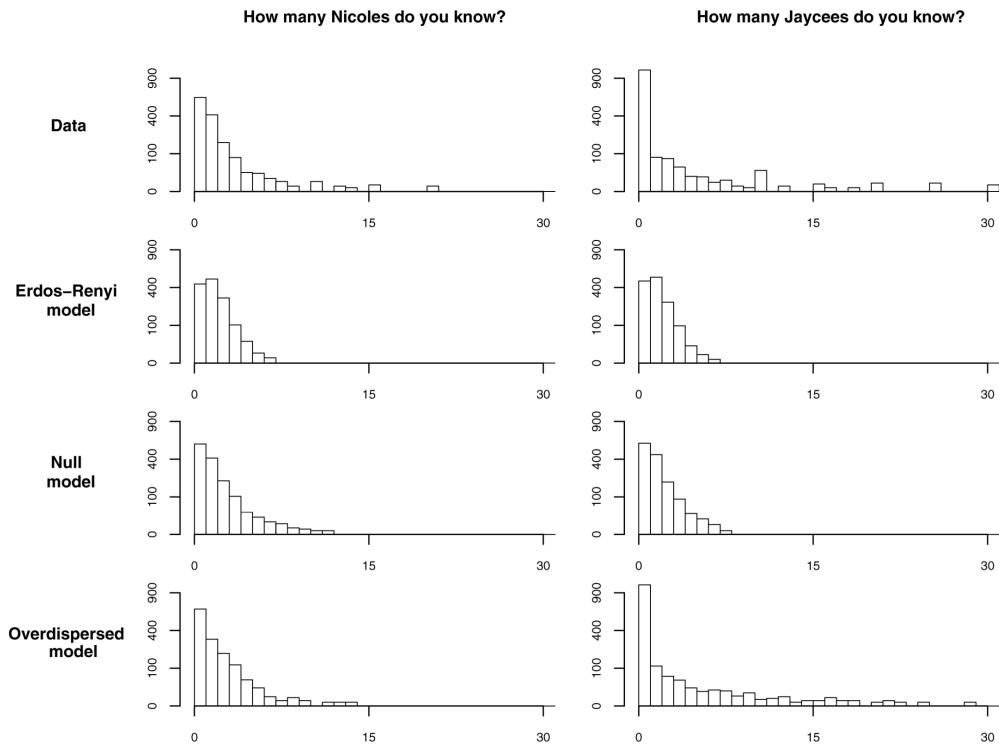
and marital and parental status, the person who is 1 point higher on the Moral Harm scale is approximately  $0.543/4 = 0.13$  or 13 percentage points less likely to own a gun.

The real challenge comes when interpreting a comparison of people who differ by one unit in some predictor while being identical in all others. For example, how do you think about a comparison of a man and a woman who are identical on all other predictors, including Moral Harm? One might suspect that women tend to be higher on the Moral Harm scale, so we’re comparing atypical members of the two sexes.

Further difficulties arise in the second regression, where we are invited to compare people who disagree on one set of survey questions but are identical in all the others. The authors of the paper write that it is likely that “the effect of gender operates indirectly through some other significant variable in the model, such as Childhood Socialization.” The evidence here is that the coefficient for Male is large in model 1 and near zero in model 2, after adjusting for Childhood Socialization and other variables. Unfortunately, it’s not so easy to make such a statement—indeed, it’s not clear what a statement such as “operates indirectly” really means. We will return to this point in the causal inference chapter.

We can then follow a similar exercise, first with students working in pairs and then convoking a class discussion, on the negative binomial models for predicting number of guns. The coefficients can now be interpreted on a logarithmic scale. For example, in model 3, the coefficient of  $-0.565$  for Moral Harm corresponds to a predicted difference of this amount in the logarithm of number of guns and thus a multiplication by  $\exp(-0.565) = 0.57$  in the expected number of guns; that is, you would expect, on average, 57% as many guns for someone who is 1 point higher on that scale, comparing to someone who is identical on all other predictors. Or, if you want a 95% confidence interval,  $\exp(-0.565 \pm 2 * 0.090) = (0.47, 0.68)$ , so that you would expect, on average, between 47% and 68% as many guns. Again, the challenges come when interpreting comparisons between people who are identical on all but one predictor.

This interpretation in terms of comparisons is so difficult that it would be good to have some other way of understanding the models in Figure 85. A causal story would be tempting (for example, in model 3, “Being married causes you, on average, to have 20% more guns”), but we won’t allow



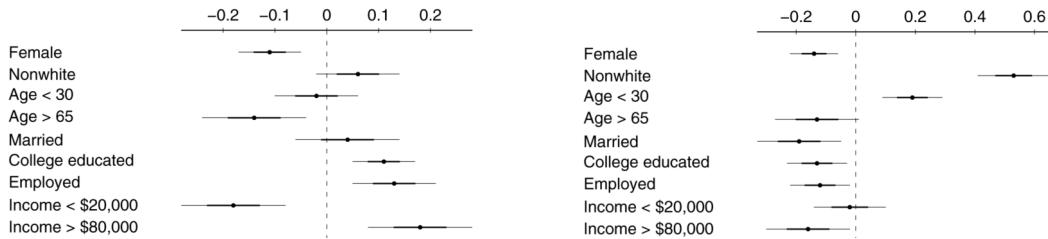
**Figure 86** Histograms (on the square-root scale) of responses to “How many people do you know named Nicole?” and “How many Jaycees do you know?” from a national survey, along with random simulations under three fitted models: The Erdős–Renyi model (completely random links), a null model (a Poisson regression), and an overdispersed (negative-binomial) regression. Each model predicts more variation than the one above, with the overdispersed model fitting the data reasonably well. The propensities to form ties to Jaycees show much more variation than the propensities to form ties to Nicoles, hence the Jaycees counts are much more overdispersed.

that! A different way of interpreting the regression is just as a model, a mathematical formula: plug in the predictors and you get a prediction.

This example is relevant to the week’s reading in giving students a chance to interpret coefficients in two related generalized linear models. It is relevant to the course as a whole by being an example of the use of multiple regression as a way of understanding observational data, without any particular causal goal. The problem remains open-ended, and we do not see the results in Figure 85 as definitive; indeed, it’s not clear what it would mean to have a definitive analysis here. Nonetheless, we can learn from these regressions. The class can discuss what sorts of analyses could clarify what is happening, what questions that might be asked here, and what future data would be worth gathering to answer these questions.

## 2. Using overdispersion in count data to estimate structure in social networks

How many people do you know? It depends on how the question is asked. One way that researchers have measured the size of people’s social networks is to ask them how many people they know named Nicole, or Christopher, or Anthony, or Stephanie. To illustrate, suppose that a student knows two people named Nicole, and 0.13% of Americans are named Nicole. Extrapolating to the entire population yields an estimate of  $2/0.0013 = 1530$  for the size of that student’s social network. A more precise estimate can be obtained by averaging these estimates using a range of different names. This is only a crude inference, but as an estimation procedure, it has the advantage of not requiring respondents to recall their entire networks.



**Figure 87** (a) Estimated coefficients  $\pm 1$  and 2 standard errors from regression of estimated log gregariousness parameters on personal characteristics; (b) estimated coefficients  $\pm 1$  and 2 standard errors from regression of residuals from the “How many males do you know incarcerated in state or federal prison?” question on personal characteristics.

The later rows of graphs in the figure display the distribution of each response under different statistical models fit to data:

- The Erdős-Renyi model,  $y_{ij} \sim \text{Poisson}(\beta_j)$ , where  $i$  is the respondent and  $j$  is the survey item. This is a model in which Nicols and Jaycees are randomly distributed in the social network and the two groups are allowed to have different sizes (hence the parameters  $\beta_1$  and  $\beta_2$ ). After fitting this model to the data and estimating  $\beta_1, \beta_2$ , we simulate 1370 responses at random from this distribution using these fitted parameters.
- The null model,  $y_{ij} \sim \text{Poisson}(\alpha_i + \beta_j)$ , which still assumes random mixing, but now allows for the gregariousness  $\alpha_i$  to vary across people. After fitting this model to the data, we simulate 1370 responses at random from this distribution allowing the  $\alpha_i$ 's to vary.
- The overdispersed model,  $y_{ij} \sim \text{negative binomial}(\alpha_i + \beta_j, \phi_j)$ . The parameter  $\phi_j$ , which varies by group, allows for a clustering or nonrandom distribution of the group within the social network. The limit  $\phi \rightarrow \infty$  corresponds to the Poisson distribution, and lower values of  $\phi$  represent clustering or overdispersion.

Descending from the second to fourth rows of Figure 86, we see increasingly accurate fits to the data. The Jaycees in particular need an overdispersed model, which makes sense given that they are a social organization, so if you know one Jaycee you’re likely to know many.

Most of our effort in this research project was spent in understanding the differences between groups, and how we can use the observed overdispersion to learn about the social network.

For the purpose of this story, though, we talk about the other part of our model—the gregariousness parameters  $\alpha_i$ . Figure 87a shows the estimated coefficients from a linear regression of the estimates of  $\log \alpha_i$  on characteristics of the survey respondents. Because the regression is on the logarithmic scale, the coefficients (with the exception of the constant term) can be interpreted as proportional differences: thus, with all else held constant, women have social network sizes 11% smaller than men, people over 65 have social network sizes 14% smaller than others, and so forth. The  $R^2$  of the model is only 10%, indicating that these predictors explain little of the variation in gregariousness in the population.

Similar regressions can be performed on the residuals from the regression. For each item  $j$ , we computed the residuals  $r_{ij} = y_{ij} - \exp(\hat{\alpha}_i + \hat{\beta}_j)$  and regressed them on individual characteristics. We did this for all 32 groups asked about in our survey. Different groups showed different patterns. One particularly interesting group is men in prison. The estimated regression coefficients from the residuals for that question are shown in Figure 87b. Being male, nonwhite, young, unmarried, and so on are associated with knowing more men than expected in prison. However, the  $R^2$  of the regression is only 11%, indicating that most of the variation in the data is not captured by these predictors. The plots in Figure 87 also demonstrate how helpful it can be to see coefficients graphed rather than presented as numbers in a table.

This story relates to the week's reading as an example of negative binomial regression for overdispersed count data. It is an interesting example in which the overdispersion is of interest for its own sake and is not just an annoyance to be adjusted for. Unfortunately, this model cannot be fit using `stan_glm`—the difficulty is that the overdispersion parameter needs to vary by item, and so we coded the model directly in Stan—but you can still use this example to understand some of the concepts of generalized linear modeling. The story is relevant to the course as a whole in showing how we can use survey data to learn about social structure.

### Class-participation activities

#### 1. How similar are you to your friends?

Political scientists and sociologists are interested in similarities within social networks: to what extent are you similar to your friends? We can assess this with a set of survey questions that have been used in social research.

The first step is to ask students to pick some questions to ask about each other. Aim for 10 questions, including some political attitudes (for example, support for the death penalty, public funding for charter schools, etc.) and some personal views (for example, preference for cats vs. dogs, coffee vs. tea, etc.) We come up with a consensus set of questions (each with a binary outcome) and write them on the board. Then we ask students to divide themselves into pairs and ask each pair to select a unique ID number, for example by going into R and typing `round(runif(1, 0, 1e6))` which will allow confidentiality of responses. Next we set up a Google form for each student to fill out with 22 items: their pair's ID number, the student's responses to the 10 questions, and the student's guess of the responses of the other student in the pair, and a measure of how well the student knows the partner (on a scale of 1 = don't know the person at all, 2 = seen the person around, 3 = acquaintance, 4 = friend, 5 = close friend).

The goal is to see how similar people are to their friends and also how similar people think their friends are to them. When filling out the form, students should honestly report their own attitudes on the 10 questions and do their best to guess their partners' attitudes. We emphasize that all responses will be confidential; the pair ID numbers will allow us to match up students in pairs but not to identify who the students are.

Usually when we ask students to work in pairs, if there is an odd number, the extra student can join a pair and form a group of three. In this one activity, however, if there is an odd number, one of the students will just have to observe without participating.

Once the students have entered their data, the instructor can download the file and do some operations in R to arrange the people in pairs and extract their responses. In the code, we arrange the people in pairs based on their ID numbers and define `ego` as the answers to the ten questions and `guess_for_alter` as the guesses of their partners' responses:

```
responses <- read.csv("Similarity.of.friends.csv")
pair_id <- responses[, "Pair.ID.number"]
print(table(pair_id))
responses <- responses[order(pair_id),]
J <- 10
n <- nrow(responses)
alter_index <- 1:n + rep(c(1, -1), n/2)
ego <- as.matrix(responses[, paste("Q", 1:J, sep="")])
guess_for_alter <-
  as.matrix(responses[, paste("Guess.of.partner.s.response.to.Q", 1:J, sep="")])
alter <- ego[alter_index,]
```

For each student, we construct two outcomes:

- `n_agree`, the number of the 10 questions for which the two students had the same opinion
- `n_perceived`, the number of the 10 questions for which there was perceived agreement (that is, the number of questions where person A thought that person B would agree with them):

```
n_agree <- rowSums(ego==alter)
n_perceived <- rowSums(ego==guess_for_alter)
```

And we create one predictor:

- `close`, the average of the two students' assessments of how well they know each other (thus, a number between 1 and 5)

We can use the `recode` function in R's `dplyr` package to conveniently convert the responses into numeric categories:

```
library("dplyr")
close_to_alter <- recode(responses[, "How.well.do.you.know.your.partner."],
  "Not at all" = 1, "Seen each other around" = 2, "Acquaintance" = 3,
  "Friend" = 4, "Close friend" = 5)
close <- (close_to_alter + close_to_alter[alter_index])/2
```

We then fit regressions predicting each of the two outcomes given the predictor. Each outcome can be viewed as a number of successes out of a known number of trials, so we use binomial regression:

```
friends_data <- data.frame(n_agree, n_perceived, close)
fit_agree <- stan_glm(cbind(n_agree, J - n_agree) ~ close,
  family=binomial(link="logit"), data=friends_data, subset=seq(2,n,2), refresh=0)
fit_perceived <- stan_glm(cbind(n_perceived, J - n_perceived) ~ close,
  family=binomial(link="logit"), data=friends_data, refresh=0)
```

The `fit_agree` regression selects the data from every second person because otherwise we would be counting each observation twice.

As usual, the instructor should ask students to discuss what the results might be before displaying the outcomes of the regressions. The next step is to plot the two fitted curves:

```
x_jitt <- runif(n, -0.2, 0.2)
plot((close + x_jitt)[seq(2,n,2)], n_agree[seq(2,n,2)]/J,
  xlim=c(1,5), ylim=c(0,1), xlab="Friendship measure", ylab="Proportion agreed")
curve(invlogit(coef(fit_agree)[1] + coef(fit_agree)[2]*x), add=TRUE, col="red")
plot(close + x_jitt, n_perceived/J, xlim=c(1, 5), ylim=c(0, 1),
  xlab="Friendship measure", ylab="Perceived agreement")
curve(invlogit(coef(fit_perceived)[1] + coef(fit_perceived)[2]*x),
  add=TRUE, col="red")
```

The data collected in class may not show clear patterns in either the graphs or the regressions. If so, that's fine, and it can be taken as an example of realistic ambiguity in results.

After looking at and analyzing the students' data, the instructor can discuss the research literature, which has found that friends are more similar than strangers, but by less than they expect.<sup>16</sup> There are various hypotheses for why people overestimate their friends' similarity to themselves. One possibility is that people see their own views as reasonable and so attribute them to others. Another possibility is that friendship colors people's judgment. Other research has found that "while

<sup>16</sup>See Miller McPherson, Lynn Smith-Lovin, and James Cook (2001), Birds of a feather: Homophily in social networks, *Annual Review of Sociology* 27, 415–444, and Sharad Goel, Winter Mason, and Duncan Watts (2010), Real and perceived attitude agreement in social networks, *Journal of Personality and Social Psychology* 99, 611–621.

individuals experience attitude homogeneity in their interpersonal networks, their networks are characterized by attitude heterogeneity,” and this is consistent with a model in which friends avoid areas of potential conflict in their conversations.<sup>17</sup> The result is that you can have an exaggerated belief in your friends’ agreement with you on contentious issues.

This activity relates to the week’s reading as an example of logistic regression with binomial data. It is relevant to the course as a whole by demonstrating data collection and analysis in the unusual setting of social networks.

## 2. Alternative models for discrete data

Here are some examples of count data: How many X’s do you know? How often have you been stopped by the police? How often have you gone to the hospital?

A class-participation activity can be built upon these by discussing alternative ways of modeling discrete data. Under the instructor’s guidance, the class can consider a particular area of interest and flesh out the details. For example, for the number of hospital visits in the past year, the class could consider regression models predicting this outcome given age, sex, and indicators for ethnicity and urban/rural/suburban.

With such a context, the class can discuss various alternative models for the outcome, including:

- Linear regression,
- Thresholding and binary logistic regression,
- Poisson regression,
- Overdispersed Poisson regression,
- Ordered logistic regression if there are not too many categories (for example, if the observed data counts range from 1 to 5) or binning larger categories (for example, coding counts of 0 as 0, counts of 1 as 1, counts of 2 through 4 as 2, counts of 5 through 9 as 3, and counts of 10 or higher as 4).

Depending on the problem being studied and the data, each of these alternatives has strengths and weaknesses, with issues including interpretability of the model, fit to data, and usefulness of predictions.

This activity relates to the week’s reading by pushing students to think about the applicability of different generalized linear models. It relates to the course as a whole in drawing connections between different families of models fit to a single (hypothetical) dataset.

## Computer demonstrations

### 1. Problem with fitting Poisson regression to simulated overdispersed data

We simulate data on the number of bacteria that can be found after a number of days in a hypothetical laboratory experiment. We first simulate data from a Poisson distribution and fit a Poisson model.<sup>18</sup> Then we simulate data from a negative binomial distribution with reciprocal overdispersion parameter  $\phi = 1$  and fit a Poisson as well as a negative binomial model. It may be worthwhile to experiment with different values of  $\phi$ .

```
library("MASS")
par(mfrow=c(1,3), mar=c(3,3,2,0), mgp=c(1.5,.5,0), tck=-.01)

# Simulate Poisson data, estimate Poisson regression, plot data, and fit
```

<sup>17</sup>Delia Baldassarri and Peter Bearman (2007), Dynamics of political polarization, *American Sociological Review* 72, 784–811.

<sup>18</sup>From Section 15.2 of *Regression and Other Stories*.

```
n <- 100
days <- runif(n, 0, 10)

bacteria <- rpois(n, 3.0*exp(0.2 * days))
fake <- data.frame(days, bacteria)
fit_pois <- stan_glm(bacteria ~ days, family=poisson(link="log"),
  data=fake, refresh=0)
print(fit_pois, digits=2)
plot(days, bacteria)
curve(exp(coef(fit_pois)[1] + coef(fit_pois)[2] * x), add=TRUE, col="red")
sims <- as.matrix(fit_pois)
n_sims <- nrow(sims)
for (i in sample(n_sims, 10))
  curve(exp(sims[i,1] + sims[i,2] * x), add=TRUE, col="red", lwd=.5)
newdata <- posterior_predict(fit_pois)
for (i in sample(n_sims, 2))
  plot(days, newdata[i,], bty="l", main="Simulated replication")

# Repeat with overdispersed data, fitting Poisson model
bacteria2 <- rnegbin(n, 3.0*exp(0.2 * days), 1)
fake2 <- data.frame(days, bacteria2)
fit2_pois <- stan_glm(bacteria2 ~ days, family=poisson(link="log"),
  data=fake2, refresh=0)
print(fit2_pois, digits=2)
plot(days, bacteria2)
curve(exp(coef(fit2_pois)[1] + coef(fit2_pois)[2] * x), add=TRUE, col="red")
sims <- as.matrix(fit2_pois)
n_sims <- nrow(sims)
for (i in sample(n_sims, 10))
  curve(exp(sims[i,1] + sims[i,2] * x), add=TRUE, col="red", lwd=.5)
newdata <- posterior_predict(fit2_pois)
for (i in sample(n_sims, 2))
  plot(days, newdata[i,], bty="l", main="Simulated replication")

# Repeat with overdispersed data, fitting negative binomial model
fit2_nb <- stan_glm(bacteria2 ~ days, family=neg_binomial_2(link="log"),
  data=fake2, refresh=0)
print(fit2_nb, digits=2)
plot(days, bacteria2)
curve(exp(coef(fit2_nb)[1] + coef(fit2_nb)[2] * x), add=TRUE, col="red")
sims <- as.matrix(fit2_nb)
n_sims <- nrow(sims)
for (i in sample(n_sims, 10))
  curve(exp(sims[i,1] + sims[i,2] * x), add=TRUE, col="red", lwd=.5)
newdata <- posterior_predict(fit2_nb)
for (i in sample(n_sims, 2))
  plot(days, newdata[i,], bty="l", main="Simulated replication")
```

## 2. Generalized linear model with offset

As explained in Section 15.2 of *Regression and Other Stories*, an *offset* in a Poisson or negative binomial model is the logarithm of a predictor that scales the expected outcome. We demonstrate here with a simple simulation.

```
library("MASS")
n <- 100
```

```
pop <- exp(runif(n, log(100), log(1e6)))
x <- runif(n, 0, 10)
a <- -7 # log(1000) = 6.9
b <- 0.1 # choose by comparing b*0 to b*10, the 2 extremes
deaths <- rnegbin(n, pop*exp(a + b*x), 2)
fake <- data.frame(x, deaths, pop)

fit_1 <- stan_glm(deaths ~ x, offset=log(pop),
  family=neg_binomial_2(link="log"), data=fake, refresh=0)
print(fit_1, digits=2)

plot(x, deaths/pop)
a_hat <- coef(fit_1)[1]
b_hat <- coef(fit_1)[2]
curve(exp(a_hat + b_hat*x), add=TRUE, col="red")
sims <- as.matrix(fit_1)
n_sims <- nrow(sims)
for (i in sample(n_sims, 10)) {
  curve(exp(sims[i,1] + sims[i,2]*x), add=TRUE, col="red", lwd=.5)
}

# Including log(pop) as a predictor
fit_2 <- stan_glm(deaths ~ x + log(pop),
  family=neg_binomial_2(link="log"), data=fake, refresh=0)
print(fit_2, digits=2)

# We can include log(pop) as a predictor and also as an offset
fit_3 <- stan_glm(deaths ~ x + log(pop), offset=log(pop),
  family=neg_binomial_2(link="log"), data=fake, refresh=0)
print(fit_3, digits=2)

# What about modeling death rate directly?
fake$death_rate <- fake$deaths/fake$pop
fit_4 <- stan_glm(log(death_rate) ~ x + log(pop), data=fake, refresh=0)
# This fails because you can't take log of 0.
```

## Drills

1. Interpret coefficients in a negative binomial or Poisson regression

For each of the following examples, interpret the coefficients and standard errors of the fitted Poisson regression.

Also, for each example it would be better to fit a negative binomial regression. If a negative binomial regression were fit instead of a Poisson regression, how would the estimated coefficients change and how would the standard errors change?

- (a) The number of awards earned by students at one high school, as predicted by the type of program in which the student was enrolled (vocational, general, or academic), and the score on their final exam in math:<sup>19</sup>

	Median	MAD_SD
(Intercept)	-5.3	0.6
progAcademic	1.1	0.4
progVocational	0.4	0.4
math	0.1	0.0

<sup>19</sup>See UCLA Statistical Consulting Group (2021), Poisson regression, <https://stats.idre.ucla.edu/r/dae/poisson-regression/>.

*Solution:*

- The intercept says that on average, under the model we would expect to see  $\exp(-5.3) = 0.005$  awards for a student in the general program (so that the indicators for academic and vocational programs are both 0) with a math score of 0. This is difficult to interpret because it might well be that no students get scores of 0 or close to that.
  - The coefficient for progAcademic says that, according to the model, on average comparing two students with identical math scores but where one is in the academic program and one is in the general program, the student in the academic program would be expected to receive  $\exp(1.1) = 3.0$  times as many awards. The standard error of 0.4 implies that we assign a roughly 68% chance to this multiplicative factor being in the range  $\exp(1.1 \pm 0.4) = (2.0, 4.5)$ .
  - The coefficient for progVocational has the corresponding interpretation, comparing two students with the same math scores but one is in the vocational program and one is in the general program. The student in the vocational program is predicted on average to receive  $\exp(0.4) = 1.5$  times as many awards, with a 68% uncertainty range of  $\exp(0.4 \pm 0.4) = (1.0, 2.3)$ .
  - The coefficient for math says that, according to the model, on average comparing two students in the same program but where one student scores 1 point higher on the final exam in math, the higher-scoring student would be expected to receive  $\exp(0.1) = 1.1$  times as many awards. This estimate is essentially impossible to interpret without having some sense of the range of math scores in the data or in the population of interest. What is relevant is  $\beta$  times a relevant difference in  $x$ , not  $\beta$  alone.
  - If a negative binomial regression were fit to these data, the estimated coefficients would be about the same. The standard errors would become bigger; there's no way of saying how much bigger they would be without working it out with the data.
- (b) The number of cyclists crossing the Brooklyn Bridge, as predicted by the high temperature (centered at the mean, in tens of degrees Fahrenheit) and the precipitation (centered at the mean, with units not specified in the data):<sup>20</sup>

	Median	MAD_SD
(Intercept)	7.85	0.00
c_HIGH_T_tens	0.12	0.00
c_PRECIP	-0.84	0.01

## 2. Interpret the parameters in ordered logistic regression

For each of the following examples, interpret the estimated coefficients and cutpoints of the fitted ordered logistic regression.

- (a) Predicting party identification (on a 3-point scale, 1 = Democrat, 2 = Independent, 3 = Republican), given age10 (age in years divided by 10) and sex:

```
stan_polr
family:      ordered [logistic]
formula:     factor(pid3) ~ age10 + male
observations: 7891
-----
          Median MAD_SD
age10    0.06   0.01
male     0.38   0.04
```

<sup>20</sup>Sachin Date (2021), The Poisson regression model, <https://timeseriesreasoning.com/contents/poisson-regression-model/>.

Cutpoints:

	Median	MAD_SD
1 2	-0.12	0.07
2 3	1.31	0.07

*Solution:*

- Cutpoints: For a person with zero values of predictors (that is, a woman with age 0), the predicted probabilities of identifying as Democrat, Independent, or Republican are  $1 - \text{logit}^{-1}(0.12) = 0.47$ ,  $\text{logit}^{-1}(0.12) - \text{logit}^{-1}(-1.31) = 0.32$ , and  $\text{logit}^{-1}(-1.31) = 0.21$ , respectively.
  - Coefficient for *age10*: Comparing two respondents who are the same sex and 10 years apart in age, the model predicts that the older person is more likely to be Republican. For example, compare a 40-year-old woman to a 50-year-old woman. For the 40-year-old, the predicted probabilities of identifying as Democrat, Independent, or Republican are  $1 - \text{logit}^{-1}(0.06 * 4 + 0.12) = 0.41$ ,  $\text{logit}^{-1}(0.06 * 4 + 0.12) - \text{logit}^{-1}(0.06 * 4 - 1.31) = 0.33$ , and  $\text{logit}^{-1}(0.06 * 4 - 1.31) = 0.26$ , respectively. For the 50-year-old, the predicted probabilities of identifying as Democrat, Independent, or Republican are  $1 - \text{logit}^{-1}(0.06 * 5 + 0.12) = 0.40$ ,  $\text{logit}^{-1}(0.06 * 5 + 0.12) - \text{logit}^{-1}(0.06 * 5 - 1.31) = 0.33$ , and  $\text{logit}^{-1}(0.06 * 5 - 1.31) = 0.27$ , respectively.
  - Coefficient for *male*: Comparing a man and a woman who are the same age, the model predicts that the man is more likely to be Republican. For example, compare a 50-year-old woman to a 50-year-old man. For the woman, the predicted probabilities of identifying as Democrat, Independent, or Republican are  $1 - \text{logit}^{-1}(0.06 * 5 + 0.12) = 0.40$ ,  $\text{logit}^{-1}(0.06 * 5 + 0.12) - \text{logit}^{-1}(0.06 * 5 - 1.31) = 0.33$ , and  $\text{logit}^{-1}(0.06 * 5 - 1.31) = 0.27$ , respectively. For the man, the predicted probabilities of identifying as Democrat, Independent, or Republican are  $1 - \text{logit}^{-1}(0.06 * 5 + 0.38 + 0.12) = 0.31$ ,  $\text{logit}^{-1}(0.06 * 5 + 0.38 + 0.12) - \text{logit}^{-1}(0.06 * 5 + 0.38 - 1.31) = 0.34$ , and  $\text{logit}^{-1}(0.06 * 5 - 1.31) = 0.35$ , respectively.
- (b) Predicting party identification (on a 3-point scale), given *age10* (age in years divided by 10), sex, and education (on a 5-point scale from 1 = no high school degree to 5 = postgraduate degree):

```
stan_polar
  family:      ordered [logistic]
  formula:     factor(pid3) ~ age10 + male + education
  observations: 7408
```

	Median	MAD_SD
age10	0.07	0.01
male	0.38	0.04
education	-0.11	0.02

Cutpoints:

	Median	MAD_SD
1 2	-0.43	0.08
2 3	1.01	0.08

<sup>21</sup>We use question VCF0838, "By law, when should abortion be allowed," from the 2020 American National Election Study, and here is the exact wording: "There has been some discussion about abortion during recent years. Which one of the opinions on this page best agrees with your view? You can just tell me the number of the opinion you choose. 1. By law, abortion should never be permitted. 2. The law should permit abortion only in case of rape, incest, or when the woman's life is in danger. 3. The law should permit abortion for reasons other than rape, incest, or danger to the woman's life, but only after the need for the abortion has been clearly established. 4. By law, a woman should always be able to obtain an abortion as a matter of personal choice." We removed nonresponses from our analysis.

- (c) Predicting attitude on abortion (1 = should never be legal, 2 = should be illegal in most circumstances, 3 = should be legal in most circumstances, 4 = should always be legal), given age10 (age in years divided by 10), sex, education (on a 5-point scale), and party identification (on a 3-point scale):<sup>21</sup>

```
stan_polr
family:      ordered [logistic]
formula:     factor(abortion) ~ age10 + male + education + pid3
observations: 7098
-----
          Median MAD_SD
age10      -0.09   0.01
male        0.02   0.05
education   0.17   0.02
pid3       -1.07   0.03

Cutpoints:
          Median MAD_SD
1|2 -4.51   0.12
2|3 -2.82   0.11
3|4 -2.11   0.11
```

## Discussion problems

### 1. Identification in linear models

Consider a regression model with two identical predictors, for example  $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \text{error}_i$ , for  $i = 1, \dots, n$  where  $x_1 \equiv x_2$ . In this case, the coefficients  $b_1$  and  $b_2$  are said to be *not identified* from the data. It is possible to fit such models using `stan_glm` but the identification there is coming from the prior. A similar problem arises if one of the predictors is a constant, in which case the data do not separately identify the intercept and the coefficient for the constant predictor.

More complicated examples of lack of identification can arise when multiple predictors are linearly dependent. This can arise, for example, if a dataset includes people from four different occupational categories, and indicators for all four categories are included as regression predictors. In this case, the standard approach is to include indicators for three of the categories and consider the other category as default; see page 138 in Section 10.4 of *Regression and Other Stories*.

A good topic for discussion is to consider various ways that nonidentification can arise in applications. In some cases, nonidentification can be *structural*, for example if a model includes three predictors, number of boys, number of girls, and number of children, then  $x_1 + x_2 - x_3 = 0$  by construction. In other settings, nonidentification can arise with the data at hand, for example if certain categories that are theoretically possible happen to be empty. This is related to the problem of complete separation in logistic regression, as discussed in Section 14.6 of *Regression and Other Stories*.

### 2. Functional forms for nonlinear models

Tricky issues of identification can arise with nonlinear models. For example, what is the problem with predicting  $y$  from  $x$  using the model,  $y = a + b * \exp(-cx + d) + \text{error}$ ? This is an entry point to a discussion of different possible nonlinear functional forms. Students will have learned about quadratic or polynomial models such as  $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \text{error}$ —again, we are considering  $x$  as the predictor,  $y$  as the outcome, and the parameters  $b_0, b_1, b_2, b_3$  as constants to be estimated from data—but polynomials are typically not the most effective ways of modeling real-world relationships. The quadratic model has the problem that its rise and fall are symmetric; also it blows up at the extremes. Declining exponential models can be useful because they flatten out in the limit:  $y = b * \exp(-cx)$  for  $x_i \geq 0$  starts at  $b$  and decreases to 0 as  $x$  increases, with a

scale of  $1/c$ . Students can learn by drawing these curves for different values of  $b$  and  $c$ . The next question is how to write the equation of a curve that starts at zero and increases with an asymptote. Here is an answer:  $y = b * (1 - \exp(-cx))$ . This can be made non-monotonic by adding a linear trend:  $y = a + b * (1 - \exp(-cx)) + dx$ . Again, graphing this for different choices of parameters will provide insight.

## 4.18 Design and sample size decisions

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
The multiverse and the feedback loop	Design an experiment from scratch	Design analysis by simulation	Sample size calculation for proportions	Designing a survey
Lucky golf balls and implausible effect sizes	Hypothetical study of left-handedness	Design for estimating interactions	Sample size calculation for averages	Designing future studies

### Reading

Chapter 16 of *Regression and Other Stories*: Design and sample size decisions

### Pre-class warmup assignments

#### 1. Sample size and statistical power: comparison of proportions

Suppose you are comparing two manufacturing processes, with the goal of reducing the rate of defects in a product from 1% under process A to 0.8% under process B. For each design below, how large must  $n$  be for your experiment to have 80% power of finding a difference that is “statistically significant” at the conventional 95% level, under this assumed treatment effect.

- (a) Produce  $n/2$  products under each process and compare the defect rates.
- (b) Produce  $2n/3$  products under process A and  $n/3$  products under process B.
- (c) Produce 20 000 products under process A and  $n$  products under process B.
- (d) Produce 2000 products under process A and  $n$  products under process B.

#### 2. Sample size and statistical power: regression

A linear regression is performed to estimate the effect of a treatment, yielding the following result:

```
formula:      post_test ~ z + pre_test
observations: 100
predictors:   3
-----
           Median MAD_SD
(Intercept) 23.6   10.9
z            1.4    4.0
pre_test     0.7    0.2

Auxiliary parameter(s):
  Median MAD_SD
sigma 20.1    1.4
```

Here,  $z$  is the treatment indicator, and the coefficient of  $z$  is the treatment effect. Suppose are designing a new study on a similar population and you think the true treatment effect will be 5.0.

- (a) How large a sample size would you need to have approximately 80% power of finding a difference that is “statistically significant” at the conventional 95% level?
- (b) How large a sample size would you need to have approximately 80% power of finding a difference that is “statistically significant” at the 90% level?

### Homework assignments

1. (a) Logistic regression and choice models (Exercise 15.10 of *Regression and Other Stories*)  
Using the individual-level survey data from the election example described in Section 10.9 of *Regression and Other Stories* (data in the folder NES), fit a logistic regression model for the choice of supporting Democrats or Republicans. Then interpret the output from this regression in terms of a utility/choice model.
2. (a) Sample size calculations for estimating a proportion (Exercise 16.1 of *Regression and Other Stories*)
  - i. How large a sample survey would be required to estimate, to within a standard error of  $\pm 3\%$ , the proportion of the U.S. population who support the death penalty?
  - ii. About 14% of the U.S. population is Latino. How large would a national sample of Americans have to be in order to estimate, to within a standard error of  $\pm 3\%$ , the proportion of Latinos in the United States who support the death penalty?
  - iii. How large would a national sample of Americans have to be in order to estimate, to within a standard error of  $\pm 1\%$ , the proportion who are Latino?
- (b) Sample size calculations for estimating a difference (Exercise 16.2 of *Regression and Other Stories*)  
Consider an election with two major candidates, A and B, and a minor candidate, C, who are believed to have support of approximately 45%, 35%, and 20% in the population. A poll is to be conducted with the goal of estimating the difference in support between candidates A and B. How large a sample would you estimate is needed to estimate this difference to within a standard error of 5 percentage points? (Hint: consider an outcome variable that is coded as +1, -1, and 0 for supporters of A, B, and C, respectively.)
- (c) *In pairs:* Working through your own example (Exercise 15.18 of *Regression and Other Stories*)  
Continuing the example from the final exercises of the earlier chapters, fit a generalized linear model that is not a logistic regression, graph the data and fitted model, and interpret the estimated parameters and their uncertainties.

### Stories

1. The multiverse, the statistical significance filter, and the feedback loop

Section 4.5 of *Regression and Other Stories* discusses a study conducted in 2012 that claimed that single women during certain times of the month were 20 percentage points more likely to support Barack Obama for president. As we wrote in the book,

“There are so many different things that could be compared, but all we see is some subset of the comparisons. Some of the choices available in this analysis include the days of the month characterized as peak fertility, the dividing line between single and married (in this particular study, unmarried but partnered women were counted as married), data exclusion rules based on reports of menstrual cycle length and timing, and the decision of which interactions to study. Given all these possibilities, it is no surprise at all that statistically significant comparisons turned up; this would be expected even were the data generated purely by noise.”

We looked at this question more carefully by going back to the data in this study, considering different options in the data coding and analysis, and performing a “multiverse analysis” by looking at the possible results.<sup>22</sup>

<sup>22</sup>Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel (2016), Increasing transparency through a multiverse analysis, *Perspectives on Psychological Science* 11, 702–712.

1. Exclusion criteria based on cycle length (3 options)
  2. Exclusion criteria based on “How sure are you?” response (2 options)
  3. Cycle day assessment (3 options)
  4. Fertility assessment (4 options)
  5. Relationship status assessment (3 options)
- 168 possibilities (after excluding some contradictory combinations)

**Figure 88** Options in the analysis of the ovulation-and-voting study. We reanalyzed the data in each of these 168 ways to see how sensitive the inference was to these choices of data coding.

Figure 88 shows the options we considered in our reanalyses. These are really only a small subset of the coding and modeling choices, but they still give a sense of some of the forking paths. For each of these choices, we chose the options based on this and other papers in the literature. For example, one of the papers by these authors assigned days 7–14 to the high-fertility group and days 17–25 to the low-fertility group, another used days 9–17 and 18–25, another used 8–14 for high fertility and 1–7 and 15–28 for low fertility, and another used 9–17 and 1–8, 18–28. Similarly, there are different possible choices in excluding women with irregular cycles, categorizing relationship status, and so on. We ended up with 168 options, and for each we performed the analysis performed in the original article, which was essentially a linear regression on the binary outcome (prefer Barack Obama or Mitt Romney for president) on ovulation status, relationship status, and their interaction. The key parameter here is the coefficient on the interaction, and for each of the 168 analyses we computed its estimate and standard error.

Of all these analyses, only three were “statistically significant” in the sense of having a 95% interval excluding zero, and the analysis that appeared in the published paper is one of these three. This is not to say that the authors performed all 168 analyses and chose the one whose estimate was the most standard errors away from zero; rather, they made various coding and analysis decisions with the data in front of them, and it makes sense that they would make choices that yielded the sharpest results. But such an analysis procedure can easily produce apparently statistically significant results from pure noise.

This sort of thing can result in a feedback loop by which effect sizes are overestimated, leading to overly optimistic designs for new studies, leading to noisy data followed by a search for statistically significant results, which will then be extreme overestimates of effect sizes, leading to overconfidence in future research, and so forth.

This story is relevant to the week’s reading by connecting choices in design, data processing, and analysis to the way that a statistical study is interpreted. It relates to the course as a whole as an example of the feedback loop by which quantitative misconceptions (in this case, of large effects of politically irrelevant stimuli) can be propagated through flawed quantitative analyses.

## 2. Lucky golf balls and implausible effect sizes

In June 2021, the site Golf.com ran an article, “‘Lucky’ golf items might actually work,” pointing to a research paper published in 2010 in which a team of psychologists claimed that “superstition improves performance.”<sup>23</sup> Empirical support for this statement came from an experiment on 28 university students who were given the task of attempting 10 short golf putts. The students were randomly assigned to the treatment (told they’d been given a “lucky ball”) or the control condition (told their ball was ordinary). The average number of shots made was 6.42 out of 10 in

<sup>23</sup>Luke Kerr-Dineen (2021), “Lucky” golf items might actually work, according to study, <https://golf.com/instruction/lucky-golf-balls-study-play-smart/>. The discussion here is taken from Andrew Gelman (2021), The so-called “lucky golf ball”: The Association for Psychological Science promotes junk science while ignoring the careful, serious work of replication, <https://statmodeling.stat.columbia.edu/2021/12/20/not-replicable-but-citable/>.

the treatment group and 4.75 out of 10 in the control group. The standard deviation of number of shots made was 1.88 in the treatment group and 2.15 in the control group, hence the observed difference of  $6.42 - 4.75 = 1.67$  with a standard error of  $\sqrt{1.88^2/14 + 2.15^2/14} = 0.76$ , and a 95% confidence interval of  $[1.67 \pm 2 * 0.76] = [0.91, 2.43]$ . The interval excludes zero, so, following conventional practice, this was considered as strong evidence in favor of the claim of improved performance.

But there's a problem here. An increase of 1.67 successes—that's a  $1.67/4.75 = 35\%$  improvement—all from using a ball that you're told is “lucky”—that's an implausibly large effect. Any average effect of such a belief would realistically be much lower. Suppose, for example, that a belief in the lucky ball increased the probability of success by 5%, which would correspond to  $0.05 * 4.75 = 0.24$  shots out of 10. Then the existing experiment, which yields a standard error of 0.76, is way too noisy to detect such an effect. As discussed in Chapter 16 of *Regression and Other Stories*, to get 80% power we need an effect size of 2.8 standard errors away from zero, hence a standard error of  $0.24/2.8 = 0.09$ . The existing experiment has a standard error that is  $0.76/0.09 = 8$  times too high, which implies that if the design of the study were kept as is, we'd want at least 64 times the sample size, thus 1800 students in the study rather than 28.

To go back to the study that actually happened, the estimate is just too noisy to be useful.

There's only one missing piece in the story. If, as we claim above, the study was underpowered, far too noisy to learn anything useful, then how did the authors manage to find a “statistically significant” result? Sure, you can get a 95% interval to exclude zero just by luck—even if there is no effect at all, it will happen 5% of the time from chance alone—but the published article had four experiments, each with its own statistically significant finding. It seems unlikely that all four of these comparisons could have attained statistical significance by chance alone.

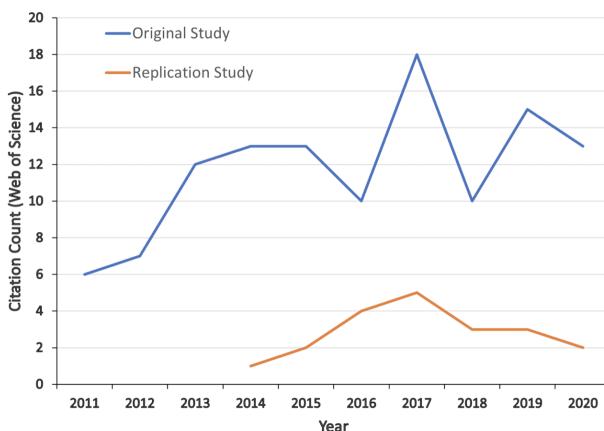
At this point we usually point to forking paths—alternative analyses that could give different results. But this study seems so simple. Here is the description from the published article:<sup>24</sup>

“We recruited 28 university students (12 males, 16 females) as participants and randomly assigned them to a superstition-activated or a control condition. Participants were asked to engage in a 10-trial putting task. A pretest revealed that more than 80% of our participant population believed in good luck, so to activate the superstition, we linked the concept of good luck to the ball participants used during the task. Specifically, while handing the ball over to the participants, the experimenter said, ‘Here is your ball. So far it has turned out to be a lucky ball’ (superstition-activated condition) or ‘This is the ball everyone has used so far’ (control condition). Finally, participants performed the required 10 putts from a distance of 100 cm. We used the number of hits as our central dependent measure, with ‘hits’ defined as successful putts (when the ball actually ended up where it was supposed to be). As predicted, participants performed better when playing with an ostensibly lucky ball ( $M = 6.42$ ,  $SD = 1.88$ ) rather than a neutral ball ( $M = 4.75$ ,  $SD = 2.15$ ).”

With such a simple design and analysis, where could there possibly be forking paths? We project the above description onto the screen and ask students to consider the question in pairs.

What is the answer? It's hard to know, but here are a few possibilities. First, we are told of a pre-test where the participants were asked if they believed in good luck. There may have been other questions on the pre-test too. This provides lots of potential forking paths. For example, the analyses could have been restricted just to the people who reported believing in good luck, or just to men, or just to women, or the analysis could have looked at the interaction of sex with the treatment. We are also told that the number of hits was the “central dependent measure,” which suggests that other outcomes could have been used, for example the average distance of putt from

<sup>24</sup>Lysann Damisch, Barbara Stoberock, and Thomas Mussweiler (2010), Keep your fingers crossed!: How superstition improves performance, *Psychological Science* 21, 1014–1020.



**Figure 89** Number of citations of the “lucky golf ball” paper and the later replication study that revealed no effect. Sadly, the erroneous paper continues to be cited in the academic literature and promoted in the news media, while the failed replication is barely noticed.

the hole, or maybe something simpler such as just counting the number of students in each group who had at least a 50% success rate. Another possibility is forking paths in data inclusion. We are told of these 28 students, but perhaps this was just one of several groups that were analyzed. Another potential forking path is the task itself. Perhaps an earlier study was done with puts at a distance of 200 cm, the results were not statistically significant, and researchers decided that the task was too difficult, so then they tried a shorter distance which then worked as planned. Similarly, had the results not shown up in 10 puts, they could have extended the data collection.

This all may sound like fanciful speculation—but really we have no idea what was going on. Consider, for example, the following alternative description that we made up:

“We recruited 60 university students (28 males, 32 females) as participants in two waves of a study and randomly assigned them to a superstition-activated or a control condition. In a pre-test, 47 of these students stated that they believed in luck. All participants were asked to engage in a 15-trial putting task from a distance of 150 cm. As our central dependent measure, we used total success, defined as more than half the attempted puts ending up where they were supposed to be. Consistent with our hypothesis, participants who believed in luck performed better when playing with an ostensibly lucky ball, while there was no difference among those expressed no belief in luck.”

That sounds just as reasonable as the original, right? There are many researcher degrees of freedom, and it could have seemed reasonable to the researchers at the time to navigate through the paths to find a statistically significant result.

One way to get a handle on this problem is to perform a replication study, and this was done a few years later: a team of researchers performed a replication on 124 students and another replication with a slightly stronger intervention on 111 students. The results were consistent with chance: in the first experiment, the difference was 0.11 successes (out of 10 tries), with a standard error of 0.37; the second experiment found an average difference of 0.10 with a standard error of 0.40.

Given our earlier design analysis, it is unsurprising that a sample of 124 or 111 is not nearly large enough to reliably detect a realistic effect size here.

The design analysis, followed by the failed replication attempts, makes us doubt there is anything beyond noise in the original finding. In that first article, the authors wrote, “This study is in line with prior research examining the influence of seemingly irrelevant factors, such as one’s arm position, specific colors, or preperformance routines, on intellectual and physical performance.” Unfortunately, these earlier studies may well have been flawed in the same way.

Does the problem of forking paths imply that nothing can be trusted? No. One way we can get around the problem of forking paths is to decide the baseline analysis ahead of time. This strategy can work well in replication—that’s what was done in the replication studies just mentioned—but is not so practical for a new project where, realistically, some exploration is necessary to understand the data. In settings where we cannot preregister the analysis, we recommend performing multiple analyses and not laying all your bets down on whether a particular comparison or coefficient is “statistically significant.”

This story is relevant to the week’s reading in demonstrating the value of design analysis for understanding what can be learned from a study, in particular the challenge of interpreting results when the underlying effect size is likely to be small. It relates to the course as a whole as an example of the connections between theoretical models, data collection and analysis, and scientific conclusions.

There is a sad postscript to this story. Someone tracked the citations to the “lucky golf ball” article and the followup replication study. Unfortunately, the original study continues to be cited; see Figure 89. This is a case where statistics did (a small amount of) scientific damage, or at least wasted some people’s time. Without the infrastructure of randomized experimentation, statistical modeling, and hypothesis testing, we doubt that a claimed effect of lucky golf balls would have been taken seriously.

### Class-participation activities

#### 1. Design analysis for an experiment

The class can work together to design an experiment. The instructor should start by projecting Figure 90 onto the screen to show the steps of a simulation-based design analysis. The starting point is to consider an area of application with some treatment of interest and then a pre-test and post-test measurement. For example, this could be a study of a hypothetical advertising program designed to increase vaccination rates, so that the individual data points in the study are cities, the pre-treatment variable  $x$  is the vaccination rate in the city last year, each city receives the advertising program ( $z = 1$ ) or not ( $z = 0$ ), and the outcome  $y$  is the vaccination rate a year later. Or it could be a study on individuals, for example a physical therapy program to increase the range of motion of people who are recovering from wrist injuries.

1. Consider a pre-test measurement  $x$ , a treatment  $z$ , and a post-test outcome  $y$
2. Set a distribution for the pre-test measurement  $x$
3. Assume 50/50 randomized assignment for the treatment,  $z$
4. Specify values for the parameters in the model  $\log y = \beta_0 + \beta_1 \log x + \beta_2 z + \text{error}$
5. Choose a provisional sample size  $n$
6. Simulate  $x$ ,  $z$ , and  $y$  (in that order)
7. Plot the data and see if they make sense
8. Fit the linear regression of  $\log y$  on  $\log x$  and  $z$  and look at the parameter estimates and uncertainties
9. Look at the standard error for the treatment effect and alter the sample size accordingly to aim for a standard error that is equal to the hypothesized effect size divided by 2.8
10. Loop it: perform 100 simulations, save the results, and look at the distribution of estimates and uncertainties

Figure 90 *Steps for choosing the sample size for a hypothetical experiment. These should be projected onto onto the screen, and then students can work in pairs to go through these steps.*

In any case, the pre-test and post-test measurements should be positive continuous numbers (for example, rates, costs, spending, or physical measurements) so that we can fit a multiplicative model, that is, a linear model on the log scale. The class can work through all the steps in Figure 90, with the simulation in the final step confirming that the chosen sample size is sufficient.

This activity relates to the week's reading as an example of designing a study using simulation. It relates to the course as a whole in requiring numbers to be assigned to the model based on assumptions regarding a realistic problem.

## 2. Sample size calculation for a hypothetical study of left-handedness

Suppose we want to design a study comparing the handedness of artists to the general population, and to do so we conduct a study with  $n$  artists and  $n$  others. How large a study would we need?

We should be able to figure this out from first principles and common knowledge. Start by considering the measurement. We'll assume that each person in the study is given a handedness questionnaire as in Figure 20, giving them a score ranging from  $-1$  for pure left-handers to  $+1$  for pure righties. Call this measurement  $y$ . We would then compute the difference  $\bar{y}_1 - \bar{y}_2$ , which has standard error  $\sqrt{s_1^2/n + s_2^2/n}$ , where  $s_1$  and  $s_2$  are the standard deviations of the handedness scores in the two groups.

So to figure this out we just need an assumption about the mean and standard deviations of the handedness scores of artists and others. How to get that?

To start with, we know that approximately 90% of people are right-handed, and that non-right-handers are roughly uniformly distributed across the handedness spectrum. For simplicity we shall construct a discrete distribution of handedness, with possible values  $-1.0, -0.9, -0.8, \dots, 0.9, 1.0$ . We'll assume that the 90% of people who are right-handers are divided evenly among scores 0.8, 0.9, and 1.0, and that the remaining 10% of people have scores uniformly distributed between  $-1.0$  and  $0.7$ . This distribution is shown in Figure 91a.

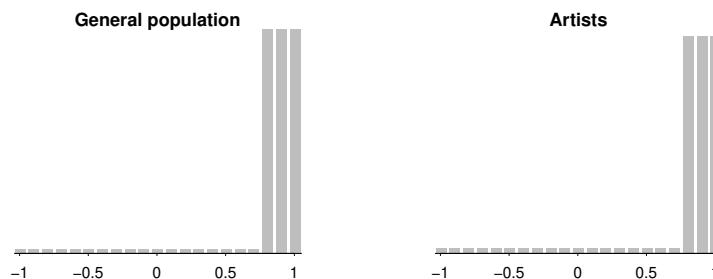
What about artists? Suppose they are three percentage points more likely to be lefties. We shall assume that this is attained by having 87% being divided evenly among scores 0.8, 0.9, and 1.0, with the remaining 13% uniformly distributed between  $-1.0$  and  $0.7$ , as shown in Figure 91b.

We can compute the mean and standard deviation of these distributions:

```

h <- seq(-1, 1, 0.1)
p <- rep(c(0.1/18, 0.9/3), c(18, 3))
mean_h <- sum(h*p)
sd_h <- sqrt(sum((h - mean_h)^2*p))
p_art <- rep(c(0.13/18, 0.87/3), c(18, 3))
mean_h_art <- sum(h*p_art)
sd_h_art <- sqrt(sum((h - mean_h_art)^2*p_art))

```



**Figure 91** Hypothesized distribution of handedness scores (a) for the general population and (b) for artists, under the assumption that artists are 3 percentage points more likely to be left-handers. This does not look like much on the graph, but it corresponds to a big difference (from 10% mixed- or left-handers to 13%).

Course of study	Sample size	Percentage left-handed
Behavioral sciences	90	8.89%
Humanities	51	9.80%
Sciences	92	4.35%
Other arts and sciences	156	7.05%
Business	241	9.54%
Music	47	14.89%
Design and art	147	12.24%
Engineering	75	10.67%
Nursing	71	4.23%
Other	75	9.38%

**Figure 92** Data from a survey of 1045 students in an introductory psychology course, used to compare the rate of left-handedness (“determined by the self-reported hand used to perform a simple drawing task”) among different majors. The comparisons look interesting at first, but there is too much uncertainty here to draw any firm conclusions, as can be seen from a careful look at uncertainties; see Figure 93.

The resulting values are  $\text{mean\_h} = 0.795$  and  $\text{sd\_h} = 0.36$  for the general population and  $\text{mean\_h\_art} = 0.764$  and  $\text{sd\_h\_art} = 0.41$  for the artists.

Now we can perform a design analysis. The difference between the average handedness score of  $n$  artists and the average handedness of  $n$  others would have an approximately normal distribution with mean  $0.795 - 0.764 = 0.031$  and standard deviation  $\sqrt{0.41^2/n + 0.36^2/n} = 0.55/\sqrt{n}$ . To have 80% power to detect this difference with a 95% confidence interval, it is necessary for the effect size to be more than 2.8 times the standard deviation of the estimate (see Section 16.2 of *Regression and Other Stories*), thus  $0.031 > 2.8 * 0.55/\sqrt{n}$ ; that is,  $n > (2.8 * 0.55/0.031)^2 = 2467$ . So we would need a sample size of at least 2500 to be able to reliably see this difference.

We can also understand this by simulating data from our hypothesized distributions of handedness. We’ll first simulate a study with  $n = 100$  in each group:

```
n <- 100
y_controls <- sample(h, n, p, replace=TRUE)
y_arts <- sample(h, n, p_art, replace=TRUE)
diff <- mean(y_controls) - mean(y_arts)
se_diff <- sqrt(sd(y_controls)^2/n + sd(y_arts)^2/n)
print(round(c(diff, se_diff), 3))
```

This should be too small, and indeed it is. We can further check by looping this 50 times:

```
for (loop in 1:50){
  n <- 100
  y_controls <- sample(h, n, p, replace=TRUE)
  y_arts <- sample(h, n, p_art, replace=TRUE)
  diff <- mean(y_controls) - mean(y_arts)
  se_diff <- sqrt(sd(y_controls)^2/n + sd(y_arts)^2/n)
  print(round(c(diff, se_diff), 3))
}
```

The estimates are hopelessly noisy and almost never more than two standard errors from zero. Redoing with  $n = 2500$  gives more reasonable results.

<sup>25</sup>John Peterson (1979), Left-handedness: Differences between student artists and scientists, *Perceptual and Motor Skills* 48, 961–962.

Course of study	Sample size	Percentage left-handed	± standard error
Behavioral sciences	90	9%	±3%
Humanities	51	10%	±4%
Sciences	92	4%	±2%
Other arts and sciences	156	7%	±2%
Business	241	10%	±2%
Music	47	15%	±5%
Design and art	147	12%	±3%
Engineering	75	11%	±2%
Nursing	71	4%	±4%
Other	75	9%	±3%

**Figure 93** Repeat of Figure 92 including standard errors. The differences between the percentages in the table are not large once we consider the uncertainties; a sample size of 1045 is too small to identify differences in proportion left-handed among these groups. Also, recognizing the large uncertainties, we have rounded each number to the nearest percentage point, as it would be silly, for example, to report a percentage as 8.89% when the uncertainty is 3 percentage points.

Googling revealed a study with some data; see Figure 92.<sup>25</sup> At first the results in this table may look interesting—compare, for example, the music and arts majors with over 12% left-handers to the sciences with less than 5%. Calculation of standard errors, however, reveals these differences to be consistent with noise. Figure 93 demonstrates by repeating the table including standard errors using the formula  $\sqrt{p(1-p)/n}$  for each group. For example, the comparison between the science majors and the music majors is 10 percentage points but with a standard error of  $\sqrt{2^2 + 5^2} = 5$  percentage points, so even this difference is only 2 standard errors away from zero.

At first this might be considered a respectable finding, indeed attaining the conventional level of “statistical significance,” but consider all the other comparisons that could have been made. Before the data had been collected, it could seem natural to include engineering with the sciences and include the arts with music. The resulting comparison is then  $(92 * 4.35\% + 75 * 10.67\%)/(92 + 75) = 7.2\%$  for science/engineering and  $(47 * 14.89\% + 147 * 12.24\%)/(47 + 147) = 12.9\%$  for music/arts; the difference is 5.7% with standard error  $\sqrt{0.072(1 - 0.072)/(92 + 75) + 0.129(1 - 0.129)/(47 + 147)} = 3.2\%$ , which could easily be explained by chance.

In short, the noise in the estimation of the proportions in Figure 93 overwhelms any possible signal, which is no surprise given our design calculations earlier. The simulation revealed that sample sizes of 100 per group are not nearly enough to detect realistic differences in handedness.

This activity is relevant to the week’s reading in that students are performing a design analysis from first principles, using the conventional criterion of statistical power. It relates to the course as a whole in connecting a claim of statistical significance from the scientific literature to statistical theory explaining why such a claim is overstated.

### Computer demonstrations

1. Design analysis for a simple experiment, including with a pre-treatment predictor

Consider the case of an experimental sports regimen that is expected to decrease heart rate by 3 beats per minute. Assume that heart rates are normally distributed with mean 80 and standard deviation 10. Also assume that  $\sigma_1 = \sigma_2 = \sigma$  and  $n_1 = n_2 = n/2$ , for control and treatment groups, respectively.

First calculate the required  $n$  for a simple experiment to achieve 80% power. Then show the effect

of including a pre-treatment predictor for physical activity. As in Section 16.6 of *Regression and Other Stories*, the variation in  $y$  is constructed to be the same across both setups.

```
# Calculate required n
theta <- -3 # treatment effect: difference in heart rates
sigma <- 10 # variation in treatment and control group
required_n <- ceiling((5.6 * sigma / theta) ^ 2) # rounding up

# Validate calculation
loop <- 1000
ledger <- array(NA, dim=c(loop, 2)) # store estimate and se of coef
for (i in 1:loop) {
  y_if_control <- rnorm(required_n, 80, sigma)
  y_if_treatment <- y_if_control + theta
  z <- sample(c(0,1), required_n, replace=TRUE)
  y <- ifelse(z == 1, y_if_treatment, y_if_control)
  ledger[i,1] <- mean(y[z==1]) - mean(y[z==0]) # coef
  ledger[i,2] <- sqrt((sd(y[z==1])^2/length(y[z==1])) +
    (sd(y[z==0])^2/length(y[z==0]))) # se
}
mean(abs(ledger[,1]/ledger[,2]) >= 1.96) # power (note "abs()")

# Include pre-test, using latent variable representing physical activity
library("rstanarm")
n <- required_n
ledger <- array(NA, dim=c(loop, 2)) # store estimate and se of coef
for (i in 1:loop) {
  x <- rnorm(n, 20, 6) # a physical activity scale
  y_if_control <- (100 - x) + rnorm(n, 0, 8) # heart rate
  y_if_treatment <- y_if_control + theta
  z <- sample(c(0,1), n, replace=TRUE)
  y <- ifelse(z == 1, y_if_treatment, y_if_control)
  fake <- data.frame(x, y, z)
  fit <- stan_glm(y ~ z + x, data=fake, refresh=0)
  ledger[i,] <- summary(fit)[["z",c("mean", "sd")]]
  if(i %% 5 == 0) cat("Rep:", i, "\n")
}
mean(abs(ledger[,1]/ledger[,2]) >= 1.96) # power
```

2. You need 16 times the sample size to estimate an interaction that is half the size as the main effect

You can see the challenge of estimating interactions, as discussed in Section 16.4 of *Regression and Other Stories*, by simulating a hypothetical experiment on the effect of canvassing on feeling thermometer scores toward Joe Biden.<sup>26</sup> First, consider an intervention that improves feelings by 10 points for everybody. You can calculate the standard error of the estimate of the main effect and see what sample size would be required to achieve a power of 80%.

Then introduce an interaction that is half the size of the main effect: women's feelings increase by 5 points more than men's. Using the same sample size as before, the interaction term that is twice as large as the standard error of the main effect.

As discussed on page 302 in Section 16.4 of *Regression and Other Stories*, to be able to detect the interaction effect, the experiment needs to have a sample size 16 times what is required to detect the main effect. We demonstrate with this code:

<sup>26</sup>Pew Research Center (2020), Perceptions of Trump and Biden, <https://www.pewresearch.org/politics/2020/10/09/perceptions-of-donald-trump-and-joe-biden/>.

```
# Simple case: main effect only (10 points)
n <- 100
theta <- 10
sigma <- 20
y_control <- rnorm(n, 45, sigma) # Pew, fall 2020
y_treatment <- y_control + theta
z <- sample(c(0,1), n, replace=TRUE)
y <- ifelse(z==1, y_treatment, y_control)
fake <- data.frame(y, z)
fit <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit, digits=2)

# Calculate the standard error analytically
se <- sqrt((sigma^2/(n/2)) + (sigma^2/(n/2)))
print(se)

# Calculate the sample size needed for 80% power
n_required <- (5.6 * sigma / theta)^2
print(n_required)

# Repeat with interaction half the size of main effect (5 points)
# ... "Female" is set at (-0.5, 0.5) instead of (0, 1)
n <- 100
female <- sample(c(-0.5, 0.5), n, replace=TRUE)
y_control <- rnorm(n, 45, 20)
y_treatment <- y_control +
  ifelse(female==0.5, (theta + theta/4), (theta - theta/4))
z <- sample(c(0,1), n, replace=TRUE)
y <- ifelse(z==1, y_treatment, y_control)
fake <- data.frame(y, z, female)
fit <- stan_glm(y ~ z + female + z:female, data=fake, refresh=0)
print(fit, digits=2)
```

If  $n = 125$  is enough to detect the main effect with 80% power, then  $n = 16 * 125 = 2000$  is needed for the interaction effect. You can show this using the following code:

```
tries <- 50
ledger <- array(NA, dim=c(tries, 2))
for (i in 1:tries) {
  n <- 2000 # required_n * 16
  female <- sample(c(-0.5, 0.5), n, replace=TRUE)
  y_control <- rnorm(n, 45, 20)
  y_treatment <- y_control +
    ifelse(female==0.5, (theta + theta/4), (theta - theta/4))
  z <- sample(c(0,1), n, replace=TRUE)
  y <- ifelse(z==1, y_treatment, y_control)
  fake <- data.frame(y, z, female)
  fit <- stan_glm(y ~ z + female + z:female, data=fake, refresh=0)
  ledger[i,] <- summary(fit)[["z:female", c("mean", "sd")]]
  if(i %% 5 == 0) cat("Rep:", i, "\n")
}
print(mean(abs(ledger[,1] / ledger[,2]) >= 1.96)) # power
```

## Drills

1. Sample size calculation for proportions

Calculate the sample size needed to achieve the stated statistical goal:

- (a) Estimate the proportion of Democratic voters, with a standard error of 1 percentage point.  
*Solution:* Start by guessing  $n = 100$ , then the standard error of the estimated proportion is approximately  $\sqrt{0.5 * 0.5 / 100} = 0.05$ . To get this down to 1 percentage point, we need to lower the standard error by a factor of 5, so we need to increase the sample size by a factor of  $5^2$ , thus the required sample size is 2500.
- (b) Estimate the difference in proportions of voters with conservative political ideology, comparing married and divorced people, with a standard error of 2 percentage points, assuming an equal number of people in each sample.
- (c) Estimate the difference in proportions of citizens in the United States and Canada who have personally experienced corruption, with an 80% probability of achieving “statistical significance,” from a survey with an equal number of people in each sample and a hypothesized true difference in proportions of 5 percentage points. For this problem you must also hypothesize a baseline proportion.

## 2. Sample size calculation for averages

Calculate the sample size needed to achieve the stated statistical goal:

- (a) Estimate the mean college admissions test scores of a population of high school students, with a standard error of 10 points. Assume the scores have a standard deviation of 150 in the population.  
*Solution:* Start by guessing  $n = 100$ , then the standard error of the estimated average is  $150 / \sqrt{100} = 15$ . To get this down to 10, we need to increase the sample size by a factor of  $(15/10)^2$ , thus the required sample size is 225.
- (b) Compare the average heights of women in two different countries, estimating the difference to within a standard error of 0.5 cm. Assume the population of women’s heights in each country has a standard deviation of 5 cm and that your sample sizes are equal for the two countries.
- (c) Estimate the difference in average happiness levels, between Republicans and Democrats, with an 80% probability of obtaining “statistical significance.” Assume that happiness is measured on a 1–10 scale and that the underlying average difference is 0.2 points on that scale. For this problem you must also hypothesize a population standard deviation for each group.

## Discussion problems

### 1. Choosing a survey design to estimate question-wording effects

Suppose you are conducting a survey and are concerned about the possible effects of the wording of one particular question.

You are offered the choice of doing one of two experiments: (a) Within-subject design: Put the two different wordings on the same survey form (randomizing the order of the two questions) and compare responses to the two wordings. (b) Between-subject design: Randomly give one wording to half the respondents and the other wording to the other half. Compare the average responses under the two wordings.

Give a reason why you might prefer design (a). Think about how you could perform a simulation on the computer to compare the two designs.

### 2. Implications and potential escapes from the trap of designing future studies based on past “statistically significant” estimates

Consider a world in which researchers report “statistically significant” estimates and then use these to design future studies. Work out the implications of this decision. Is there a way for the self-correcting nature of science to escape from this trap? How could you write a simulation to study this?

## 4.19 Poststratification and missing-data imputation

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Estimating state-level opinion	Generalizing from class to population	Regression and post-stratification	Poststrat and weighted averages	Network sampling
Environmental Sustainability Index	Experimental design and effect sizes	Random imputation	Methods for imputation	Problems with missing data

### Reading

Chapter 17 of *Regression and Other Stories*: Poststratification and missing-data imputation

### Pre-class warmup assignments

#### 1. Poststratification

For each problem, suppose an experiment has been performed in a set of middle schools, to study a new teaching style by comparing it to existing approaches, and that the results will be used to estimate the average effect, assuming a population that is 35% sixth-graders, 35% seventh-graders, and 30% eighth-graders.

- Suppose the estimated effect of the new style on standardized test scores is  $2.5 \pm 1.1$  points in sixth grade,  $-0.9 \pm 1.5$  points in seventh grade, and  $1.9 \pm 1.2$  points in eighth grade. Give the estimate and standard error of the average effect in the population.
- Suppose the estimated effect of the new style on standardized test scores is  $3.5 \pm 2.1$  points in sixth grade,  $-1.9 \pm 2.5$  points in seventh grade, and  $X \pm Y$  points in eighth grade. Suppose the estimate and standard error of the average effect in the population is 2.0 points with a standard error of 3.5. Then what are  $X$  and  $Y$ ?

#### 2. Missing data

Consider the following very small dataset:

	x	y	z
1	NA	NA	1
2	NA	-0.60	NA
3	NA	NA	NA
4	NA	NA	1
5	NA	0.30	0
6	1.11	NA	0
7	0.45	NA	2
8	-1.55	1.40	1
9	0.14	-0.38	1
10	0.11	-0.82	1

You are planning to fit a regression of  $y$  on  $z$  and a regression of  $y$  on  $x$  and  $z$ .

- Which data points would be included in a complete-case analysis?
- Which data points would be included in an available-case analysis?
- Which data points will be included if you fit these regressions in R?

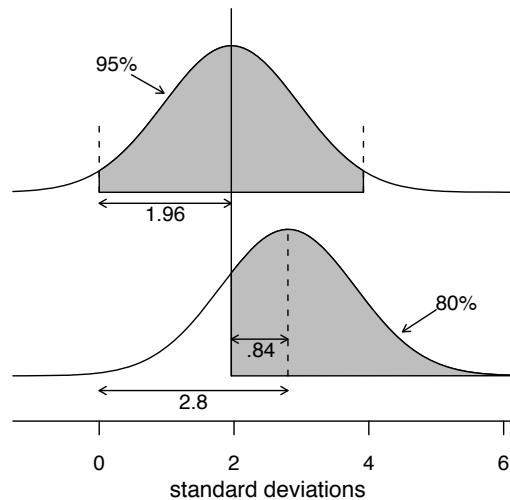


Figure 94 Sketch illustrating that, to obtain 80% power for a 95% confidence interval, the true effect size must be at least 2.8 standard errors from zero (assuming a normal distribution for estimation error). The top curve shows that the estimate must be at least 1.96 standard errors from zero for the 95% interval to be entirely positive. The bottom curve shows the distribution of the parameter estimates that might occur, if the true effect size is 2.8. Under this assumption, there is an 80% probability that the estimate will exceed 1.96. The two curves together show that the lower curve must be centered all the way at 2.8 to get an 80% probability that the 95% interval will be entirely positive.

### Homework assignments

1. (a) Power (Exercise 16.3 of *Regression and Other Stories*)

Following Figure 94, determine the power (the probability of getting an estimate that is “statistically significantly” different from zero at the 5% level) of a study where the true effect size is  $X$  standard errors from zero. Answer for the following values of  $X$ : 0, 1, 2, and 3.

(b) Power, type M error, type S error (Exercise 16.4 of *Regression and Other Stories*)

Consider the experiment shown in Figure 95 where the true effect could not realistically be more than 2 percentage points and is estimated with a standard error of 8.1 percentage points.

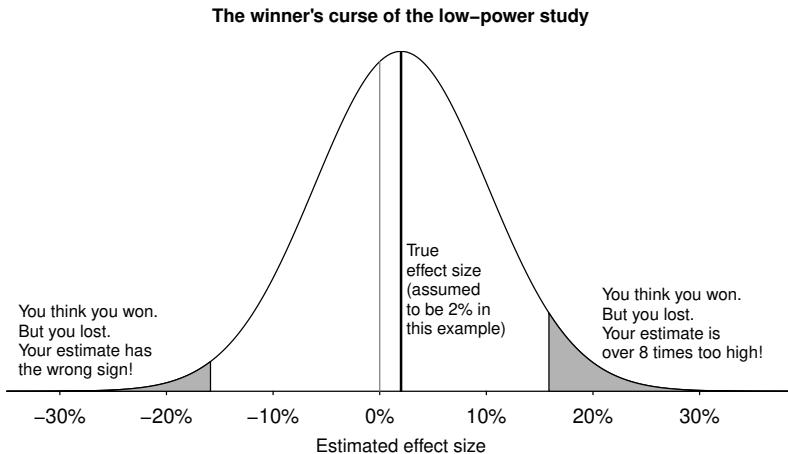
- i. Assuming that the estimate is unbiased and normally distributed and the true effect size is 2 percentage points, use simulation to answer the following questions: What is the power of this study? If only “statistically significant” results are reported, what is the expected type M error rate and what is the type S error rate?
- ii. Assuming that the estimate is unbiased and normally distributed and the true effect size is no more than 2 percentage points in absolute value, what can you say about the power, expected type M error, and type S error rate?

(c) Decline effect (Exercise 16.7)

After a study is published on the effect of some treatment or intervention, it is common for the estimated effect in future studies to be lower. Give five reasons why you might expect this to happen.

2. (a) Regression and poststratification (Exercise 17.1 of *Regression and Other Stories*)

Section 10.4 of *Regression and Other Stories* presents some models predicting weight from height and other variables using survey data in the folder Earnings. But these data are not representative of the population. In particular, 62% of the respondents in this survey are women, as compared to only 52% of the general adult population. We also know the approximate distribution of heights in the adult population: normal with mean 63.7 inches



**Figure 95** When the effect size is small compared to the standard error, statistical power is low. In this diagram, the bell-shaped curve represents the distribution of possible estimates, and the gray shaded zones correspond to estimates that are “statistically significant” (at least two standard errors away from zero). In this example, statistical significance is unlikely to be achieved, but in the rare cases where it does happen, it is highly misleading: there is a large chance the estimate has the wrong sign (a type S error), and in any case the magnitude of the effect size will be vastly overstated (a type M error) if it happens to be statistically significant. Thus, what would naively appear to be a “win” or a lucky draw—a statistically significant result from a low-power study—is, in the larger sense, a loss to science and to policy evaluation.

and standard deviation 2.7 inches for women, and normal with mean 69.1 inches and standard deviation 2.9 inches for men.

- i. Use poststratification to estimate the average weight in the general population, as follows:
  - (i) fit a regression of linear weight on height and sex, (ii) use `posterior_epred` to make predictions for men and women for each integer value of height from 50 through 80 inches, (iii) poststratify using a discrete approximation to the normal distribution for heights given sex and the known proportion of men and women in the population. Your result should be a set of simulation draws representing the population average weight. Give the median and mad sd (see Section 5.3 of *Regression and Other Stories*) of this distribution: this represents your estimate and uncertainty about the population average weight.
  - ii. Repeat the above steps, this time including the `height:female` interaction in your fitted model before poststratifying.
  - iii. Repeat (a) and (b), this time performing a regression of `log(weight)`, but still with the goal of estimating average weight in the population, so you will need to exponentiate your predictions in step (ii) before poststratifying.

**(b) In pairs:** Working through your own example (Exercise 16.14 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, think of a new data survey, experiment, or observational study that could be relevant and perform a design analysis for it, addressing issues of measurement, precision, and sample size. Simulate fake data for this study and analyze the simulated data.

## Stories

1. Using regression and poststratification to estimate public opinion by state

Section 17.1 of *Regression and Other Stories* introduces the method of regression and poststratification to adjust for differences between sample and population. There is a more advanced method,

multilevel regression and poststratification (MRP) that combines these ideas with what is called “small-area estimation” to get estimates for subpopulations for which we have small sample sizes. The idea of MRP is to build a regression model to make predictions for these subpopulations, “partially pooling” between the model and the data in each state.

MRP was first developed in the 1990s to estimate state-level opinion from national polls, and it is still used for this and related purposes. For example, in the 2019 British election, the *London Times* and *Guardian* referred to the seat-by-seat forecasts produced by an “MRP election polls.”

When using MRP to estimate state-level public opinion in the United States, the regression model should include state-level predictors. Consider a small state such as Vermont or Wyoming: a national poll of 5000 people might only include 10 respondents from each of these states, and the MRP-based estimates for these states will depend strongly on the model. Without state-level predictors, these estimates would be partially pooled toward the national average, which would not be appropriate. It typically makes sense to include as a predictor the Republican vote share in the state in the most recent presidential election: this way, the estimate for Vermont is partially pooled toward the other strongly Democratic-leaning states and the estimate for Wyoming is partially pooled toward the strongly Republican states.

But that one predictor won’t always do the job. For example, suppose you’re interested in attitudes toward gun control. Two natural state-level predictors are Republican vote share and percent rural. These variables are also highly correlated. But look at Vermont: it’s one of the most Democratic states and also the most rural. Vermont also is a small state, so the MRP inference for Vermont will depend strongly on the fitted model, which in turn will depend strongly on the coefficients for Republican vote share and percent rural. You may need some strong priors on these state-level coefficients to get a stable answer when fitting the regression.<sup>27</sup>

To elaborate, consider the following four MRP models:

- (a) No state-level predictors. This is bad because for states without much data, estimates are pooled toward the national mean. This is clearly the wrong thing to do for Wyoming, say, as this is a small state where attitudes toward gun control are probably far from the national average.
- (b) Republican vote share as a state-level predictor. Now the estimates are pooled toward the state-level regression model, which will estimate a negative effect of state-level Republican vote on gun control attitude. This will do the right thing for Wyoming. This model will partially pool Vermont to other strongly Democratic states such as California, Maryland, and Hawaii.
- (c) Percent rural as a state-level predictor. This should also fit the data well, with voters in more rural states being less likely to support gun control, and it should also do the right thing for Wyoming. This model will partially pool Vermont to other rural states such as Montana, North Dakota, and Mississippi.
- (d) Include Republican vote share and percent rural as two state-level predictors. This will do the right thing again for Wyoming and will split the difference on Vermont. The estimate for Vermont should also have a higher uncertainty, reflecting inferential uncertainty about the relative importance of the two state-level predictors.

Of all of these options, we think the last is the best—but this presupposes that we’re willing to use strong priors to control the estimation. If we’re only allowed to use weak priors—or, worse, not allowed to use any priors at all—then including both predictors could give noisy results.

This is where Bayesian inference comes in. It’s not just that we can now use prior information. It’s also that, by allowing the use of prior information, the Bayesian approach also opens the door to including more information into the model in the form of predictors.

<sup>27</sup> Andrew Gelman (2021), State-level predictors in MRP and Bayesian prior, <https://statmodeling.stat.columbia.edu/2021/04/08/state-level-predictors-in-mrp-and-bayesian-prior/>.

To put it another way, the first three models above are all special cases of the fourth model, but with very strong priors where one or two of the coefficients are assumed to be exactly zero. Paradoxically, putting Bayesian priors in the fourth model allows us to fit a bigger, more general model than would otherwise be realistically possible.

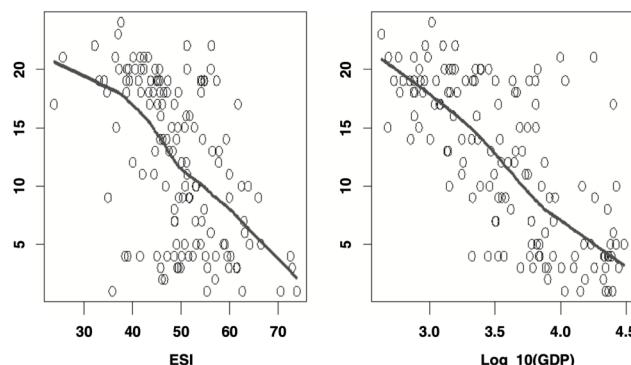
This story relates to the week's reading in providing a subtlety to regression and poststratification beyond the example in the textbook. It is relevant to the course as a whole in connecting technical decisions of statistical modeling and adjustment with subject-matter concerns—in this case, the unusual position of Vermont as a rural state whose residents mostly vote for Democrats. Similar issues arose in other MRP models when considering unusual states. For example, in a model for predicting attitudes on social issues, one state-level predictor that was used was percentage evangelical Protestant. But the model did not fit for Utah, a state with a large number of Mormons, a different conservative denomination. The problem was fixed by changing the state-level predictor to be the percentage of evangelical Protestants plus the percentage of Mormons in the state. In this case, it would not have worked to include these two as separate predictors, because pretty much the only information available to estimate the coefficient for percent Mormon would've come from Utah, which would ruin the whole partial-pooling approach.

## 2. Challenges with imputations of the Environmental Sustainability Index

The environmental sustainability index (ESI) was created as a measure of overall progress toward environmental sustainability and was designed in 1992 to permit systematic and quantitative comparison between nations. The ESI is a scaled linear combination of 64 variables of environmental concern. Environmental measures (such as oxide emissions and concentration) are included along with political indicators (such as civil liberty and level of corruption) that are relevant to environmental sustainability. The ESI, like other indices of environmental concern (such as the environmental wellbeing index and the human development index), condenses dissimilar social and physical metrics into cohesive summaries for national level comparisons.

Environmental data are often dissimilarly reported across regions or nations—rendering their quality poor, missing, or so incomparable that variables need to be treated as missing. When imputing missing values, we want to take into account context from the data that are observed. For example, if a variable is more likely to be missing for countries with low values of per capita gross domestic product, and this gross domestic product predictor is available for all countries, then we can use it as a predictor in a regression imputation.

Figure 96 shows the aggregate pattern of missingness in our data. Lower-income countries and countries with less environmental sustainability (as measured by our index) had more missing data.



**Figure 96** Aggregate pattern of missingness in the Environmental Sustainability Index. Each dot is a country; the plots show the number of missing variables (out of 64) in the data for the country, plotted vs. the country's environmental sustainability index and its per capita gross domestic product.

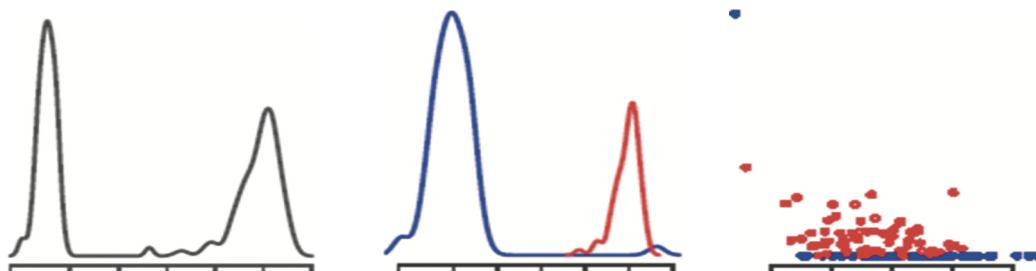


Figure 97 *Missing-data imputation for a variable in the Environmental Sustainability Index relating to water pollution: (a) distribution of observed and imputed data for all 142 countries in the data, (b) separating into imputed (the mode on the right half of the density plot) and observed (the large mode on the left of the plot and the small bump on the far right), (c) observed and imputed values plotted vs. index estimates for each country. These graphs show a problem, and so the imputation model was changed.*

We imputed missing values for the 2002 ESI using a three-step process, first using a statistical model to impute missing values, second exploring the observed and imputed data to reveal problems with the imputations, and in the third step improving the imputations. The result was far from perfect, and we see the main value of this story in demonstrating the bad news that imputation can create problems and the good news that sometimes these problems can be detected.<sup>28</sup> At first the idea of detecting imputation problems might seem impossible—after all, the whole point is that the values are missing, so how can they possibly be checked?—but it is indeed possible to diagnose problems with imputed data by looking at coherence with observed data.

Figure 97 shows an example, the observed and imputed data for a measure of the industrial and organic pollutants per available freshwater. Most of the observed data are in a narrow range. The exception is for Kuwait, which, as a net importer of freshwater, stands out as the point in the upper left of the scatterplot. The graphs show the problem with a naive default imputation model that fits a simple distribution to the observed data.

This story relates to the week’s reading by showing how problems in imputations can be diagnosed. It relates to the course as a whole as an example of the way that statistical analysis can be involved in the measurement process. We can follow up with a discussion of models that go into other summary measurements such as the Consumer Price Index and various political indexes.

### Class-participation activities

#### 1. Generalizing from the class to a larger population

For this activity, the instructor conducts a survey of the class and then attempts to use these responses to generalize to the larger population, poststratifying on age, sex, and education.

The first step is to choose three survey questions of interest—for example, one issue attitude, one experiential question, and one pop culture question—based on suggestions from the class. The next step is to put these on a Google form, along with background variables: questions on age, sex, and education.

Students fill out the form, and then the instructor can download the data and conduct the analysis. The first step is fitting a regression predicting each response given the demographics. The second step is to get predictions for all demographic categories in the population. Here we follow standard practice and bin age into the categories 18–29, 30–44, 45–64, and 65+, and bin education into

<sup>28</sup>Kobi Abayomi, Andrew Gelman, and Marc Levy (2008), Diagnostics for multivariate imputations, *Journal of the Royal Statistical Society C* 57, 273–291.

the categories no high school degree, high school degree or equivalent, some college, four-year college degree, and post-college education, so that we have  $4 \times 2 \times 5 = 40$  poststratification cells defined by age, sex, and education. The third step is to average this over the population. We have prepared the poststratification table ahead of time using the American Community Survey five-year sample.<sup>29</sup>

Here is our script for recoding the survey responses and reading in the census data:

```
responses <- read.csv("Poststratification example.csv")
  # Students' survey responses as saved on Google form
n <- nrow(responses)
library("dplyr")
age <- recode(responses$Age, "18-29" = 1, "30-44" = 2, "45-64" = 3, "65+" = 4)
male <- recode(responses$Sex, "Male" = 1, "Female" = 0)
educ <- recode(responses$Education, "No high school degree" = 1,
  "High school degree or equivalent" = 2, "Some college" = 3,
  "Four-year college degree" = 4, "Postgraduate degree" = 5)
census <- read.csv("census2.csv")
```

Suppose the first survey question is a binary response on attitude to gun control. Then here is code for regression and poststratification:

```
responses$gun_attitude <- ifelse(responses$Q1=="Yes", 1, 0)
fit_gun <- stan_glm(gun_attitude ~ factor(age) + male + factor(education),
  data=responses, family=binomial(link="logit"), refresh=0)
print(fit_gun)
pred_gun <- posterior_epred(fit_gun, newdata=census)
pred_gun_poststrat <- pred_gun %*% census$N / sum(census$N)
print(c(mean(pred_gun_poststrat), sd(pred_gun_poststrat)))
```

Unfortunately, this code, when applied to data from a class of university students, will produce an error. The problem is that there will be no data on the highest age categories and lowest education categories. So the usual regression can't even be fit!

The class can discuss possible solutions to this problem, along with the related scenario where very few responses are available in some categories. The problem with very few data can be seen by augmenting the dataset with some fake responses corresponding to people in the missing age and education categories and then fitting the model. The resulting predictions will have big uncertainties.

Beyond issues of uncertainty, we also need to be concerned with bias, to the extent that the sample (in this case, the students in the class) is unrepresentative of the population, even within these demographic categories. A possible solution is to supplement the data with survey responses from other sources.

The point is that there is no reasonable way of generalizing to the larger population when the same is so unrepresentative—at least not without outside information regarding responses in the missing groups.

This activity relates to the week's reading on poststratification, and it relates to the course as a whole by giving us a chance to discuss modeling assumptions in the context of a real applied problem, albeit simplified.

## 2. Experimental design and effect sizes

This activity demonstrates a basic principle of efficient design of experiments using a hypothetical study of people who are caring for their elderly relatives, where some training by a social worker

<sup>29</sup>See Juan Lopez-Martin, Justin Phillips, and Andrew Gelman (2022), Multilevel regression and poststratification case studies, <https://bookdown.org/jl5522/MRP-case-studies/>.

```
1. z <- sample(rep(c(0, 10), c(n/2, n/2)), n)
2. z <- sample(rep(c(0, 50), c(n/2, n/2)), n)
3. z <- sample(rep(c(0, 200), c(n/2, n/2)), n)
4. z <- sample(rep(c(0, 5, 10), c(n/3, n/3, n/3)), n)
5. z <- sample(rep(c(0, 25, 50), c(n/3, n/3, n/3)), n)
6. z <- sample(rep(c(0, 100, 200), c(n/3, n/3, n/3)), n)
7. z <- sample(runif(n, 0, 10))
8. z <- sample(runif(n, 0, 50))
9. z <- sample(runif(n, 0, 200))
```

Figure 98 *Code snippets representing nine different treatment assignment mechanisms for the caregiver design activity. Each pair of students will pick a random integer between 1 and 9 and use this to simulate the treatments for the n caregivers in their simulated-data experiment. The treatment z here represents the number of hours spent by a social worker for each caregiver in the hypothetical experiment.*

can reduce the stress of the caregiver. Imagine an experiment in which 60 caregivers are studied, with each being given a treatment  $z$  representing some number of hours of training by a social worker, and a year later the stress level of the caregiver is measured, perhaps by some combination of a questionnaire and observation by a trained social worker, yielding a continuous stress measure  $y$ . Assume for simplicity that no pre-treatment measurements are available. Further assume a linear model,  $y = a + bz + \text{error}$ . The treatments are randomly assigned to the 60 caregivers, and from the data we can estimate  $b$ , the effect on stress per hour of treatment.

In this activity, each pair of students performs an experiment—actually simulating the result of an experiment on a computer—and produces an estimate of  $b$  and a standard error. Each pair of students will simulate an experiment. The trick is that different pairs are given different distributions of treatment assignments.

We display on the screen Figure 98, which shows code for nine different treatment assignment mechanisms, and ask each each pair of students to open R on their laptops and choose a mechanism by simulating a random number between 1 and 9; in R, this would be `sample(1:9, 1)`.

Before actually doing the simulation, we ask students to discuss which of the nine designs will give the most precise estimate of the treatment effect, and which will give the noisiest estimate.

Then we ask each pair to simulate an experiment on  $n = 60$  caregivers by first simulating  $z$  from the line of code they've been given and then drawing a vector of outcome measurements from the model  $y = a + bz + \text{error}$ , with a baseline expected stress level  $a = 100$ , the average effect on stress level per hour of training being  $b = -0.1$ , and with the model error being normally distributed with mean 0 and standard deviation 40. Each pair should then fit the regression of  $y$  on  $z$  and look at their estimate and standard error.

Then compare results. Which groups had the smallest standard error? Which had the biggest? If all goes well, the smallest standard error should come from the design where half the caregivers are assigned 0 hours and half are assigned 200. This is the design that gives us the most leverage to estimate the slope.

The class can discuss the following questions: Are there problems with the design where all treatments are at the extreme values? What about potential nonlinearity in the treatment effect? Is it realistic to expect that a social worker could spend 200 hours on a single caregiver? To give good answers to these questions, students need to think carefully about what is being estimated.

This activity is relevant to the week's reading as an example of experimental design. It demonstrates the general principle that, when estimating an effect, it is best for the differences between treatment

and control group to be as large as possible. The activity relates to the course as a whole by bringing in realistic concerns about possible sizes of treatments. We can discuss with students how these issues arise in other settings.

### Computer demonstrations

#### 1. Regression and poststratification

The following code goes through a simple example of regression and poststratification, generalizing from a survey to the U.S. adult population. The survey data come from the Cooperative Election Study.<sup>30</sup> For convenience, we created a small file, `cces.csv`, with responses to just a few questions. Focus on the response to the question `pro_gun`: “Do you support the following proposal: make it easier for people to obtain concealed-carry permit?” First read in the data and compute the mean of this binary response and its standard error, for simplicity excluding missing responses:

```
cces <- read.csv("cces.csv")
ok <- !is.na(cces$pro_gun)
n <- sum(ok)
print(mean(cces$pro_gun[ok]))
print(sd(cces$pro_gun[ok])/sqrt(n))
```

The mean is 0.358—that is, 35.8% of the respondents answered Yes to this question, with a standard error of 0.002, or 0.2%. We ask the students in pairs to discuss how to think about this small uncertainty. It comes from the sample size being so large:  $n = 60\,984$ , and  $0.5/\sqrt{n} = 0.002$ . What can it mean to say that  $35.8\% \pm 0.2\%$  of Americans support concealed carry? Could this quantity ever really be estimated so accurately? Given that this example is coming up in the context of survey adjustment and missing data, students might point out that this survey is not a true random sample, and the nonsampling error could be much more than 0.2%. Beyond this, public opinion is not completely stable: Support for concealed carry could be 35.7% one day and 36.1% the next. At some point, the extra precision you get from a very large  $n$  is illusory.

In any case, you can move on to regression and poststratification. First fit a model predicting the survey response from three demographic variables:

```
fit_1 <- stan_glm(pro_gun ~ age + male + educ,
  data=cces, family=binomial(link="logit"), algorithm="optimizing")
print(fit_1)
```

We recommend fitting the model using the `optimizing` setting because it’s faster. When `stan_glm` is run in this way, it uses a normal approximation to the posterior distribution centered at a point estimate. In this case, the sample size is large and this approximation is fine.

You can then move to the poststratification step, first computing expected predicted values for the demographic categories we have gathered from the U.S. Census:

```
pred_1 <- posterior_epred(fit_1, newdata=census)
```

The result is a matrix with simulations representing uncertainty in the proportion of Yes responses in the population within each poststratification cell, as computed from the fitted model. Generalize to the population by weighting these cell estimates by the proportion of the population in each cell, which we can do using matrix multiplication:

```
pred_1_poststrat <- pred_1 %*% census$N / sum(census$N)
```

The result is a set of simulations representing the posterior distribution of the proportion of Yes respondents in the population. Take the mean and standard deviations of these simulations to represent the posterior estimate and uncertainty of this population proportion:

<sup>30</sup>Full data are at <https://cces.gov.harvard.edu>, with a subset at <https://bookdown.org/jl5522/MRP-case-studies/>.

#### 4.19. POSTSTRATIFICATION AND MISSING-DATA IMPUTATION

235

```
print(c(mean(pred_1_poststrat), sd(pred_1_poststrat)))
```

The resulting estimate is  $0.381 \pm 0.002$ , that is,  $38.1\% \pm 0.2\%$ . This estimate is higher than the raw survey mean of 35.8% that we computed above. Why? You can understand this by going back to the fitted logistic regression:

```
family:      binomial [logit]
formula:    pro_gun ~ age + male + educ
observations: 60984
predictors:  9
-----
              Median MAD_SD
(Intercept) -1.2   0.0
age30-44     0.1   0.0
age45-64     0.2   0.0
age65+       0.1   0.0
male         0.7   0.0
educHS       0.5   0.0
educNo HS    0.7   0.1
educPost-grad -0.3  0.0
educSome college 0.4   0.0
```

Yes responses are slightly more common among older respondents, but the strongest pattern is that Yes responses are more common among men and less-educated respondents. Surveys tend to undersample men and less-educated respondents, and you can see that here:

```
print(sum(census$N[census$male==1])/sum(census$N))
print(mean(cces$male))
```

The proportion of adults who are male is 48.7% from the Census and only 42.4% in the survey.

Similarly for education:

```
print(sum(census$N[census$educ=="No HS" | census$educ == "HS"])/sum(census$N))
print(mean(cces$educ=="No HS" | cces$educ=="HS"))
```

The proportion with high school education or less is 39.7% from the Census and only 30% in the survey.

The poststratification is a projection to a population that is more male and less educated than the sample, and this has the effect of increasing the estimate of support for concealed carry. Or, to put it another way, the survey undersamples men and less educated people and, as a result, understates support for this policy.

You can do a more elaborate version by including interactions in the regression:

```
fit_2 <- stan_glm(pro_gun ~ age + male + educ + age:male + age:educ + male:educ,
  data=cces, family=binomial(link="logit"), algorithm="optimizing")
print(fit_2)
pred_2 <- posterior_epred(fit_2, newdata=census)
pred_2_poststrat <- pred_2 %*% census$N / sum(census$N)
print(c(mean(pred_2_poststrat), sd(pred_2_poststrat)))
```

The resulting estimate is  $37.6\% \pm 0.2\%$ .

This is not the final word; real surveys are adjusted for other variables, such as ethnicity, region of the country, and party identification (although that last variable is tricky as it is not recorded in the Census), but the above calculations give a sense of how poststratification works and why we do it.

## 2. Random imputation

In this demonstration, first create fake data on log wages, education, experience, and sex. Then delete some of the log wage records (turning them into missing data), and pursue three different strategies to impute the missing values. As an add-on, you could compare the imputed values.

```
# Simulate fake data (with some missing), based on a previously estimated model
n <- 1000
exper <- runif(n, 0, 30)
female <- rbinom(n, 1, 0.49)
logwage <- 0.86 + 0.10 * female + 0.10 * educ + 0.01 * exper + rnorm(n, 0, 0.3)
missing <- sample(1:n), 50
logwage[missing] <- NA
fake <- data.frame(educ, exper, female, logwage)

# Random imputation
imputed_logwage_1 <- ifelse(is.na(logwage),
  sample(logwage[!is.na(logwage)], logwage)
# Deterministic imputation
fit_imp <- stan_glm(logwage ~ female + educ + exper, data=fake)
predictors <- fake[,1:3]
pred_det <- predict(fit_imp, newdata=predictors)
imputed_logwage_2 <- ifelse(is.na(logwage), pred_det, logwage)
# Random regression imputation
pred_ran <- posterior_predict(fit_imp, newdata=predictors, draws=1)
imputed_logwage_3 <- ifelse(is.na(logwage), pred_ran, logwage)
```

## Drills

### 1. Poststratification and weighted averages

Calculate population-level averages from the information about group-level averages as well as the groups' sizes.

(a) Life expectancy in the United States:

- $\bar{y}_{\text{male}} = 76$
- $\bar{y}_{\text{female}} = 81$
- $N_{\text{male}} = 161$  million
- $N_{\text{female}} = 169$  million

*Solution:*  $(161 * 76 + 169 * 81) / (161 + 189) = 74.1$

(b) Proportion of Americans without health insurance:<sup>31</sup>

- $\bar{y}_{\text{U.S. born}} = 7.5\%$
- $\bar{y}_{\text{foreign born}} = 19.6\%$
- Proportion U.S.-born = 86.3%
- Proportion foreign-born = 13.7%

### 2. Methods for imputation

Give an advantage and a disadvantage of each of the following approaches for imputing missing responses for a question in a survey, and come up with a scenario in which you would be comfortable using it. Explain your reasoning.

<sup>31</sup>From Abby Budiman, Christine Tamir, Lauren Mora, and Luis Noe-Bustamente (2020), Facts on U.S. immigrants, 2018, <https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants-current-data/>.

(a) Simple random imputation

*Solution:* Advantage: easy to do and to explain. Disadvantage: can give unreasonable imputations (for example, if education level is being imputed at random, you could impute post-graduate education to an 18-year-old).

(b) Deterministic regression imputation

(c) Random regression imputation

### Discussion problems

1. Network sampling

A researcher at Columbia University’s School of Social Work wanted to estimate the prevalence of drug abuse problems among American Indians (Native Americans) living in the New York City area.<sup>32</sup> From the Census, it was estimated that about 100 000 American Indians lived in the area, and the researcher had a budget to interview 400. She did not have a list of the population, so instead she planned to obtain her sample using network sampling, as follows.

She started with a list of approximately 300 members of a local American Indian community organization and planned to take a random sample of 100 from this list. She then planned to interview these 100 people and ask each of these to give her the names of other Indians in the area whom they knew. The respondents would also be asked to characterize themselves and the people whose names they supplied on a 0–10 scale, where 10 is “strongly Indian-identified,” 5 is “moderately Indian-identified,” and 0 is “not at all Indian identified.” Most of the original 100 people sampled would be expected to characterize themselves near 10 on the scale, given that they belonged to an Indian community organization. The researcher then planned to sample 100 people from the combined lists of all the people referred to by the first group, and then repeated the process twice more to obtain 400 people in total.

Describe how you would use the data from these 400 people to estimate, and get an uncertainty for your estimate of, the prevalence of drug abuse problems among American Indians living in New York City. You must account for the bias and dependence of the nonrandom sampling method.

2. Missing data imputation problems

The instructor can lead the class in a discussion of some missing data problems. To start, students divide into groups of four, and each group should come up with a problem they care about involving missing data. Missingness can arise for many reasons, including survey nonresponse, censoring (for example, not knowing the duration of an event because it hasn’t ended yet), incomplete measurement (for example, economic data not being available in all countries for all years of a dataset), and others. After the groups of four have discussed for a couple of minutes, the instructor reconvenes the class to discuss a few of their examples, using this as an opportunity to bring in the ideas of Sections 17.3–17.6 of *Regression and Other Stories*.

This activity is relevant for the week’s reading on missing-data imputation, and it relates to the course as a whole in calling attention to an important practical challenge in applied statistics.

<sup>32</sup>Teresa Evans-Campbell, Taryn Lindhorst, Bu Huang, and Karina Walters (2006), Interpersonal violence in the lives of urban American Indian and Alaska Native women: Implications for health, mental health, and help-seeking, *American Journal of Public Health* 96, 1416–1422.

## 4.20 How can flipping a coin help you estimate causal effects?

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Varying treatment effects	Potential outcomes for basketball study	Data analysis for basketball activity	Average treatment effects	Randomization and ethics
Ballot-order effects	Potential outcomes for ballot-order comparisons	Sample and population average effects	Uncertainty in treatment effects	Assumptions in randomized experiments

### Reading

Chapter 18 of *Regression and Other Stories*: Causal inference and randomized experiments

### Pre-class warmup assignments

#### 1. Average causal effects

Consider the following hypothetical table of pre-treatment predictor, treatment, and potential outcomes:

Unit $i$	Age, $x_i$	Treatment, $z_i$	Potential outcomes	
			$y_i^0$	$y_i^1$
Audrey	40	0	140	135
Anna	45	1	140	135
Bob	50	1	150	140
Bill	55	0	150	140
Caitlin	60	0	160	155
Cara	65	0	160	155
Dave	70	1	170	160
Doug	75	1	170	160

- What is the treatment effect for Audrey?
- What is the sample average treatment effect?
- What is the estimated average treatment effect, if it is estimated by a simple regression of the observed outcome,  $y$ , on the treatment indicator,  $z$ ?

#### 2. Simulate randomized experiments in R

Consider the following code to simulate a randomized experiment, starting with a vector  $x$  representing a pre-treatment predictor and vectors  $y0$  and  $y1$  representing potential outcomes. In a real experiment, we would not know both  $y0$  and  $y1$ , but this sort of simulation can be useful to understand the principles.

```
n <- length(x)
z <- sample(...)
y <- ifelse(z==0, y0, y1)
expt <- data.frame(x, y, z)
fit <- stan_glm(y ~ x + z, data=expt)
print(fit)
```

The coefficient for  $z$  in this fitted model is the estimated treatment effect. Assume  $n = 100$ .

## 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

239

In the above code, the line `z <- sample(...)` is a placeholder. Replace it with correct code for each of the following scenarios:

- (a) Completely randomized experiment: 50 get treatment, 50 get control.
- (b) Complete randomization but with 33 getting treatment and 67 getting control.
- (c) Random assignment with unequal probabilities, where the probability of getting the treatment is  $\text{logit}^{-1}(x)$ .

### Homework assignments

#### 1. (a) Bias in deterministic imputation (Exercise 17.4 of *Regression and Other Stories*)

Suppose you are imputing missing responses for income in a social survey of American households, using for the imputation a regression model given demographic variables. Which of the following two statements is basically true?

- i. If you impute income deterministically using a fitted regression model (that is, imputing using  $X\beta$  rather than  $X\beta + \epsilon$ ), you will tend to impute too many people as *rich* or *poor*: a deterministic procedure overstates your certainty, making you more likely to impute extreme values.
- ii. If you impute income deterministically using a fitted regression model (that is, imputing using  $X\beta$  rather than  $X\beta + \epsilon$ ), you will tend to impute too many people as *middle class*: by not using the error term, you'll impute too many values in the middle of the distribution.

#### 2. (a) Designing an experiment (Exercise 18.1 of *Regression and Other Stories*)

Suppose you are interested in the effect of the presence of vending machines in schools on childhood obesity. What controlled experiment would you want to do (in a world without ethical, logistical, or financial constraints) to evaluate this question?

#### (b) Designing an experiment with ethical constraints (Exercise 18.2 of *Regression and Other Stories*)

Suppose you are interested in the effect of smoking on lung cancer. What controlled experiment could you plausibly perform (in the real world) to evaluate this effect?

#### (c) In pairs: Working through your own example (Exercise 17.11 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, perform multiple imputation for missing data (in earlier analyses, you may have just excluded cases with missingness) and then compare one of your earlier regression results to what you obtain with the imputed datasets.

### Stories

#### 1. Treatment effect depends on the population: coronavirus example

Part of designing a study is accounting for uncertainty in effect sizes. Unfortunately, there is a tradition in clinical trials of making optimistic assumptions in order to claim high power. Here is an example that came up in March, 2020.<sup>33</sup> A doctor was designing a trial for an existing drug that he thought could be effective for high-risk coronavirus patients. He contacted us to check his sample size calculation: under the assumption that the drug increased survival rate by 25 percentage points, a sample size of  $n = 126$  would assure 80% power. (With 126 people divided evenly in two groups, the standard error of the difference in proportions is bounded above by  $\sqrt{0.5 * 0.5/63 + 0.5 * 0.5/63} = 0.089$ , so an effect of 0.25 is at least 2.8 standard errors from zero, which is the condition for 80% power for the comparison.) When we asked the doctor

<sup>33</sup>See Jon Zelner, Julien Riou, Ruth Etzioni, and Andrew Gelman (2021), Accounting for uncertainty during a pandemic, *Patterns* 2, 100310.

Hypothetical scenario of 1000 people:

- 300 would live either way
- 450 would die either way
- 250 would be saved by the treatment

Average treatment effect: 25 percentage points

**Figure 99** Hypothetical scenario of a potentially lifesaving treatment with a 25 percentage point effect. We display this diagram to demonstrate to students that this does not represent “a 25 percentage point benefit” for each patient; rather, it’s a benefit for 25% of the patients.

how confident he was in his guessed effect size, he replied that he thought the effect on these patients would be higher and that 25 percentage points was a conservative estimate. At the same time, he recognized that the drug might not work. We asked the doctor if he would be interested in increasing his sample size so he could detect a 10 percentage point increase in survival, for example, but he said that this would not be necessary, because if it did work, he thought it would increase the survival rate by at least 25 percentage points.

It might seem reasonable to suppose that a drug might either have a large effect or not work at all. But this vision of uncertainty has problems. Suppose, for example, that the survival rate was 30% among the patients who do not receive this new drug and 55% among the treatment group. Then in a population of 1000 people, it could be that the drug has no effect on the 300 of people who would live either way, no effect on the 450 who would die either way, and it would save the lives of the remaining 250 patients. There are other possibilities consistent with a 25 percentage point benefit—for example the drug could save 350 people while killing 100—but it makes sense to stick with the simple scenario for now. In any case, the point is that the posited benefit of the drug is not “a 25 percentage point benefit” for each patient; rather, it’s a benefit for 25% of the patients. In explaining this to the class, the instructor can construct on the board the chart shown in Figure 99 and elaborate as necessary.

From that perspective, once we accept the idea that the drug works on some people and not others—or in some comorbidity scenarios and not others—we realize that “the treatment effect” in any given study will depend entirely on the patient mix. There is no underlying number representing the effect of the drug. Ideally one would like to know what sorts of patients the treatment would help, but in a clinical trial it is enough to show that there is some clear average effect. The point is that if you consider the treatment effect in the context of variation between patients, this can be the first step in a more grounded understanding of effect size.

This story relates to the week’s reading as an example where the potential-outcome framework is directly helping us understand what is going on.<sup>34</sup> It is relevant to the course as a whole in connecting this model of causal inference to experimental design in a context of varying treatment effects, which in turn relates to more general issues in experimental design and extrapolation.

## 2. Effects of ballot order on voting

Did Donald Trump win the 2016 election because his name came first in the ballot in key states? We’re moderately skeptical, but let’s look at the evidence.<sup>35</sup>

<sup>34</sup>See also Andrew Gelman (2021), A counterexample to the potential-outcomes model for causal inference, <https://statmodeling.stat.columbia.edu/2021/07/26/causal-counterexample/>.

<sup>35</sup>See Andrew Gelman (2017), Did Trump win because his name came first in key states? Maybe, but I’m doubtful, <https://statmodeling.stat.columbia.edu/2017/02/27/name-first-doubtful/>; Andrew Gelman (2018), A Bayesian take on ballot order effects, <https://statmodeling.stat.columbia.edu/2018/11/21/bayesian-take-ballot-order-effects/>; Andrew Gelman (2019), Ballot order update, <https://statmodeling.stat.columbia.edu/2019/04/25/ballot-order-update/>; and Andrew Gelman (2019), Ballot order effects in the news: I’m skeptical of the claimed 5% effect, <https://statmodeling.stat.columbia.edu/2019/11/15/ballot-order-effects-in-the-news-im-skeptical-of-the-claimed-5-effect/>.

#### 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

241

The claim came from a BBC news report in early 2017:<sup>36</sup>

“One of the world’s leading political scientists believes Donald Trump most likely won the US presidential election for a very simple reason . . . his name came first on the ballot in some critical swing states.

Jon Krosnick has spent 30 years studying how voters choose one candidate rather than another, and says that ‘at least two’ US presidents won their elections because their names were listed first on the ballot, in states where the margin of victory was narrow. . . . ‘There is a human tendency to lean towards the first name listed on the ballot,’ says Krosnick, a politics professor at Stanford University. ‘And that has caused increases on average of about three percentage points for candidates, across lots of races and states and years.’ . . . When an election is very close the effect can be decisive, Krosnick says—and in some US states, such as Pennsylvania, Michigan and Wisconsin, the 2016 election was very close.”

As noted in the news article, Trump seems to have been listed first on the ballot in Michigan and Wisconsin.

What about the other close states in that election? We looked up their ballot rules on the internet. In Minnesota, it looks like Trump was first on the ballot, and he did almost come from behind to win that state in 2016.

Florida and Pennsylvania appeared to list the candidate of the governor’s party first, which would put Trump first in Florida and Clinton first in Pennsylvania. New Hampshire rotates candidate name order, which would imply no large aggregate effect there.<sup>37</sup>

Suppose ballot order gave Trump the win in Michigan, Wisconsin, and Florida. That’s  $16+10+29 = 55$  electoral votes. On the other side, there were no close states that Clinton won where she had a ballot-order advantage. Take away 55 of Trump’s electoral votes and he no longer has the victory (assuming all electoral voters voted as pledged). We tend to think of all these little things as averaging out, but they don’t have to. The number of swing states is small.

So, yeah, maybe Krosnick is right on this one. It all comes down to Florida: if the ballot-order effect were enough to swing that state, it also likely would have swung the closer elections in Michigan and Wisconsin.

In 2016, Trump beat Clinton in Florida by 1.2% of the vote. Could ballot order have been enough to cause a 1.2% swing? Maybe so, maybe not. The research is mixed. Analyzing data from California elections where a rotation of candidate orders was used across assembly districts, a study by Krosnick and others found large effects including in the 2000 presidential race. But a different analysis of California elections a few years later concluded that “in general elections, ballot order significantly impacts only minor party candidates, with no detectable effects on major party candidates.”<sup>38</sup> The authors of that California study also point out that the analysis of Krosnick et al. is purely observational. That said, much can be learned from observational data. Krosnick et al. analyzed data from the 80 assembly districts but it doesn’t look like they adjusted for previous election results in those districts, which would be the obvious thing to do in such an analysis. So their problem was not so much in using observational data but in not adjusting for a key pre-treatment predictor.

Ballot-order effects have been studied using various data. An analysis of Australian elections

<sup>36</sup>BBC News (2017), Did Trump win because his name came first in key states?, <https://www.bbc.com/news/magazine-39082465/>.

<sup>37</sup>Data from New Hampshire are analyzed by Bo MacInnes, Joanne Miller, Jon Krosnick, Clifton Below, and Miriam Lindner (2021), Candidate name order effects in New Hampshire: Evidence from primaries and from general elections with party column ballots, *PLoS One* 16, e0248049. However, we are not convinced by their conclusions regarding name-order effects in the general elections for president.

<sup>38</sup>Daniel Ho and Kosuke Imai (2008), Estimating causal effects of ballot order from a randomized natural experiment: The California alphabet lottery, 1978–2002, *Public Opinion Quarterly* 72, 216–240.

found that “being placed first on the ballot increases a candidate’s vote share by about 1 percentage point.”<sup>39</sup> A study of local contests in the United States found ballot order effects of 4 to 5 percentage points on city council and school board elections, but this is not so relevant for the presidential race.<sup>40</sup> A literature search found many papers on ballot-order effects but mostly on local elections or primary elections, where such effects would be expected to be larger. Some of the research goes back to the early 1900s.<sup>41</sup>

So, putting all the evidence together: what do we think? As noted above, it all comes down to Florida. In 2000, Florida was extremely close—the best estimate has Gore winning by only about 30 000 votes—according to political scientist Walter Mebane, the votes were lost “primarily due to defective election administration in the state”—and had ballot order been randomized Gore could well have won by even more, enough for the state to have counted in his favor in the electoral college.<sup>42</sup>

In 2016, maybe, maybe not. Based on the literature we’ve seen, a swing of 1 percentage point seems to be on the border of what might be a plausible ballot-order effect for the general election for president, maybe a bit on the high end given our current level of political polarization. So we think Krosnick is overstating the case, but it is just possible that the ballot order effects were large enough that, had the ballots been randomized, Clinton could have won Florida as well as Michigan, Wisconsin, and thus the electoral college.

This story is relevant to the week’s reading because it demonstrates the potential-outcome model of causality: causal inference is about predicting what would’ve happened under alternative states of the world. It relates to the course as a whole in that we sift through results from various studies of different data sources in order to adjudicate a research and policy claim.

### Class-participation activities

#### 1. Potential outcomes and treatment assignments for basketball training

Chapter 18 of *Regression and Other Stories* has a recurring example with a table of pre-treatment variables  $x$ , treatments  $z$ , and potential outcomes  $y^0$  and  $y^1$  for a hypothetical blood pressure study with 8 people. As a class-participation activity, we set up a similar structure with the students in our class, using an example of sports training.

We explain to the class that we would be setting up a potential-outcome table for a hypothetical sports example as displayed in Figure 100. Students are told to imagine themselves going to the basketball court in a month and taking 50 free throws, with  $y^0$  being the number of shots they would make if they do no preparation and  $y^1$  being the number of shots they would make if they were to go to the court and practice for 15 minutes each day for a month. This is not a true experiment because we are asking students to guess how they *might* respond—we are not actually performing the treatments or measuring the outcome. The point here is to have that full table to demonstrate the principles of causal inference. We ask students to complete the exercise with a pre-treatment variable, an assessment of their athleticism on a 1–10 scale. We give them the URL for a Google form asking their name, age, self-assessed athleticism, and the two potential outcomes; we then download the data and project all the information on to the screen, where it remains during our discussion.

The next step is to consider several different designs—approaches for assigning treatment or control to the participants in this hypothetical experiment. For each design, we compute the simple

<sup>39</sup>Amy King and Andrew Leigh (2009), Are ballot order effects heterogeneous?, *Social Science Quarterly* 90, 71–87.

<sup>40</sup>Marc Meredith and Yuval Salant (2013), On the causes and consequences of ballot order effects, *Political Behavior* 35, 175–197.

<sup>41</sup>See R. Darcy and Ian McAllister (1990), Ballot position effects, *Electoral Studies* 9, 5–17.

<sup>42</sup>Walter Mebane (2004), The wrong man is President! Overvotes in the 2000 presidential election in Florida, *Perspectives on Politics* 2, 525–535.

## 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

243

- Potential outcomes,  $y$ :
- $y^1$ : number of free throws you make out of 50 tries, if  $z = 1$
  - $y^0$ : number of free throws you make out of 50 tries, if  $z = 0$
- Treatment,  $z$ :
- $z = 1$ : practice for 15 minutes each day for a month
  - $z = 0$ : no practice
- Pre-treatment predictors,  $x$ :
- Age
  - Self-assessed athleticism (on 1–10 scale)

Figure 100 *Outline of a hypothetical experiment on sports training, used as an example of potential outcomes. We display this outline to students to help them understand the different variables in the model.*

estimate  $\bar{y}^1 - \bar{y}^0$  and its standard error. But we first calculate the true average treatment effect by averaging the values of  $y^1 - y^0$  for everyone in the class. This is something that would not usually be known, because in general it is not possible to give anyone both treatments.

We then consider some designs, each of which can be implemented in R:

- An independent randomization in which each participant is equally likely to receive treatment or control. To get things started, we simulated this basic design, computed the resulting  $\bar{y}^1 - \bar{y}^0$  and its standard errors, and compared this result to the students' expectations.
- A biased treatment assignment where participants get to decide whether to do the treatment or the control. It seems plausible that people who are more athletic would be more likely to do regular practices, and we can simulate such a design using a rule such as  $\Pr(z = 1) = \text{logit}^{-1}(x - 5.5)$ . We ask students how they think this will bias the treatment estimate: What direction will the bias be and how large? We then simulate this treatment assignment rule, compute the resulting  $\bar{y}^1 - \bar{y}^0$  and its standard errors, and compare the result to the students' expectations.
- We next construct a rule that's biased in the other direction. Suppose the treatment is not chosen by the participants but instead is assigned by a coach, who gives the practice to those who seem to really need it. This can be simulated using the rule  $\Pr(z = 1) = \text{logit}^{-1}(-(x - 5.5))$ . We again ask the class to guess the direction and magnitude of the bias and follow up with the simulation.
- What about an alternating design, where participants are alternately given control and treatment, using the order in which the forms were returned?
- If there is time, students can be asked if there are any other designs they'd be interested in considering, for example a non-randomized design without a known treatment assignment rule.

This activity relates to the week's reading by bringing to life the contingent nature of potential outcomes. It relates to the course as a whole by showing the connection between biased data collection and biased inference.

### 2. Potential outcomes and randomization

A class can experience the sampling distribution of a randomized experiment by having each pair

<sup>43</sup>Jim Saunders (2020), Court refuses to reconsider ballot order ruling, <https://cbs12.com/news/local/court-refuses-to-reconsider-ballot-order-ruling/>.

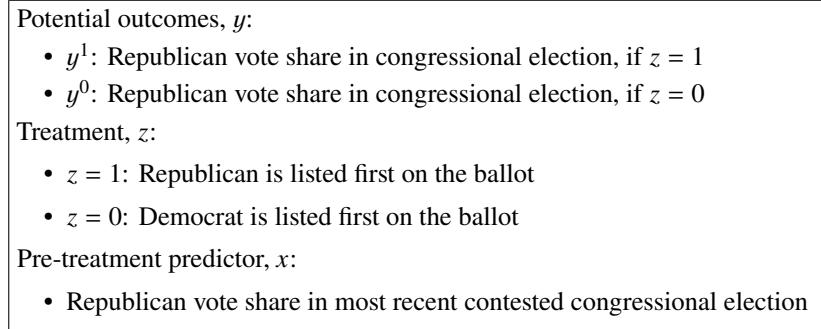


Figure 101 Structure of a hypothetical study to estimate the effect of ballot order on congressional election outcomes.

dist	r2020	rprev	yearprev	dist	y0	y1	dist	y0	y1	z	y
1	0.655	0.671	2018	1	0.635	0.655	1	0.635	0.655	1	0.655
2	1.000	0.674	2018	2			2				
3	0.571	0.576	2018	3	0.551	0.571	3	0.551	0.571	0	0.551
4	0.611	0.668	2018	4	0.591	0.611	4	0.591	0.611	0	0.591
5	0.349	0.332	2018	5	0.329	0.349	5	0.329	0.349	1	0.349
.	.	.	.	.	.	.	.	.	.	.	.

Figure 102 (a) First few rows of data from the file `florida2020data.csv` showing the Republican vote share for each of the state's 27 U.S. congressional districts in 2020, along with the Republican vote share in that district in the most recent previous contested congressional election, and the year of that election. In most cases, the most recent contested election occurred in 2018, but in some districts it was in 2016 or 2014. (b) Potential outcomes under the control (Democratic candidate listed first on the ballot in that district) or treatment (Republican listed first), under the assumption that the treatment effect is exactly 2% of the vote in every district, excluding the election in district 2, which was uncontested. (c) Possible outcome of an experiment in which ballot position is assigned randomly by district.

of students perform their own experiment on the computer and then comparing the results. For example, consider a hypothetical ballot-order effect in Florida. In that state, the candidate of the governor's party is listed first on all ballots.<sup>43</sup>

But suppose that Florida followed California's rule and randomly chose the order of candidates in districts.<sup>44</sup> This would be fairer; also it would provide data that would allow us to estimate ballot-order effects. To get a sense of this, we constructed a dataset of Florida's 27 congressional districts in the 2020 election that you can use as the basis for a simulated experiment.

The instructor can start by displaying Figure 101, which places the problem in the potential-outcome framework. The next step is to input the data and take a look.

```
data2020 <- read.csv(paste0(
  "http://www.stat.columbia.edu/~gelman/regression_course/",
  "florida2020data.csv"
))
print(data2020)
```

The left panel of Figure 102 shows the first few rows of the dataset, which has district number, Republican share of the two-party vote in 2020, Republican vote share in that district in the most recent previous contested congressional election, and the year of that election. In most cases, the most recent contested election occurred in 2018, but in some districts it was in 2016 or 2014.

<sup>44</sup>California Secretary of State (2015), Randomized alphabet, <https://www.sos.ca.gov/elections/randomized-alphabet/>; see also Darren Grant (2016), The ballot order effect is huge: Evidence from Texas, *Public Choice* 172, 441–442.

## 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

245

In this activity, students will consider alternative outcomes in 2000 under different assignments of ballot order. In doing this, set aside district 2 and another district that had an uncontested election in 2020, as there is no ballot order effect if there is only one major-party candidate. It is simplest to work with a smaller dataset including only the elections that were contested by both parties in 2020:

```
contested <- (data2020$r2020 > 0) & (data2020$r2020 < 1)
data <- data2020[contested, ]
```

Next, potential outcomes can be defined under the assumption that being listed first on the ballot is worth 1% of the vote. In that case, switching the ballot order in a congressional race from Republican first to Democrat first would decrease Republican vote share by 2%. You can define the outcome under the control condition  $y^0$  if the Democratic candidate had been listed first on the ballot and the treatment condition  $y^1$  if the Republican had been listed first, which is what actually happened in the Florida races:

```
data$y1 <- data$r2020
data$y0 <- data$y1 - 0.02
```

The center panel of Figure 102 shows the result for the first few districts in our dataset. You can then simulate a randomized treatment assignment and the resulting election outcome:

```
data$z <- rbinom(nrow(data), 1, 0.5)
data$y <- ifelse(data$z==1, data$y1, data$y0)
```

The first line of code assigns the treatment randomly; the second line chooses the appropriate potential outcome to be observed. The right panel of Figure 102 shows one possible result.

The way we do this in class is to share the URL of the data file and go through these steps on our computer, displaying on the screen, and then asking students to pair up, and each pair should go into R, copy in the code, and simulate their own experiment on the 25 Florida districts that were contested in 2020.

Each pair, once they have simulated the treatment assignment and created the observed  $y$ , should estimate the treatment effect. Each pair can then announce their estimate and standard error.

Here are two ways of estimating the treatment effect given data  $y$  (but not the potential outcomes  $y^0$  and  $y^1$ , as that would be cheating). First is the simple difference between treatment and control groups, equivalently the regression of the outcome on the treatment indicator:

```
fit <- stan_glm(y ~ z, data=data, refresh=0)
print(fit, digits=2)
```

We can do better by adjusting for the pre-treatment variable:

```
data$x <- data$rperv
fit2 <- stan_glm(y ~ z + x, data=data, refresh=0)
print(fit2, digits=2)
```

Figure 103 shows the sorts of results you can expect to see. In this case, the pre-treatment variable  $x$  is a very strong predictor—past results are a good predictor of future elections—and this can be seen in the smaller standard error for  $z$  after adjusting for  $x$  and also the much smaller residual standard deviation.

You can also do a similar simulation but replacing the constant 2% treatment effect by an effect that varies between 0 and 4% of the vote:

```
data$y0 <- data$y1 - runif(nrow(data), 0, 0.04)
data$y <- ifelse(data$z==1, data$y1, data$y0)
print(stan_glm(y ~ z, data=data, refresh=0), digits=2)
print(stan_glm(y ~ z + x, data=data, refresh=0), digits=2)
```

#### 4. WEEK BY WEEK: THE SECOND SEMESTER

```

family: gaussian [identity]
formula: y ~ z
observations: 25
predictors: 2
-----
      Median MAD_SD
(Intercept) 0.53  0.04
z            0.08  0.05

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.13  0.02

family: gaussian [identity]
formula: y ~ z + x
observations: 25
predictors: 3
-----
      Median MAD_SD
(Intercept) 0.07  0.02
z           0.01  0.01
x           0.89  0.04

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.02  0.00

```

**Figure 103** *Estimating the effect of a treatment (in this case, ballot order) on Republican vote share in Florida congressional elections based on a hypothetical experiment. This figure, to be displayed to the class, shows two estimates: (a) regression of outcome on treatment indicator, which is equivalent to average difference between treatment and control groups; (b) regression also adjusting for pre-treatment variable (in this case, Republican vote share in a previous election). From our construction of the data, we know that the true treatment effect is  $-0.02$ , but the key point here is that the estimated treatment effect (the coefficient of  $z$ ) has a much lower standard error after adjusting for  $x$ .*

This activity relates to the week's reading as an example of potential outcomes and randomized experiments. It relates to the course as a whole by illustrating the way that such experiments can be used to estimate a causal effect, and also demonstrating the challenges of such estimation when the effect size is small, data are variable, and the sample size is low.

### Computer demonstrations

#### 1. Data analysis for the basketball training activity

Here is the code for cleaning and analyzing the data collected from students in the potential outcomes activity on page 242. The demonstration starts with the data file from the students' survey responses and then goes from there.

```

# Read in and clean data
data <- read.csv("Potential outcomes sports training example.csv")
n <- nrow(data)
x <- data[, "x...Athleticism..1...not.athletic.to.10...very.athletic."]
y1 <- data[, "y1...Number.of.shots..out.of.50..you.would.make.with.training"]
y0 <- data[, "y0...Number.of.shots..out.of.50..you.would.make.with.NO.training"]
true_effect <- mean(y1 - y0)
print(true_effect)

# Flip coin to assign treatment or control to each student
z <- rbinom(n, 1, 0.5)
y <- ifelse(z==1, y1, y0)
minidata <- data.frame(x, y, z)

# Simple comparison and se
est <- mean(y[z==1]) - mean(y[z==0])
s1 <- sd(y[z==1])
s0 <- sd(y[z==0])
n1 <- sum(z==1)
n0 <- sum(z==0)
se <- sqrt(s1^2/n1 + s0^2/n0)

```

## 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

247

```
print(c(est, se))

# Equivalently, run regression and look at coef and se for z
fit <- stan_glm(y ~ z, data=minidata, refresh=0)
print(fit)

# Biased treatment assignment
curve(invlogit(x - 5.5), from=1, to=10)
p_treat <- invlogit(x - 5.5)
z <- rbinom(n, 1, p_treat)
y <- ifelse(z==1, y1, y0)
minidata2 <- data.frame(x, y, z)
fit2 <- stan_glm(y ~ z, data=minidata2, refresh=0)
print(fit2)

# Adjust for pre-treatment variable: this estimate should be better (on average).
fit2a <- stan_glm(y ~ z + x, data=minidata2, refresh=0)
print(fit2a)

# Biased in the other direction
curve(invlogit(-(x - 5.5)), from=1, to=10)
p_treat <- invlogit(-(x - 5.5))
z <- rbinom(n, 1, p_treat)
y <- ifelse(z==1, y1, y0)
minidata3 <- data.frame(x, y, z)
fit3 <- stan_glm(y ~ z, data=minidata3, refresh=0)
print(fit3)

# Adjust for pre-treatment variable: this estimate should be better (on average).
fit3a <- stan_glm(y ~ z + x, data=minidata3, refresh=0)
print(fit3a)
```

### 2. Sample and population average treatment effects

Consider the treatment “publicly available tutoring,” which might have less of an effect on wealthy people who might already be drawing on other resources. Imagining that wealthy people are nevertheless overrepresented in an experiment that assesses the impact of tutoring, you can show how the sample average treatment effect (SATE) differs from the population average treatment effect (PATE):

```
# Publicly available tutoring
n <- 1000
wealthy <- c(rep(0, 800), rep(1, 200))
treatment_effect <- ifelse(wealthy == 1, 5, 15)
print(treatment_effect)
y_0 <- rnorm(n, 100, 15)
y_1 <- y_0 + treatment_effect
pop <- data.frame(y_0, y_1)
PATE <- mean(pop$y_1) - mean(pop$y_0)
print(PATE)

# Sample, with wealthy people overrepresented
sample_n <- 50
selected <- c(sample(1:800, sample_n/2), sample(801:1000, sample_n/2))
sample <- pop[selected, ]
sample$z <- sample(c(0,1), sample_n/2, replace=TRUE)
treated <- sample$y_1[sample$z==1]
```

```
control <- sample$y_0[sample$z==0]
SATE <- mean(treated) - mean(control)
print(SATE)
```

## Drills

### 1. Average treatment effects

Suppose you run this code:

```
fit <- stan_glm(post_test ~ z + pre_test + z:pre_test, data=expt)
print(fit)
```

and get the following result:

```
formula:      post_test ~ z + pre_test + z:pre_test
observations: 100
predictors:   3
-----
           Median MAD_SD
(Intercept) 23.6   10.9
z            10.4    4.0
pre_test     0.7    0.2
z:pre_test   -0.4   0.3

Auxiliary parameter(s):
           Median MAD_SD
sigma 20.1    1.4
```

You also have a data frame, pop, representing the population.

- (a) Give R code to compute the estimated sample average treatment effect (SATE).

*Solution:*  $10.4 - 0.4 * \text{mean(expt$pre\_test)}$

- (b) Give R code to compute the estimated population average treatment effect (PATE).

- (c) Give R code to compute the estimated PATE among students who scored more than 50 on the pre-test.

### 2. Average treatment effects with uncertainty

Suppose you run the following to estimate the effect of a treatment,  $z$ :

```
fit <- stan_glm(post_test ~ z + pre_test + z:pre_test, data=expt)
```

You also have a data frame, pop, representing the population.

- (a) Give R code to compute the estimate and standard error of the sample average treatment effect (SATE), using `posterior_epred`:

*Solution:*

```
expt_0 <- expt
expt_0$z <- 0
expt_1 <- expt
expt_1$z <- 1
treatment_effect <- posterior_epred(fit, newdata=expt_1) -
  posterior_epred(fit, newdata=expt_0)
SATE <- rowMeans(treatment_effect)
print(c(mean(SATE), sd(SATE)))
```

- (b) Give R code to compute the estimate and standard error of the population average treatment effect (PATE)

#### 4.20. HOW CAN FLIPPING A COIN HELP YOU ESTIMATE CAUSAL EFFECTS?

249

- (c) Give R code to compute the estimate and standard error of the PATE among students who scored more than 50 on the pre-test.

#### Discussion problems

##### 1. Randomization and ethics

Consider a medical experiment in which people are randomly assigned to existing protocol or a new treatment. Several ethical challenges can arise, for example: (1) the experimental treatment could be risky; (2) from the other direction, if the new treatment is believed to be better, it could seem unfair to give someone the control; (3) it is unclear what standard of evidence should be required for the treatment to be deemed effective enough to be approved for public use; (4) it is unclear how to balance risks and benefits; (5) the new treatment could be very slightly more effective but much more expensive; (6) it must be determined how to compare to treatments that have not been evaluated using randomized experiments. Discuss these issues, which can also apply to educational and social experiments.

##### 2. Assumptions in randomized experiments

Take a particular example of a randomized experiment in an area of interest to you—this could be a real experiment that has been done, or a hypothetical experiment on some treatment or exposure—and consider each of the following issues: ignorability, efficiency, the stable unit treatment value assumption, external validity, internal validity, missing data, and noncompliance.

## 4.21 Causal inference using regression on the treatment variable

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Pest control experiment	Adjustments in causal inference	Benefits of pre-treatment data	Average effect and post-stratification	Causal logistic regression
Social penumbras	Average treatment effect	Combining pre-treatment predictors	Average effect for nonlinear models	Holding all else equal?

### Reading

Chapter 19 of *Regression and Other Stories*: Causal inference using regression on the treatment variable

### Pre-class warmup assignments

#### 1. Gain scores

Consider the following hypothetical scenario of pre-treatment measurements and potential outcomes:

	$x$	$z$	$y^0$	$y^1$
Aurelia	11	?	13	17
Bora	14	?	14	16
Chen	13	?	17	18
D'Angelo	14	?	19	19
Eddie	9	?	14	13
Farah	11	?	9	10
Goldie	7	?	13	13
Harpreet	10	?	15	19

- (a) Use R to randomly assign treatment to four out of the eight people.
- (b) Given your treatment assignment, calculate the gain score  $g_i = y_i - x_i$  for each person  $i$ .
- (c) Given the data you would see after your treatment assignment, estimate the treatment effect in two ways, first by regressing  $y$  on  $z$ , then by regressing  $y$  on  $z$  and  $x$ .

#### 2. Poststratification

Consider the following fitted model:

	Median	MAD_SD
(Intercept)	23.6	10.9
$z$	10.4	4.0
pre_test	0.7	0.2
$z:pre\_test$	-0.4	0.3

Auxiliary parameter(s):

Median	MAD_SD
sigma	20.1 1.4

- (a) What is the estimated treatment effect?

#### 4.21. CAUSAL INFERENCE USING REGRESSION ON THE TREATMENT VARIABLE

251

- (b) Suppose the estimated population average treatment effect is 7.1. What, then, can you say about the population?

#### Homework assignments

1. (a) Average treatment effects (Exercise 18.4 of *Regression and Other Stories*)

The following table displays a hypothetical experiment on 8 people. Each row of the table gives a participant and her pre-treatment predictor  $x$ , treatment indicator  $z$ , and potential outcomes  $y^0$  and  $y^1$ .

	$x$	$z$	$y^0$	$y^1$
Anna	3	0	5	5
Beth	5	0	8	10
Cari	2	1	5	3
Dora	8	0	12	13
Edna	5	0	4	2
Fala	10	1	8	9
Geri	2	1	4	1
Hana	11	1	9	13

- Give the average treatment effect in the population, the average treatment effect among the treated, and the estimated treatment effect based on a simple comparison of treatment and control.
- Simulate a new completely randomized experiment on these 8 people; that is, resample  $z$  at random with the constraint that equal numbers get the treatment and the control. Report your new randomization and give the corresponding answers for (a).

- (b) Potential outcomes (Exercise 18.5 of *Regression and Other Stories*)

The following tables displays a hypothetical experiment on 2400 people. Each row specifies a category of person, as defined by a pre-treatment predictor  $x$ , treatment indicator  $z$ , and potential outcomes  $y^0, y^1$ . For simplicity, we assume unrealistically that all the people in this experiment fit into these eight categories.

Category	# people in category	$x$	$z$	$y^0$	$y^1$
1	300	0	0	4	6
2	300	1	0	4	6
3	500	0	1	4	6
4	500	1	1	4	6
5	200	0	0	10	12
6	200	1	0	10	12
7	200	0	1	10	12
8	200	1	1	10	12

In making the table, we are assuming omniscience, so that both  $y^0$  and  $y^1$  are known for all observations. But the (non-omniscient) investigator would only observe  $x$ ,  $z$ , and  $y^z$  for each unit. For example, a person in category 1 would have  $x=0, z=0, y=4$ , and a person in category 3 would have  $x=0, z=1, y=6$ .

- What is the average treatment effect in this population of 2400 people?
- Another summary is the mean of  $y$  for those who received the treatment minus the mean of  $y$  for those who did not. What is the relation between this summary and the average treatment effect (ATE)?
- Is it plausible to believe that these data came from a completely randomized experiment? Defend your answer.
- For these data, is it plausible to believe that treatments were assigned using randomized blocks conditional on given  $x$ ? Defend your answer.

2. (a) Gain scores (Exercise 19.3 of *Regression and Other Stories*)

In the discussion of gain-score models in Section 19.3 of *Regression and Other Stories*, we noted that if we include the pre-treatment measure of the outcome in a gain score model, the coefficient on the treatment indicator will be the same as if we had just run a standard regression of the outcome on the treatment indicator and the pre-treatment measure. Show why this is true.

(b) *In pairs:* Working through your own example (Exercise 18.18 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, frame a substantive question in terms of the effect of a binary treatment. For this example, explain what are the outcome variable  $y$ , the treatment variable  $z$ , the pre-treatment variables  $x$ , and the potential outcomes  $y^0$  and  $y^1$ .

## Stories

1. Pest control experiment: estimating a multiplicative treatment effect

Years ago we were working with a group that was estimating the effect of pest control treatments. They were planning to perform an experiment in a number of roach-infested apartments, proceeding as follows:

- (a) In each apartment, a pre-test measurement  $x$  was taken by setting out several traps and counting the number of roaches that are captured.
- (b) The apartments were randomly assigned to a treatment or control group ( $z = 1$  or  $z = 0$ ). In the treatment group, residents were given advice on pest control and, in addition, a crew came into the apartment to clean it, put out poison, and seal all cracks in the walls; this took a full day and cost about \$700. The control group just got the advice.<sup>45</sup>
- (c) A month later, a post-test measurement  $y$ , the number of roaches was taken in each apartment.

Figure 104 summarizes.

The efficacy of the intervention can be estimated by a regression of the post-test measurement on the treatment indicator and the pre-test measurement. A simple linear regression,  $y \sim x + z$ , would be a reasonable start, but it would make more sense to estimate multiplicative effects (for example, “the treatment reduces infestation on average by 40%” rather than “the treatment reduces infestation on average by 15 roaches”), and for that purpose it would make sense to fit a model on the log scale, where we’d first define  $\log_x = \log(x)$  and  $\log_y = \log(y)$  and then fit  $\log_y \sim \log_x + z$ . The trouble here is that some apartments have zero roaches, and you can’t take the logarithm of 0. The data are counts, so it makes sense to fit a negative binomial model with a logarithmic link:  $y \sim \log_x + z$ ,  $\text{family}=\text{neg\_binomial\_2(link="log")}$ , which gives us the best of both worlds: the effect is on the multiplicative scale (just exponentiate the regression coefficient; for example if the coefficient of  $z$  is  $-0.30$ , then the treatment effect is  $\exp(-0.30) = 0.74$ , so the treatment reduces the number of roaches by 26%, on average), and we can model the data as is, including zero cases.

There’s still a problem here for any cases with zero roaches in the pre-treatment measurement, because it’s not possible to take  $\log(x)$  if  $x = 0$ . In practice this might not be a problem because any apartments being included in a roach-control experiment probably will have a high level of infestation at the beginning of the study. Just in case, though, we can work with  $\log(x + 1)$  as our predictor. This is a sloppy solution—really we’d like to use what is called a measurement-error model, where we keep  $\log(x)$  as our regression predictor but consider the observed number of roaches as a noisy measure of  $x$ . But that sort of model goes beyond the scope of this course.

This story relates to the week’s reading by demonstrating a real experiment with pre-test, treatment,

<sup>45</sup>Andrew Gelman (2000), Should we take measurements at an intermediate design point?, *Biostatistics* 1, 27–34.

Potential outcomes, $y$ :
• $y^1$ : number of roaches in your apartment, if $z = 1$
• $y^0$ : number of roaches in your apartment, if $z = 0$
Treatment, $z$ :
• $z = 1$ : cleaning/poison/sealing and pest control advice
• $z = 0$ : pest control advice
Pre-treatment predictor, $x$ :
• Number of roaches measured before treatment

Figure 104 *Outline of an experiment on cockroach infestation. We display to the class to reinforce the sequence of pre-treatment variables, treatment assignment, and potential outcomes.*

and post-test measurements. It is relevant to the course as a whole in combining issues of data collection and measurement with modeling, inference, and decision making.

## 2. Social penumbras and causal effects

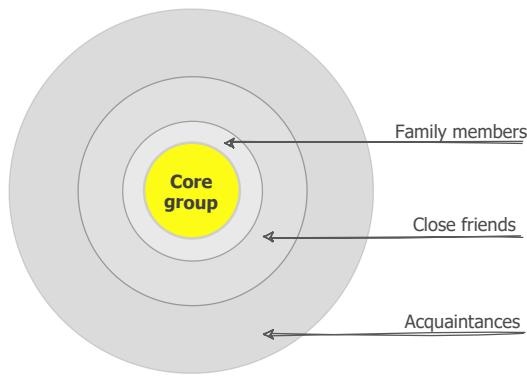
The political influence of a group is typically explained in terms of its size, its geographic concentration, or the wealth and power of the group's members. This example introduces another dimension, the “penumbra,” defined as the set of individuals in the population who are personally familiar with someone in that group. Distinct from the concept of an individual's social network, penumbra refers to the circle of close contacts and acquaintances of a given social group. Figure 105 illustrates.<sup>46</sup>

In 2013 we conducted a survey of American adults, asking them how many people they knew in each of 14 groups, separately asking about close family, close friends, and others, defined as “people that you know their name and would stop and talk to at least for a moment if you ran into the person on the street or in a shopping mall.” Figure 106 shows the size of the different penumbras of each group, along with the size of the group indicated in yellow. The groups are ordered by increasing group size. This figure can be displayed on the screen so that students can discuss it in pairs and note whatever interesting patterns they see. We are now ready for a class discussion. One important feature is that similarly-sized groups can have much differently sized penumbras. Compare, for example, the sizes of the Muslim and gay penumbras. This could have some impact on the political clout of these groups.

Figure 107 shows some statistical summaries of the survey responses for these 14 groups, along with 8 names that we included for comparison: Rose, Emily, Bruce, Walter, Tina, and Kyle, chosen to represent a balance of male, female, young, middle aged, and old, along with Jose and Maria to target the Hispanic population. Students should look in pairs at these graphs and share anything interesting that they notice.

For each respondent and each item, we created a “penumbra score” by counting 1 point for knowing at least one friend with the relevant characteristic, at least one family member, and at least one other person. The penumbra score is thus a number between 1 and 3. For each item, we fit a regression predicting this penumbra score from income, education, ethnicity, and sex. Figure 108 shows coefficients for the regression predictors. These different demographic profiles could be related to groups' political profiles. For example, people with high incomes and high education are less likely to know welfare recipients, uninsured people, and unemployed people. At the opposite end, the penumbras of National Rifle Association (NRA) members, Muslims, and recent immigrants have high shares of upper-income and well-educated people.

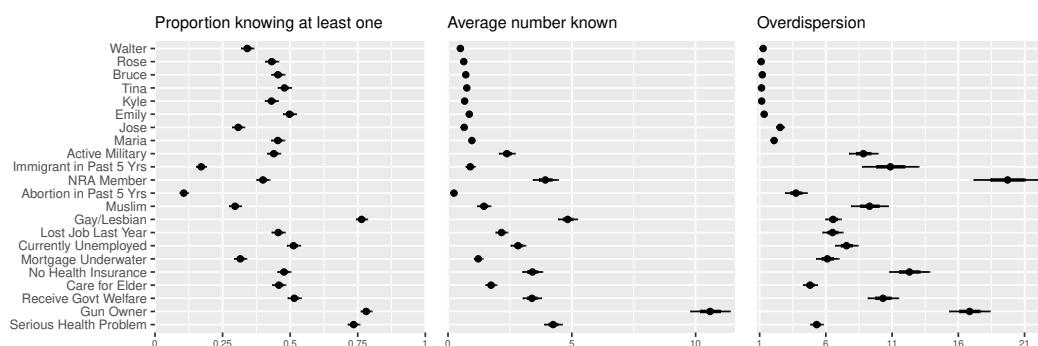
<sup>46</sup>Andrew Gelman and Yotam Margalit (2020), Social penumbras predict political attitudes, *Proceedings of the National Academy of Sciences* 118 (6), e2019375118.



**Figure 105** Sketch of the social penumbra of a group in the population, which we estimate using a series of survey questions on a sample of Americans.



**Figure 106** Core groups and penumbra sizes for the groups asked about in our survey, listed in increasing order of size.



**Figure 107** Some summaries of the penumbras of the 14 groups asked about in our survey, along with 8 names that we included for comparability to earlier research on the topic.

## 4.21. CAUSAL INFERENCE USING REGRESSION ON THE TREATMENT VARIABLE

255

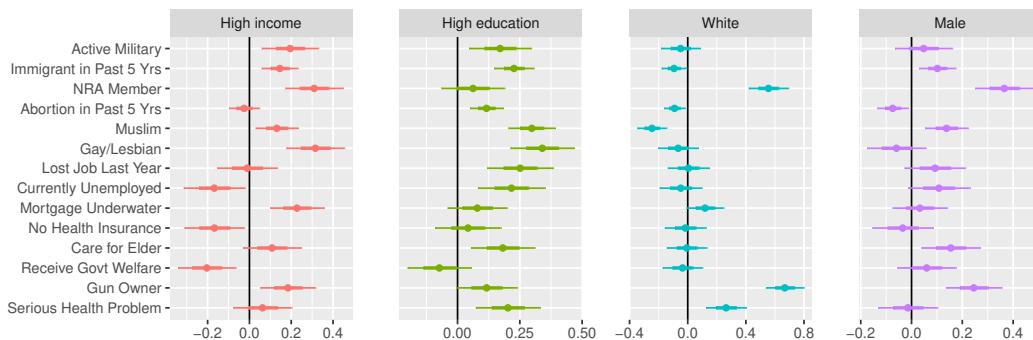


Figure 108 Estimated coefficients  $\pm 1$  and  $2$  standard errors from regressions of penumbra score on indicators for education, income, sex, and ethnicity.

In our survey, we also asked about attitudes on several issues that were related to the groups being studied. There was a question about support of the United States using military force, a question about whether immigration levels should be increased, decreased, or kept about the same, a question about whether Muslims should be given extra security checks at airports, questions about assault weapons, the legality of abortion, extension of unemployment benefits, and so on.

We were interested in seeing how entering a penumbra was predictive of changes in attitudes. So we recontacted the survey respondents from the online panel a year later, again asking the penumbra and issue attitude questions. For each issue question, we then computed the change in attitude—if each question was on a 1–4 scale, the change would be something between  $-3$  and  $3$ . We then fit a logistic regression on the subset of people who were not in the relevant penumbra, predicting the change in attitude on an indicator for whether the respondent entered that penumbra during the second wave, also adjusting for attitude in the first wave, party identification, and demographics. The regression looks like:

```
stan_glm(I(attitude2 - attitude1) ~ in_penumbra2 + attitude1 + partyid +
    age + male + education + black + hispanic, subset=!in_penumbra1)
```

The instructor should type this expression on the board and go over each term with the class.

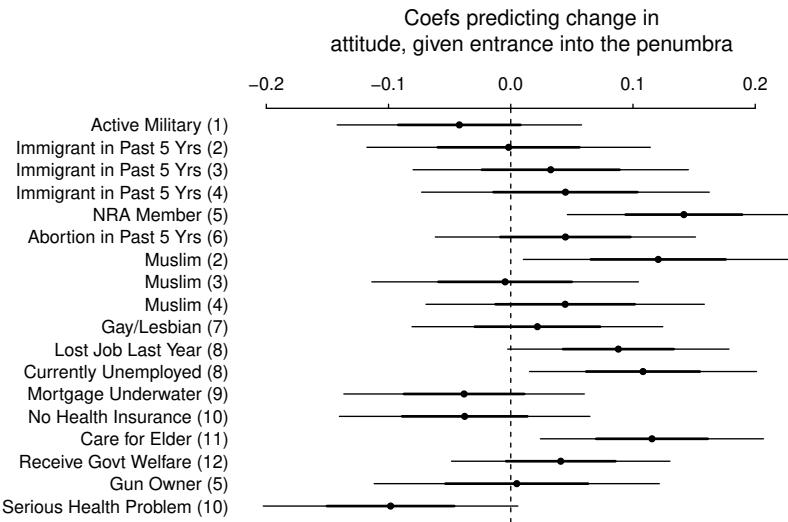
We were interested in the coefficient for `in_penumbra2`, that is, looking at people who were the same at time 1 in issue attitude, partisanship, demographics, and comparing those who were not in the penumbra of the group in question at both times 1 and 2, to those who were not in the penumbra at time 1 and entered it at time 2. Figure 109 shows the results: most of the estimates are positive, indicating that entering the penumbra is generally associated with a change in relevant political attitudes, but there is wide uncertainty.

The high uncertainty makes it challenging to interpret the results. For example, entering the NRA penumbra is associated with a shift toward opposing an assault weapons ban, but there is no clear change associated with knowing a gun owner. Some of this can be understood as simple sampling error, as we discuss next.

The estimates in Figure 109 are on the order of 0.05. Students should discuss in pairs to think about what this means. The survey responses are mostly on 4-point scales, so you can interpret an effect of 0.05 as 5% of the people shifting by 1 point on the scale, or some net shift of 5%, for example 15% shifting in a positive direction and 10% shifting in a negative direction.

Is it plausible that effects would be so small? Maybe so. Consider that many of these are

<sup>47</sup> Justin McCarthy (2001), Record-high 70% in U.S. support same-sex marriage, <https://news.gallup.com/poll/350486/record-high-support-same-sex-marriage.aspx/>.



**Figure 109** For each issue item, estimated coefficient  $\pm 1$  and 2 standard errors of penumbra membership in a logistic regression predicting change in issue attitude between the two waves of the survey, given demographics and attitude in the first wave.

high-profile issues for which most people will not change their opinion in any given year. For example, one of the biggest shifts in opinion in recent decades was the increase in support for same-sex marriage, an increase from 27% in 1996 to 70% in 2021, an average increase of 1.7 percentage points per year.<sup>47</sup> Even in a fast-moving issue with real flux, the number of people who change their views in any single year will be small. And so for most issues we would expect much smaller changes from one year to the next.

In addition, entering a penumbra will not necessarily have a predictable effect. And penumbra measurement is itself noisy, and some of the apparent change in penumbra membership can arise from variation in recall.

Why are the standard errors in the fitted regressions so large? There was only a one-year gap between the two waves of the survey. People enter penumbra all the time, but not so many in any given year. In a survey with 1700 respondents, there might be about 800 people who are not in a particular penumbra at time 1, and perhaps 100 of these enter the penumbra between times 1 and 2. So then our regression is comparing the 100 people in the “treatment group” with 700 “controls,” which is not a lot of data to measure something as noisy as a single issue attitude.

The wide uncertainties in Figure 109 imply a resolution limit of our study. Pointing to these results, we ask students to discuss in pairs how the resolution could be improved. The obvious answer is to increase the sample size, but this is costly: to divide the standard errors in half, you need to multiply the sample size by 4. Other approaches would be to wait longer between waves and to measure issue attitudes more carefully.

This story relates to the week’s reading in that it culminates in a regression model for causal inference. The story relates to the course as a whole with its use of many different analyses to capture different aspects of social science data, in this case using negative binomial regression for the overdispersion model whose results are shown in Figure 107 and linear regression for penumbra scores in Figure 108 and issue attitudes in Figure 109. The story also gives a lesson in the challenges of measurement and the limitations of statistical analysis.

### Class-participation activities

#### 1. Regression to adjust for pre-treatment characteristics

Return to the example of the potential outcomes for basketball training from page 242. In that activity, students filled out a survey form giving their age, a self-assessment of athleticism ( $x$ , on a 1–10 scale), and their guess of the number of shots they would make in 50 tries, with ( $y^1$ ) and without ( $y^0$ ) a month of practice. Students considered various treatment assignment mechanisms and the biases they yielded for the treatment effect as estimated by the average difference between treatment and control observations. This new activity continues the example with an exploration of what happens to the bias when there is adjustment for  $x$ .

For each of several possible assignments of the treatment variable,  $z$ , you can consider three estimates of the average treatment effect: (i) the difference  $\bar{y}^1 - \bar{y}^0$ , (ii) the regression of  $y$  on  $z$  (which should give the same estimate as the difference), and (iii) the regression of  $y$  on  $z$  and  $x$  (here looking at the coefficient on  $z$ ).

Several treatment assignment rules can be considered. For each, the instructor should first lead students in a discussion of the rule and then have them guess the direction and magnitude of the bias under each of the three estimates. During all this time, the data from their survey are projected on the screen, thus providing all the information necessary to compute the biases. Here are some possible treatment assignment rules:

- Independent randomization: flipping a coin (actually, simulating random binary variables with probability 0.5) to randomly assign treatment or control to each participant in the study.
- Complete randomization: half the participants get one treatment and half get the other.
- A biased design: participants choose treatment or control, and more athletic people are more likely to choose to practice basketball, hence a rule such as  $\Pr(z = 1) = \text{logit}^{-1}(x - 5.5)$ .
- A rule that's biased in the other direction: treatment or control is assigned by a coach who gives the practice to those who seem to really need it, hence a rule such as  $\Pr(z = 1) = \text{logit}^{-1}(-(x - 5.5))$ .
- A design based on participants' ages.

The basic idea is that if the design depends on a pre-treatment variable, then that variable should be included as a regression predictor. In particular, you would expect the designs that select based on athleticism to yield very biased estimates in the regression of  $y$  on  $z$  but with much less bias in the regression of  $y$  on  $z$  and  $x$ .

This example relates to the week's reading in demonstrating the value in causal inference of adjusting for pre-treatment predictors. It relates to the course as a whole by showing a connection between data collection and data analysis in the context of causal inference.

#### 2. Understand the “average treatment effect”

This activity explores the idea of the average treatment effect by considering effects on individuals.<sup>48</sup> We first describe the general problem in the context of an example and then ask students to work in pairs to come up with their own example.

Here's the background. In statistics and econometrics there's lots of talk about the average treatment effect. We've often been skeptical of the focus on the average treatment effect, for the simple reason that, if you're talking about an average effect, then you're recognizing the possibility of variation; and if there's important variation (enough so that we're talking about “the average

<sup>48</sup>From Andrew Gelman (2020), Understanding the “average treatment effect” number, <https://statmodeling.stat.columbia.edu/2020/06/30/ate/>. See also Andrew Gelman, Jessica Hullman, and Lauren Kennedy (2023), Causal quartets: Different ways to attain the same average treatment effect, <https://arxiv.org/abs/2302.12878/>.

effect” rather than simply “the effect”), then maybe we care enough about this variation that we should be studying it directly, rather than just trying to reduce-form it away.

But that’s not the whole story. Consider an education intervention. Sure, the treatment effect will vary. But if the treatment will be applied to all the students in a school district, then, yeah, let’s poststratify and estimate an average effect: this seems like a relevant number to know.

What we want to consider here is interpreting that number. It’s something that came up in the discussion of growth mindset.

The reported effect size was 0.1 points of grade point average (GPA). GPA is measured on something like a 1–4 scale, so 0.1 is not so much; indeed, one commenter wrote, “I hope all this fuss is for more than that. Ouch.”

Actually, though, an *average* effect of 0.1 GPA is a lot. One way to think about this is that it’s equivalent to a treatment that raises GPA by 1 point for 10% of people and has no effect on the other 90%. That’s a bit of an oversimplification, but the point is that this sort of intervention might well have little or no effect on most people. In education and other fields, we try lots of things to try to help students, with the understanding that any particular thing we try will not make a difference most of the time. If mindset intervention can make a difference for 10% of students, that’s a big deal. It would be naive to think that it would make a difference for everybody: after all, many students have a growth mindset already and won’t need to be told about it.

That’s all a separate question from the empirical evidence for that 0.1 increase. Our point here is that thinking about an average effect can be misleading.

Or, to put it another way, it’s fine to look at the average, but let’s be clear on the interpretation.

This comes up in a lot of cases. Various interventions are proposed, and once the hype dies down, average effects will be small. There’s no one-quick-trick or even one-small-trick that will raise GPA by 1 point or that will raise incomes by 42% (to use one of our recurring cautionary tales). An intervention that raised average GPA by 0.1 point or that raised average income by 4.2% would still be pretty awesome, if what it’s doing is acting on 10% of the people and having a big benefit on this subset. Researchers and policymakers try different interventions with the idea that maybe one of them will help any particular person.

Again, this discrete formulation oversimplifies—it’s not like the treatment either works or doesn’t work on an individual person. It’s just helpful to understand average effects as compositional in that way. Otherwise you’re bouncing between the two extremes of hypothesizing unrealistically huge effect sizes or else looking at really tiny averages. Maybe in some fields of medicine this is cleaner because you can really isolate the group of patients who will be helped by a particular treatment. But in social policy this is much harder. Even in marketing, efforts to target individuals or small subpopulations are not always successful.

Having discussed all this, we ask students to work in pairs to come up with examples of individual and average treatment effects in areas of interest to them, and then we talk about one or two of these together as a class. This activity relates to the week’s reading by bringing specificity to the theoretical concept of average treatment effect, and it is relevant to the course as a whole in connecting mathematical models to numbers from specific applications.

## Computer demonstrations

### 1. Benefits of pre-test measurements

Consider a hypothetical study of interest to the class, for example a political persuasion experiment that will shift opinions on average by 5 percentage points. To do the simulation you need to create a fake world. For this particular example, assume that, at baseline, 40% of the population supports

## 4.21. CAUSAL INFERENCE USING REGRESSION ON THE TREATMENT VARIABLE

259

some particular issue and that, in the absence of the treatment, people's opinions will not change. We further assume that the treatment will shift 5% of the people from No to Yes.

Consider an experiment in which 1000 people get the treatment and 1000 are in the control group.

```
n <- 1000
se <- sqrt(0.5^2/n + 0.5^2/n)
print(se)
```

The standard error of this experiment is 0.022, or 2.2 percentage points. Let's simulate some fake data and do the analysis:

```
y_0 <- rbinom(1, n, 0.40)
y_1 <- rbinom(1, n, 0.45)
diff <- y_1/n - y_0/n
print(c(diff, se))
```

As discussed in Section 7.3 of *Regression and Other Stories*, this comparison can be expressed as a regression. First simulate the 2000 responses:

```
z <- rep(c(0,1), c(n,n))
y <- rbinom(2*n, 1, 0.40 + 0.05*z)
fake <- data.frame(y, z)
```

Print this data frame in the console to check that it makes sense. Then fit the regression:

```
fit <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit, digits=3)
```

The errors in this regression are not normally distributed but we don't really care for the purpose of estimating the average effect.

That is all fine, but you can do better if opinions are first measured *before* the experiment and then used as a pre-treatment predictor. Consider the following hypothetical population:

Share of population	Pre-test response	Post-test response if control	Post-test response if treated
0.40	Yes	Yes	Yes
0.55	No	No	No
0.05	No	No	Yes

The last row in this table represents the persuadables. From a statistical perspective, this is a best-case scenario, but let's go with it. Now simulate some data. Let  $x$  be the pre-treatment opinion:

```
x <- rbinom(2*n, 1, 0.40)
z <- rep(c(0,1), c(n,n))
y <- ifelse(z==0, x, ifelse(x==1, 1, rbinom(n, 1, 0.05/0.60)))
fake <- data.frame(x, y, z)
```

This was kind of yucky, but it did the job. And now it's time to fit the regression:

```
fit_1 <- stan_glm(y ~ z, data=fake, refresh=0)
print(fit_1, digits=3)
fit_2 <- stan_glm(y ~ x + z, data=fake, refresh=0)
print(fit_2, digits=3)
```

The standard error is now much lower.

This is the best case, where the pre-test is a nearly perfect measurement, but still. The instructor can discuss how this works and then ask the students why this sort of pre-test measurement is not always done in a comparative study.

This demonstration relates to the week's reading in the connection between data and the underlying object of study. It relates to the course as a whole in linking data collection to analysis.

## 2. Combining pre-treatment predictors

The following code shows a simple example of combining several pre-treatment predictors into a single score. The first step is to fit a linear model with no interactions. We next use the coefficients from the fitted model to create the combined score. Finally, we include the combined score along with the treatment indicator in a regression with interactions. The advantage of this approach is that the final model is relatively simple, as long as we are comfortable interpreting the linear combination of the pre-treatment predictors as a summary score.

We demonstrate with simulated data:

```
n <- 100
z <- rnorm(n, 0, 1)
x2 <- rnorm(n, 0, 1)
x3 <- rnorm(n, 0, 1)
x4 <- rnorm(n, 0, 1)
x5 <- rnorm(n, 0, 1)
b0 <- 0; b1 <- 1; b2 <- 2; b3 <- 3; b4 <- 4; b5 <- 5
sigma <- 10
y <- b0 + b1*z + b2*x2 + b3*x3 + b4*x4 + b5*x5 + rnorm(n, 0, sigma)
fake <- data.frame(z, x2, x3, x4, x5, y)
fit <- stan_glm(y ~ z + x2 + x3 + x4 + x5, data=fake, refresh=0)
print(fit)
b_hat <- coef(fit)

fake$u <- b_hat["x2"]*x2 + b_hat["x3"]*x3 + b_hat["x4"]*x4 + b_hat["x5"]*x5
fit_2 <- stan_glm(y ~ z + u, data=fake, refresh=0)
print(fit_2)
fit_3 <- stan_glm(y ~ z + u + z:u, data=fake, refresh=0)
print(fit_3)
```

## Drills

### 1. Average treatment effect and poststratification

For each of these models, give R code to compute the sample average treatment effect, ignoring any uncertainty in the coefficient estimates. In all examples,  $z$  is a binary variable and you want to compare  $z = 0$  to  $z = 1$ . Further suppose that the data for the regression are in a data frame, `expt`.

(a)  $y = 1 + 2x + 3z + 4xz$

*Solution:* `mean(3 + 4*expt$x)`

(b)  $y = 1 + 2x + 3z$

(c)  $y = 1 + 2x_1 + 3x_2 + 4x_1x_2 + 5z$

(d)  $y = 1 + 2x_1 + 3x_2 + 4x_1x_2 + 5z + 6x_1x_2z$

### 2. Average treatment effect for nonlinear models

For each of these models, give R code to compute the sample average treatment effect, ignoring any uncertainty in the coefficient estimates. In all examples,  $z$  is a binary variable and you want to compare  $z = 0$  to  $z = 1$ . Further suppose that the data for the regression are in a data frame, `expt`.

## 4.21. CAUSAL INFERENCE USING REGRESSION ON THE TREATMENT VARIABLE

261

(a)  $\Pr(y = 1) = \text{logit}^{-1}(1 + 2x + 3z)$

*Solution:* `mean(invlogit(1 + 2*expt$x + 3) - invlogit(1 + 2*expt$x))`

(b)  $\Pr(y = 1) = \text{logit}^{-1}(1 + 2x + 3z + 4xz)$

(c)  $\Pr(y = 1) = \text{logit}^{-1}(1 + 2x_1 + 3x_2 + 4x_1x_2 + 5z)$

(d)  $\log y = 1 + 2x + 3z + 4xz$

### Discussion problems

#### 1. Causal inference using logistic regression for a binary outcome

We have focused on using linear models to estimate causal effects, starting with the additive model,  $y = \beta_0 + \beta_1 x + \beta_2 z + \text{error}$ , and continuing to the interaction model,  $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \text{error}$ , with pre-treatment predictor  $x$ , treatment indicator  $z$ , and outcome measurement  $y$ . But what if your outcome is binary (live or die, vote or don't vote, pass or fail, etc.)? Then it would seem natural to model  $y$  given  $x$  and  $z$  using logistic regression. How would it work to estimate the treatment effect in this way? What is the estimated causal effect? Be careful with notation.

#### 2. Holding all else equal?

Consider the following statement from a well-known psychology researcher:<sup>49</sup>

“Education is an important determinant of income—one of the most important—but it is less important than most people think. If everyone had the same education, the inequality of income would be reduced by less than 10%. When you focus on education you neglect the myriad other factors that determine income. The differences of income among people who have the same education are huge.”

We think we know what he’s saying—if you regress income on education and other factors, and then you take education out of the model,  $R^2$  decreases by 10%. Or something like that. Not necessarily  $R^2$ , maybe you fit the big model, then get predictions for everyone putting in the mean value for education and look at the standard deviation of incomes or the Gini index or whatever. Or something else along those lines.

But we have problems with the counterfactual: “If everyone had the same education . . .” First, if everyone had the same education, the world would be much different and we don’t see why the regressions on which Kahneman is relying would still be valid. Second, is it even possible for everyone to have the same education? Different students have different interests and their educational trajectories will differ. You could imagine everyone having the same number of years of education, but that seems like a different thing entirely.

As noted, we get the point that income is determined by lots of other factors than education, but we think the quoted statement is too casual with the causality. And, without the causal punch, the statement doesn’t seem so impressive. Take this as a warning about the pitfalls that can arise from processing undigested regression results.

For discussion, consider this example in two different ways. First, suppose you are interested in what would happen to incomes and income inequality if everyone had the same education. How could you model or estimate this? Second, consider the descriptive research alluded to in the quotation. How could you describe these findings in a way that conveys why they are interesting but without using inappropriate causal language?

<sup>49</sup>Daniel Kahneman (2011), Focusing illusion, <https://www.edge.org/response-detail/11984/>. The discussion here is taken from Andrew Gelman (2011), D. Kahneman serves up a wacky counterfactual, [https://statmodeling.stat.columbia.edu/2011/05/13/d\\_kahneman\\_serv/](https://statmodeling.stat.columbia.edu/2011/05/13/d_kahneman_serv/).

## 4.22 Causal inference as prediction

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
No effect of heart stents?	Components of an observational study	Playing with least squares	Experimental design	Individual and average effects
The freshman fallacy	Study makers vs. study breakers	Don't adjust for intermediate outcomes	Adjusting for post-treatment variables	Nudge meta-analysis

### Reading

Chapters 18 and 19 of *Regression and Other Stories* again

### Pre-class warmup assignments

#### 1. Simulate potential outcomes

Consider a randomized experiment with pre-treatment variable  $x$ , treatment indicator  $z$ , and potential outcomes  $y^0$  and  $y^1$ . Assume that, under the control condition, outcome follows a linear model,  $y = 0.1 + 0.2x + \text{error}$ . Further assume that the expected treatment effect is  $0.3 + 0.4x$ . Simulate a randomized experiment on 100 people in the following steps:

- Simulate the pre-treatment variables for the 100 people whose  $x$  values are uniformly distributed between 0 and 10.
- Simulate the treatment assignment, with 50 people randomly selected to get the treatment and 50 getting the control.
- Simulate both potential outcomes for all 100 people, assuming independent errors that are normally distributed with mean 0 and standard deviation 0.3.
- Compute the observed  $y$  for each person by picking the appropriate potential outcome given the treatment assignment.

#### 2. Regression modeling for average causal effects

Suppose you have pre-treatment variables  $x_1, x_2$ , treatment indicator  $z$ , and outcome  $y$  on a collection of people in an observational study. Give R code to fit regression models for the following tasks:

- Estimating a constant treatment effect, adjusting for  $x_1$  and  $x_2$
- Estimating a constant treatment effect, adjusting for  $x_1, x_2$ , and their interaction.
- Estimating a treatment effect that linearly depends on  $x_1$  and  $x_2$ , adjusting for  $x_1, x_2$ , and their interaction
- If  $y$  is binary, estimating a constant treatment effect on the logit scale, adjusting for  $x_1, x_2$ , and their interaction

### Homework assignments

#### 1. (a) Pre-test and post-test (Exercise 19.4 of *Regression and Other Stories*)

100 students are given a pre-test, then a treatment or control is randomly assigned to each, then they get a post-test. You are given the following regression model:

$$\text{post\_test} = a + b * \text{pre\_test} + \theta * z + \text{error},$$

where  $z = 1$  for treated units and 0 for controls. Further suppose that `pre_test` has mean 40 and standard deviation 15. Suppose  $b = 0.7$  and  $\theta = 10$  and the mean for `post_test` is 50 for the students in the control group. Further suppose that the residual standard deviation of the regression is 10.

- i. Determine  $a$ .
  - ii. What is the standard deviation of the post-test scores for the students in the control group?
  - iii. What are the mean and standard deviation of the post-test scores in the treatment group?
- (b) Causal inference using logistic regression (Exercise 19.5 of *Regression and Other Stories*)

Suppose you have fit a model,

```
fit <- stan_glm(y ~ z + age + z:age,  
family=binomial(link="logit"), data=mydata)
```

with binary outcome  $y$ , treatment indicator  $z$ , and  $age$  measured in years. Give R code to produce an estimate and standard error of average treatment effect in a large population, given a vector `n_pop` of length 82 that has the number of people in the population at each age from 18 through 99.

2. (a) Sketching the regression model for causal inference (Exercise 19.6 of *Regression and Other Stories*)

Assume that linear regression is appropriate for the regression of an outcome,  $y$ , on treatment indicator,  $z$ , and a single confounding covariate,  $x$ . With pen on paper, sketch hypothetical data (plotting  $y$  vs.  $x$ , with treated and control units indicated by circles and dots, respectively) and regression lines (for treatment and control group) that represent each of the following situations:

- i. No treatment effect,
  - ii. Constant treatment effect,
  - iii. Treatment effect increasing with  $x$ .
- (b) Linearity assumptions and causal inference (Exercise 19.7 of *Regression and Other Stories*)
- Consider a study with an outcome,  $y$ , a treatment indicator,  $z$ , and a single pre-treatment predictor,  $x$ . Draw a scatterplot of treatment and control observations that demonstrates each of the following:
- i. A scenario in which the difference in means estimate would not capture the true treatment effect but a regression of  $y$  on  $x$  and  $z$  would yield the correct estimate.
  - ii. A scenario in which a linear regression would yield the wrong estimate but a nonlinear regression would yield the correct estimate.

## Stories

1. Apparent null effects in a study of heart stents

This example is discussed at the end of Section 3.5 of *Regression and Other Stories* and in more detail in a research article in which we wrote,<sup>50</sup>

“The study [of a heart treatment called ‘stents’] included approximately 200 patients and was notable for being a blinded experiment in which half the patients received stents and half received a placebo procedure in which a sham operation was performed. In followup, patients were asked to guess their treatment and of those who were willing to guess only 56% guessed correctly, indicating that the blinding was largely successful.

The summary finding from the study was that stenting did not ‘increase exercise time

<sup>50</sup>Andrew Gelman, John Carlin, and Brahmajee Nallamothu (2019), Objective Randomised Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina (ORBITA) and coronary stents: A case study in the analysis and reporting of clinical trials, *American Heart Journal* 214, 54–59.

by more than the effect of a placebo procedure' with the mean difference in this primary outcome between treatment and control groups reported as 16.6 seconds with a standard error of 12.7 . . .”

The average gain score (increase in exercise times from the beginning to end of the study) was 28.4 seconds for the treatment group and 11.8 seconds for the control group, yielding a difference in gain scores of 16.6 seconds. Compared to the standard error of 12.7 seconds (computed using the usual formula,  $\sqrt{\sigma_C^2/n_C + \sigma_T^2/n_T}$ , where  $\sigma_C$  and  $\sigma_T$  are the standard deviations of the gain scores in the control and treatment groups, respectively, and  $n_C$  and  $n_T$  are the sample sizes), this difference is not “statistically significant”: the estimate is less than 2 standard errors from zero.

Following usual practice, the non-statistically-significant result was characterized as a null finding. For example, here is how the study was reported in the *New York Times*:<sup>51</sup>

“Heart Stents Fail to Ease Chest Pain . . . When the researchers tested the patients six weeks later, both groups said they had less chest pain, and they did better than before on treadmill tests. But there was no real difference between the patients, the researchers found. Those who got the sham procedure did just as well as those who got stents. . . . ‘It was impressive how negative it was,’ Dr. Redberg said of the new study . . .”

However, the estimate using gain in exercise time died not make full use of the data that were available on differences between the comparison groups at baseline. In particular, the treatment and placebo groups differed in their pre-treatment levels of exercise time, with the treatment group having values that were 38.0 seconds higher, on average. This sort of difference is no surprise—randomization assures balance only in expectation—but it is important to adjust for this discrepancy in estimating the treatment effect.

If we label  $y$  as the post-test outcome and  $x$  as the pre-test measure, then the estimated effect from the gain score is,

$$\text{gain score estimate} = (\bar{y}_T - \bar{x}_T) - (\bar{y}_C - \bar{x}_C).$$

But this overcorrects for differences in pre-test scores, because of the familiar phenomenon of “regression to the mean” (see Chapter 6 of *Regression and Other Stories*): just from natural variation, one would expect patients with lower scores at baseline to improve, relative to the average, and patients with higher scores to regress downward.

The optimal linear estimate of the treatment effect is,

$$\text{adjusted estimate} = (\bar{y}_T - \beta \bar{x}_T) - (\bar{y}_C - \beta \bar{x}_C),$$

where  $\beta$  is the coefficient of  $x$  in a regression of  $y$  on  $x$  and the treatment indicator  $z$ . The gain score estimate is a special case of the regression estimate corresponding to  $\beta = 1$ . Given that the pre-test and post-test measurements are on the same scale and we would expect them to be positively correlated and have nearly identical variances, we can anticipate that the estimated  $\beta$  will be less than 1, which will reduce the correction for difference in pre-test and thus increase the estimated treatment effect while also decreasing the standard error. As a result, an adjusted analysis of these data would be expected to yield a stronger estimated effect.

Indeed, this is what happened. The estimate of  $\beta$  from these data is 0.87, yielding an adjusted mean difference of 21.3 (quite a bit higher than the raw difference in gain scores of 16.6) with a standard error of 12.5 (very slightly lower than 12.7, the standard error of the difference in gain scores). The estimate is not quite 2 standard errors away from zero: the ratio of estimate divided by standard error is 1.7. The  $p$ -value from this adjusted analysis is 0.09; as anticipated, it is lower than the  $p = 0.20$  from the unadjusted analysis.

<sup>51</sup>Gina Kolata (2017), “Unbelievable”: Heart stents fail to ease chest pain, *New York Times*, 2 Nov, <https://www.nytimes.com/2017/11/02/health/heart-disease-stents.html>.

Despite moving closer to the conventional 0.05 threshold, the  $p$ -value of 0.09 remains above the traditional level of significance at which one is taught to reject the null hypothesis. A potential blockbuster reversal with an adjusted analysis—“Statistical Sleuths Turn Reported Null Finding into a Statistically Significant Effect”—does not quite materialize.

Yet within different conventions for scientific reporting, this experiment could have been presented as positive evidence in favor of stents. In some settings, a  $p$ -value of 0.09 is considered to be statistically significant.

Here is the main reason we’re telling the story of the stents experiment. It is a well-known statistical fallacy to take a result that is not statistically significant and report it as zero, as was essentially done here based on the  $p$ -value of 0.20 for the primary outcome. Had this comparison happened to produce a  $p$ -value of 0.04, would the headline have been, “Confirmed: Heart Stents Indeed Ease Chest Pain”? A lot of certainty seems to be hanging on a small bit of data.

Beyond exercise time, other signals from ORBITA seemed to suggest consistent improvements in the physiological parameter of ischemia through endpoints such as fractional flow reserve, instantaneous wave-free ratio, and stress echo. Actually, findings from the stress echo highlight a potentially important avenue into an alternative presentation of these results. There is no question that some physiological changes are being made by stents, with large and stable effects seen on echo measures. As is often the case, the null hypothesis that these physical changes should make absolutely zero difference to any downstream clinical outcomes seems farfetched. A simple way to tackle this question is to report uncertainty intervals around the mean differences and not to focus on whether the intervals happen to include zero.

The larger question has to be about balancing the long-term benefits of stents with risks of the operation, and in making such decisions it is important to move beyond a simplistic binary summary saying that the treatment works or does not work. In reality, treatment effects are variable and there is unavoidable uncertainty about the magnitude and direction of the effect. Statistics cannot always resolve this uncertainty, and in particular it is a mistake to conclude that an effect is negligible or zero just because it does not reach some threshold of “statistical significance.”

This example relates to the week’s reading as an example of the application of causal inference, and it is relevant to the course as a whole in demonstrating the challenge of moving from a statistical analysis to larger conclusions and decisions.

## 2. The Freshman Fallacy: Interactions and average treatment effects

In a discussion of an article, “Women are more likely to wear red or pink at peak fertility,” published in the journal *Psychological Science* in 2013, which was based on two samples—a self-selected sample of 100 women from the Internet, and 24 undergraduates at the University of British Columbia—we wrote,<sup>52</sup>

“[There is a problem with] representativeness. What color clothing you wear has a lot to do with where you live and who you hang out with. Participants in an Internet survey and University of British Columbia students aren’t particularly representative of much more than . . . participants in an Internet survey and University of British Columbia students.”

In response, we received this in an email from a prominent psychology researcher whom we had never met:

“Complaining that subjects in an experiment were not randomly sampled is what freshmen do before they take their first psychology class. I really \*hope\* you [know] why that is an absurd criticism—especially of authors who never claimed that their study generalized to

<sup>52</sup>Andrew Gelman (2013), Too good to be true, *Slate*, 24 Jul, <https://slate.com/technology/2013/07/statistics-and-psychology-multiple-comparisons-give-spurious-results.html>.

1. Essentially no effect, with patterns in data coming from noise or measurement artifacts
2. Large and variable effects that depend strongly on the person and context
3. Large and consistent effects

**Figure 110** Three scenarios for the treatment effect of the ovulation-and-clothing study. If the effect is estimated with no interactions and then used to draw conclusions about the general population, this is equivalent to assuming scenario 3. The instructor can project this onto the screen to help focus class discussion.

all humans. (And please spare me ‘but they said men and didn’t say THESE men’ because you said there were problems in social psychology and didn’t mention that you had failed to randomly sample the field. Everyone who understands English understands their claims are about their data and that your claims are about the parts of psychology you happen to know about.”)

Just because a freshman might raise a question, that does not make the issue irrelevant! Freshmen can be pretty thoughtful sometimes. And we hope they remain skeptical of these studies even after they take their first psychology class. Like these freshmen, we are skeptical about generalizing to the general population based on 100 people from the internet and 24 undergraduates.

We have no doubt that the authors, and anyone else who found this study to be worth noting, are interested in some generalization to a larger population—certainly not “all humans” (as claimed by our correspondent), but some large subset of women of childbearing age, some subset that includes college students in Canada and women of various ages who are on the Mechanical Turk experimental platform. The abstract to the paper simply refers to “women” with no qualifications.

Why should generalization be a problem? The issue is subtle. We will elaborate on the representativeness issue using some (soft) mathematics.<sup>53</sup>

Let  $\theta$  be the parameter of interest, in this case the difference in the probability of wearing red or pink shirts, comparing women in two different parts of their menstrual cycle, among the women who are wearing shirts and have regular menstrual periods.

The concern is that, to the extent that  $\theta$  is not very close to zero, that it can vary by person and by context. For example, perhaps  $\theta$  is a different sign for college students who don’t want to get pregnant, as compared to married women who are trying to have kids. Perhaps  $\theta$  is much different for single women than women with partners.

Consider three possible scenarios:

- (1) *Essentially no effect.* Women’s clothing colors have very low correlations with the time of the month, and anything that you find in data will likely come from sampling variability or measurement artifacts.
- (2) *Large and variable effects.* Results will depend strongly on what population is studied and on the social context. There is no reason to trust generalizations from an unrepresentative sample. The college freshmen are right.
- (3) *Large and consistent effects.* If the parameter  $\theta$  is large and pretty much the same sign everywhere, then a sample of college students or internet participants is just fine (measurement issues aside).

These are summarized in Figure 110, which the instructor can project onto the screen as reference for the students. Our correspondent, and the authors of the published article in question, implicitly

<sup>53</sup> Andrew Gelman (2013), Does it matter that a sample is unrepresentative? It depends on the size of the treatment interactions, <https://statmodeling.stat.columbia.edu/2013/09/04/does-it-matter/>.

assumed the third scenario. Until you make that assumption, you can't really generalize beyond people who are like the ones in the study.

In this particular case, we expect the authors gained confidence in their results because they appeared in two very different populations. They saw a large estimate of  $\theta$  among the group of internet participants and a large estimate among the college students, hence this is some evidence that  $\theta$  is large in general. This is a good idea in general—two case studies is a good way to get started in looking at variation—but in this particular case we don't trust it because the sample sizes are so small and the data analysis rules were flexible enough to allow researchers to find statistical significance even from noise: the “forking paths” problem discussed in Section 4.5 of *Regression and Other Stories*.

Representativeness of samples is something that statisticians and economists have thought a lot about, and which we discuss in the causal inference chapters in *Regression and Other Stories*. Phrases such as “local average treatment effect” recognize that treatment effects vary, and this leads to interest in looking at where an intervention is more or less effective.<sup>54</sup>

Researchers in medicine and public health are also acutely aware of variation in treatment effects and the need to consider what population is being studied when an effect is being estimated. In medicine and public health (unlike in psychology), it tends to be expensive to add people to a study. Researchers want to maximize their power for a given cost, and so they often make an effort to restrict enrollment to the subset of people who they believe are most likely to respond to the treatment. This results in, among other things, the “decline effect” when a successful experimental treatment is applied to the general population.

But, as indicated by the quotation from the psychology professor above, not all researchers are aware of the potential for treatment interactions; they seem to be implicitly operating under scenario 3 in Figure 110, in which effects are universal, or at least where there is no reason to be concerned about extrapolating from 24 college students to “women” in general. The point of this story is to explain why such a generalization can be a mistake. This is a case in which professional expertise can be a bad thing; the intuition of a college freshman can be more valid than the experience of an experienced and much-published researcher.

We hope that the next time a freshman comes to our correspondent with a complaint about subjects in an experiment not being randomly sampled, he (our correspondent) will not merely dismiss the complaint but instead discuss with the student its relevance under scenarios 1, 2, and 3.

The next year, the authors of the ovulation-and-clothing paper published a new article, “The impact of weather on women’s tendency to wear red or pink when at high risk for conception.” The trouble is that the weather factor came up in this new article but nowhere else. There’s an essentially unlimited list of factors that could be picked up to explain any pattern in the dress of college students: if not weather, there’s family structure such as number of younger and older siblings, relationship status of the students, age, political orientation, socioeconomic status, height, weight, general health, sexual history, and so on—all of which are plausibly related to the topic of the study, which is described as “a desire to increase one’s sexual appeal.” This long list of factors represents a large number of forking paths and, in addition, many potential interactions that would make it difficult to generalize from a nonrepresentative sample.

At this point the question arises, how could such a study ever be done in a reasonable way? It might be that the project is hopeless, as it’s plausible that there are no clear differences in patterns of dress between different parts of the menstrual cycle—that is, scenario 1 in Figure 110. If there’s no strong effect to find, then it’s game over.

But if there are real patterns—scenario 2 or 3—we would recommend a within-person design,

<sup>54</sup>See, for example, Rajeev Dehejia (2003), Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data, *Journal of Business and Economic Statistics* 21, 1–11.

interviewing people repeatedly over a series of months, measuring ovulation as accurately as possible, and looking at other outcomes too, not just what shirt the student is wearing. One of the difficulties is that the hypothesis is so vague. Maybe it would make sense to formally have a two-stage design where the first stage is exploratory and the researchers use the analyses from these data to formulate a specific hypothesis that they could test in the second stage. Think seriously about effect size and do a corresponding design analysis. It's not easy, but if you think the topic is important, that's the way to study it. It's a mistake to try to do science by taking sloppy measurements and then trying to come up with hypotheses from the data.

This story relates to the week's reading because it involves the importance of varying treatment effects and interactions. It relates to the course as a whole because it demonstrates the challenge of certain conventional approaches to statistical analysis of experimental data.

### Class-participation activities

#### 1. Components of an observational study

This activity is done with students in groups of four, with two students coming up with a setting for an observational study and the other students asking questions to clarify. For each study, there needs to be a population of units, a sample, a pre-treatment measurement  $x$ , a treatment  $z$  and its assignment rule, and an outcome  $y$ ; see Figure 111. All must be clearly defined, and  $z$  must be a treatment, that is, something that could possibly be done.

Here is a simple example, a hypothetical study of the effectiveness of a new plan for teaching statistics:

- Population: all students who might study statistics in high school
- Sample: students in statistics classes in the 18 schools in a city where the study is conducted
- Pre-treatment measurement: a mathematics pre-test given to all students
- Treatment or exposure: for a year, some students get the traditional teaching method, some get the new method
- Treatment assignment: teachers choose the teaching methods for their own classes. (For this activity we want to define an observational study, not an experiment.)
- Outcome: a standardized test taken by all students at the end of the year.

The instructor can go through this example to lead off the activity, to give a sense of how the steps in Figure 111 could go. Analysis of this example study would involve challenges including clustering (data collection on students but treatment assigned at the classroom level), omitted variables (perhaps the teachers who are willing to try a new method are already better at their jobs), and generalization from sample to population. That's fine; the focus here is on carefully laying out all the components of an observational study.

Or consider a study of the effects of smoking on lung function. The population could be all adults in the United States, the sample could be some large number of people in a survey, the pre-treatment measurement could be parents' lifespans, the treatment could be the lifetime number of cigarettes smoked or the equivalent, as measured with a set of survey questions, exposure is self-selected, and the outcome could be some measure of lung function. Challenges would include omitted variables, measurement error, and missing data, but, again, there is a pretty direct mapping from applied example to the observational-study framework.

For a more challenging example, consider a hypothetical study of the effects of income on happiness. There will be some difficulties of measuring income and happiness, but the real challenge we want to emphasize here is that "income" is not a treatment or exposure, in that it is not something that could be done. To talk about "the effect" of income, it is not enough to simply

1. Population
2. Sample
3. Pre-treatment measurement,  $x$
4. Treatment or exposure,  $z$
5. Treatment assignment rule
6. Outcome,  $y$

Figure 111 *Steps needed to define an observational study. This should be projected onto the screen during the class-participation activity. The challenge for each group of students is to get specific on all these steps.*

compare people who have different incomes. It is necessary to define some specific treatment, for example getting a new job, or a raise, or a tax break, or some other source of money. For another example, suppose you want to conduct a cross-country study of the effect of levels of social trust (as measured from the responses on a battery of survey questions) on the probability of civil war. Again, “social trust” is not in itself a treatment, so you would need to consider some potential treatments or exposures that would alter social trust, for example outside support for an anti-immigrant political party.

Once the instructor has displayed Figure 111 and discussed some of these challenges, the students should divide into groups of four, with one pair coming up with the example and the other pair going through the six steps and asking questions to clarify the details of the study. When students are done with this, one or two of these examples could be used for class discussion if there is time.

## 2. Role play: study makers vs. study breakers

This activity is done with students in groups of four. Two of the students will be “study makers” and should come up with an idea for an observational study, including plans for gathering data and estimating the treatment effect of interest. The other pair of students in the group should play the role of “study breakers,” raising concerns about the proposed idea. The pairs should go back and forth, discussing potential problems with the study and how they could be fixed through analysis or gathering additional data. Meanwhile, the instructor should go around the room observing the groups and helping to focus the discussions where necessary.

After the students have had a few minutes to work through their example, one of the groups should go up to the board and re-enact their discussion, and the rest of the class can participate in the conversation regarding this particular proposed observational study. The instructor can help guide the discussion by pointing out connections to the topics in the week’s readings and in the course as a whole. The goal here is not for the “study makers” or the “study breakers” to have a debate with a winner and a loser, but rather to use role play to explore in depth the challenges of designing and interpreting an observational study.

## Computer demonstrations

### 1. Playing with least squares optimization

Here you can explore the ideas of least squares regression, which is an important tool in causal inference. Following Section 8.1 of *Regression and Other Stories*, define a `rss` (residual sum of squares) function, simulate fake data, and fit a linear regression:

```
library("rstanarm")
rss <- function(x, y, a, b){ # x and y are vectors, a and b are scalars
  resid <- y - (a + b*x)
  return(sum(resid^2))
}
```

```
# Generate fake data
n <- 30
x <- runif(n, 0, 10)
y <- 5 + 3*x + rnorm(n, 0, 10)
fake <- data.frame(x, y)

# Prepare a plotting grid
par(mfrow=c(2,2))

# Plot the data
par(mar=c(3,3,2,1), mgp=c(1.5,.5,0), tck=-.01)
plot(fake$x, fake$y, pch=20, bty="l")

# Fit the model
fit <- stan_glm(y ~ x, data=fake, refresh=0)
abline(coef(fit), col="blue")
print(fit)
a_hat <- coef(fit)[1]
b_hat <- coef(fit)[2]
fit_rss <- rss(fake$x, fake$y, a_hat, b_hat)
print(fit_rss)
mtext(paste("y =", round(a_hat, 1), "+", round(b_hat, 1), "x; rss =", round(fit_rss, 1)), side=3, line=1, col="blue")
```

Then play around with other lines and compute `rss` for each:

```
plot(data$x, data$y, pch=20, bty="l")
abline(coef(fit), col="blue")
a_guess <- 5
b_guess <- 3
abline(a_guess, b_guess, col="red")
guess_rss <- rss(fake$x, fake$y, a_guess, b_guess)
mtext(paste("y =", round(a_guess, 1), "+", round(b_guess, 1), "x; rss =", round(guess_rss, 1)), side=3, line=1, col="red")
```

You can make this into a function and then try it out:

```
rss_plot <- function(fake, a_guess, b_guess) {
  plot(fake$x, fake$y, pch=20, bty="l")
  fit <- stan_glm(y ~ x, data=fake, refresh=0)
  abline(coef(fit), col="blue")
  abline(a_guess, b_guess, col="red")
  guess_rss <- rss(fake$x, fake$y, a_guess, b_guess)
  mtext(paste("y =", round(a_guess, 1), "+", round(b_guess, 1), "x; rss =", round(guess_rss, 1)), side=3, line=1, col="red")
}
```

```
rss_plot(fake, 5, 3)
```

Now you can try some more values and see that if you move the line around, `rss` will always go up.

This demonstration relates to the week's reading in giving a sense of the modularity of inferences. You can see this here with simple least squares, and the same principle applies with more complicated estimates such as the two-stage least squares used for instrumental variables. The demonstration is relevant more generally as an example of checking a mathematical formula by writing a function in R.

2. Don't adjust for intermediate outcomes

Section 19.6 of *Regression and Other Stories* discusses the problem with adjusting for post-treatment variables when using regression for causal inference. Here we demonstrate with a simulation of a hypothetical education experiment, constructed as follows:

- (a) Simulate a distribution of students' background abilities on the subject being studied. This will be an unobserved (latent) variable.
- (b) Simulate a pre-test score as ability plus random error.
- (c) Randomly assign each student to control ( $z = 0$ ) or treatment ( $z = 1$ ) with equal probabilities.
- (d) Simulate a midterm test score given ability and assuming the treatment has a positive effect of 5 points on the midterm.
- (e) Simulate a final exam score given ability and assuming the treatment has a positive effect of 10 points on the final.

The goal of this hypothetical study is to estimate the effect of the treatment on the final exam, and this can be done directly by regressing `final` on `z`, or more efficiently by regressing `final` on `z` and also adjusting for `pretest`. But if we also adjust for `midterm`, which is an intermediate outcome (a post-treatment variable), we will get the wrong answer: the coefficient of `z` in that regression is a biased estimate of the treatment effect.

Here is the code:

```
n <- 1000
ability <- rnorm(n, 40, 10)
pretest <- ability + rnorm(n, 0, 5)
z <- rbinom(n, 1, 0.5)
midterm <- ability + 10 + 5*z + rnorm(n, 0, 5)
final <- ability + 20 + 10*z + rnorm(n, 0, 5)
fake <- data.frame(pretest, midterm, final, z)

print(stan_glm(final ~ z, data=fake, refresh=0))
print(stan_glm(final ~ z + pretest, data=fake, refresh=0))
print(stan_glm(final ~ z + pretest + midterm, data=fake, refresh=0))
```

## Drills

1. Experimental design

For each hypothetical experiment described below, state whether it is a randomized block experiment, a matched-pair experiment, a cluster-randomized experiment, or a simple randomized experiment.

- (a) To identify the long-term effects of special economic zones on economic growth, 10 out of 300 districts in a country are randomly selected to become special economic zones. The observational units are districts.  
*Solution:* This is a simple randomized experiment.
- (b) To examine the effect of a special dietary regimen on physical stamina, 10 out of 100 amateur soccer clubs in Europe are being assigned professional nutritionists. The observational units are individual players.
- (c) To study the impact of canvassing on voter turnout, five neighborhoods in Brooklyn and five neighborhoods in Queens get randomly assigned to be visited by volunteers for a weekend. The observational units are neighborhoods.
- (d) To understand the impact of free food on learning outcomes, 100 high schools in Chile get paired based on a number of characteristics, and one school in each pair gets selected randomly to receive free food deliveries. The observational units are schools.

## 2. Adjusting for post-treatment variables

For each hypothetical analysis described below, explain the problem with adjusting for post-treatment variables, and how this could be fixed.

- (a) It is hypothesized that electorates engage in party balancing, so that if one party is in power, some voters switch to preferring the opposite party. Consider a study of several cities, comparing those with mayors with Democratic or Republican party affiliation, where the outcome is a survey of attitudes of city residents to the out-party, adjusting for past voting patterns in the city and some measure of the performance of the local economy.  
*Solution:* The treatment must be defined at a specific time. You could look at outcomes in the years following elections, in which case you would want to use a measure of the performance of the local economy at some time before the election.
- (b) An experiment was conducted in which different groups of people were randomly given music recommendations, and then they were followed up to see which songs they were downloading and were asked to rate each song in the study on a 1–5 scale.<sup>55</sup> Suppose a regression is fit, where each data point is a person in the experiment, the outcome is the average download frequency for a certain song, the treatment is whether this song was recommended to this person, and the analysis also adjusts for the ratings this person gave to the song.
- (c) A study is conducted comparing two methods of teaching introductory statistics. The outcome of interest is the score on a final exam, and the analysis also adjusts for students' previous grades, performance on a pre-test, and performance on a midterm exam.

## Discussion problems

### 1. Individual and average effects

We were reading a book on data science and came across the following motivation for comprehensive integration of data sources, a story that is reminiscent of parables that appear in business books:<sup>56</sup>

“By some estimates, one or two patients died per week in a certain smallish town because of the lack of information flow between the hospital’s emergency room and the nearby mental health clinic. In other words, if the records had been easier to match, they’d have been able to save more lives. On the other hand, if it had been easy to match records, other breaches of confidence might also have occurred. Of course it’s hard to know exactly how many lives are at stake, but it’s nontrivial.”

The moral of the story:

“We can assume we think privacy is a generally good thing. . . . But privacy takes lives too, as we see from this story of emergency room deaths.”

This particular story raised our suspicions. One or two patients per week is 75 in a year. Can you figure out if this number is plausible? This can be a good discussion problem involving numeracy and quantitative understanding of possible causal effects.

Here is our attempt to answer this question. To calibrate, we’d like to get a denominator, the total number of deaths each year.

We’re not sure how large the “smallish town” is. From Wikipedia’s discussion of towns in the United States: “In some instances, the term ‘town’ refers to a small incorporated municipality of

<sup>55</sup>Matthew Salganik, Peter Dodds, and Duncan Watts (2006), Experimental study of inequality and unpredictability in an artificial cultural market, *Science* 311, 854–856.

<sup>56</sup>David Crawshaw and Josh Wills (2013), Data engineering: MapReduce, Pregel, and Hadoop, in *Doing Data Science*, edited by Rachel Schutt and Cathy O’Neil. The discussion here is taken from Andrew Gelman (2014), Parables vs. stories, <https://statmodeling.stat.columbia.edu/2014/01/24/parables-vs-data/>.

less than 10,000 people, while in others a town can be significantly larger. Some states do not use the term ‘town’ at all, while in others the term has no official meaning . . .” Wikipedia then goes state by state, for example, “In Alabama, the legal use of the terms ‘town’ and ‘city’ are based on population. A municipality with a population of 2000 or more is a city, while less than 2000 is a town.”

Just to go forward on this, suppose the “smallish town” has 10 000 people. Given that people live approximately 70 years, we can very roughly estimate that 1/70 of the population is dying every year, which would be 140 deaths per year. In that case, the above-quoted claim can’t possibly be right—there’s no way that half the deaths in this town were caused by poor record-keeping between a hospital and a mental health clinic. If the town had 20 000 people (which would seem beyond the size of a town that one would call “smallish,” at least in the United States), then we’re talking 1/4 of the deaths, which still seems much too large a proportion. Even if it is a town with lots of old people, so that much more than 1/70 of the population is dropping off each year, the numbers don’t even come close to adding up.

Based on these calculations, we think there is something missing in the story that was told about the hospital records. Perhaps someone dropped a couple of zeroes when reporting some estimate, or maybe the authors heard about some case of someone dying because of poor record keeping in some town somewhere and just made up the “one or two patients died per week” because it sounded plausible to them and they weren’t thinking it through, kind of like how you might say that your car weighs 80 tons, because you’ve never really understood exactly what a ton is. Or we might be missing something here; it’s hard for us to say more, because the story is not sourced. Our point here is that it can be possible to evaluate causal claims by stepping back and thinking about statistical claims in a quantitative way.

## 2. Meta-analysis of nudge experiments

Here is the abstract of a published meta-analysis of “nudge” interventions:<sup>57</sup>

“Over the past decade, choice architecture interventions or so-called nudges have received widespread attention from both researchers and policy makers. . . . Drawing on more than 200 studies . . . we present a comprehensive analysis of the effectiveness of choice architecture interventions . . . . Our results show that choice architecture interventions overall promote behavior change with a small to medium effect size of Cohen’s  $d = 0.45$  (95% CI [0.39, 0.52]). . . . Food choices are particularly responsive to choice architecture interventions, with effect sizes up to 2.5 times larger than those in other behavioral domains. . . . Our analysis further reveals a moderate publication bias toward positive results in the literature. . . .”

Actually, though, we find no reason to expect large and consistent effects of nudge interventions. Defaults can be considered as a form of nudging, and defaults can make a big difference sometimes.<sup>58</sup> The evidence for the effectiveness of nudges other than defaults is not so clear.

How can we say this given the strong claims made in the above-quoted article? The problem is that the meta-analysis is summarizing a literature of noisy data, small sample sizes, and selection on statistical significance, hence massive overestimates of effect sizes. This is not a secret: follow the references and look at the papers in question and you will see, over and over again, that they are selecting what to report based on whether the  $p$ -value is less than 0.05. The problem here is

<sup>57</sup>Stephanie Mertens, Mario Herberz, Ulf Hahnel, and Tobias Brosch (2022), The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains, *Proceedings of the National Academy of Sciences* 119, e2107346118. For our take, see Andrew Gelman (2022), PNAS GIGO QRP WTF: This meta-analysis of nudge experiments is approaching the platonic ideal of junk science, <https://statmodeling.stat.columbia.edu/2022/01/07/pnas-gigo-qrp-wtf-approaching-the-platonic-ideal-of-junk-science/>, and Barnabás Szászi, Anthony Higney, Aaron Charlton, Andrew Gelman, Ignazio Ziano, Balázs Aczel, Daniel Goldstein, David Yeager, and Elizabeth Tipton, No reason to expect large and consistent effects of nudge interventions, *Proceedings of the National Academy of Sciences* 119, e2200732119.

<sup>58</sup>See, for example, Eric Johnson and Daniel Goldstein (2003), Do defaults save lives? *Science* 302, 1338–1339.

not the  $p$ -value but rather the selection, which induces noise (through the reduction of continuous data to a binary summary) and bias (by not allowing small effects to be reported at all).

Consider the implications of this meta-analysis and potential concerns. How can you think about treatment effects that are estimated from a messy literature, and how can these be studied going forward?

## 4.23 Imbalance and lack of complete overlap

### Plan for two classes

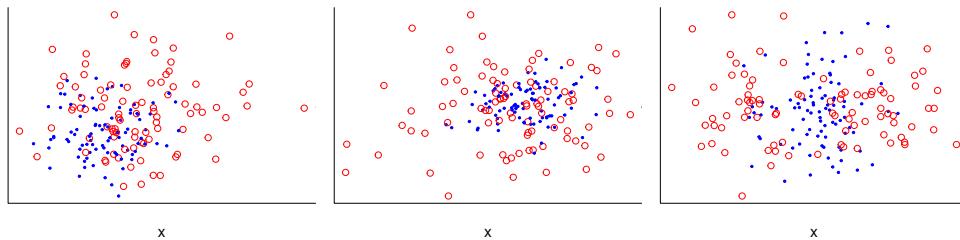
Stories	Activities	Computer demonstrations	Drills	Discussion problems
Retrospective evaluation of a policy	Imbalance and lack of overlap	Poststratification for causal inference	Ignorability of treatment assignment	Effects of campaign contributions
Postal service modeling	Victimization and views on crime policy	Measurement-error models	Imbalance and lack of overlap	Effects and variation

### Reading

Chapter 20 of *Regression and Other Stories*: Observational studies with all confounders assumed to be measured (Sections 20.1–20.6)

### Pre-class warmup assignments

1. Imbalance and lack of complete overlap
  - (a) For each of these graphs of outcome  $y$  vs. pre-treatment predictor  $x$ , describe the imbalance and lack of complete overlap. Dots and open circles represent control and treated units, respectively.



- (b) Sketch an example of two predictors,  $x_1$  and  $x_2$ , using red and blue dots for treatment and controls, where there is *balance* on each of  $x_1$  and  $x_2$  but not on their joint distribution.  
(c) Sketch an example of two predictors,  $x_1$  and  $x_2$ , using red and blue dots for treatment and controls, where there is *complete overlap* on each of  $x_1$  and  $x_2$  but not on their joint distribution.

### 2. Subclassification and average treatment effects

Suppose you are conducting an observational study comparing two different approaches to counseling of caregivers of family members. The outcome is a mental health survey of caregivers and the study also includes three pre-treatment predictors on the caregivers: age, sex, and a measure of life satisfaction.

- (a) How could you obtain estimates of the effect among all the caregivers in the study, the effect among caregivers with pre-treatment life satisfaction that was lower than average, and among caregivers with pre-treatment life satisfaction that was higher than average?
- (b) You are concerned that your study is not representative of the general population of caregivers. How can you estimate the population average treatment effect using the data from your study?

## Homework assignments

### 1. (a) Messy randomization (Exercise 19.8 of *Regression and Other Stories*)

The folder Cows contains data from an agricultural experiment that was conducted on 50 cows to estimate the effect of a feed additive on six outcomes related to the amount of milk fat produced by each cow.

Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of the cow.

Cows were initially assigned to treatments completely at random, and then the distributions of the three pre-treatment predictors were checked for balance across the treatment groups; several randomizations were tried, and the one that produced the “best” balance with respect to the three covariates was chosen. The treatment depends only on fully observed covariates and not on unrecorded variables such as the physical appearances of the cows or the times at which the cows entered the study, because the decisions of whether to re-randomize are not explained.

We shall consider different estimates of the effect of the additive on the mean daily milk fat produced.

i. Consider the simple regression of mean daily milk fat on the level of additive. Compute the estimated treatment effect and standard error, and explain why this is not a completely appropriate analysis given the randomization used.

ii. Add more predictors to the model. Explain your choice of which variables to include. Compare your estimated treatment effect to the result from (a).

iii. Repeat (b), this time considering additive level as a categorical predictor with four levels. Make a plot showing the estimate (and standard error) of the treatment effect at each level, and also showing the inference from the model fit in part (b).

### (b) Causal inference based on data from individual choices (Exercise 19.9 of *Regression and Other Stories*)

Our lives involve tradeoffs between monetary cost and physical risk, in decisions ranging from how large a car to drive, to choices of health care, to purchases of safety equipment. Economists have estimated how people implicitly trade off dollars and danger by comparing choices of jobs that are similar in many ways but have different risks and salaries. This can be approximated by fitting regression models predicting salary given the probability of death on the job (and other characteristics of the job). The idea is that a riskier job should be compensated with a higher salary, with the slope of the regression line corresponding to the “value of a statistical life.”

i. Set up this problem as an individual choice model, as in Section 15.7 of *Regression and Other Stories*. What are an individual’s options, value function, and parameters?

ii. Discuss the assumptions involved in assigning a causal interpretation to these regression models.

### 2. (a) Understanding the difference between average treatment effect in the treated and control groups (Exercise 20.3 of *Regression and Other Stories*)

Create a hypothetical dataset in which the average treatment effects on the treated and controls (ATT and ATC) are clearly different. What are the distinguishing characteristics of this dataset?

### (b) Observational studies and hypothetical experiments (Exercise 20.11 of *Regression and Other Stories*)

Consider an applied problem of interest to you with a causal effect that has been estimated using an observational study. Think about possible hypothetical experiments that could be

performed to estimate different aspects of this causal effect. Consider how the effect might vary across the population and across different implementations of the treatment.

- (c) *In pairs:* Working through your own example (Exercise 19.12 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, estimate the causal effect you defined in Exercise 18.18 of *Regression and Other Stories* using a regression of  $y$  on the treatment indicator  $z$ , at least one pre-treatment predictor  $x$ , and their interaction. Plot the data and the fitted regression lines for treatment and control, and discuss the assumptions that are required for this to be a good estimate of the causal effect of interest.

## Stories

### 1. Retrospective controlled evaluation of a policy experiment

The Millennium Villages Project (MVP) was an integrated rural development program carried out for a decade in 10 clusters of villages in sub-Saharan Africa starting in 2005 and in a few other sites for shorter durations.<sup>59</sup> The project has been controversial, both in its conception and in evaluation of its effects. The starting point for the controversy was the project's approach of economic and social development catalyzed by foreign aid, which has been criticized as a doomed-to-fail relic of a bygone paternalistic era. In addition, the project was criticized for not being designed as a randomized controlled trial. The MVP stands out as a high-profile project organized by an academic economist that did not include a control group.

At the inception of the MVP, two reasons were given for not designing the MVP as a randomized controlled trial. First, the MVP used a basket of many interventions that had already been shown to work, often through previous controlled trials. The main focus of the MVP was on the feasibility of implementing the package of proven interventions within the specified budget and timeline, a concern for which a control group is not relevant. Second, the MVP did not have an adequate project budget to engage systematically with control sites, especially to be able to offer those other sites the package of interventions at a later date. From a pragmatic, political, and ethical point of view, the MVP was therefore wary of identifying and engaging actively with non-project sites.

Eventually, however, the leaders of the project decided to perform an empirical assessment of its effectiveness, a retrospective controlled evaluation that was constructed by comparing outcomes in the 10 MVP sites to 10 other locations within the countries being studied. We retrospectively selected comparison villages that best matched the project villages on possible confounding variables based on data from 2005, thus only matching on pre-treatment information. Cross-sectional survey data on 40 outcomes of interest were collected from both the project and the comparison villages in 2015. Using these data, as well as on-site spending data collected during the project, we estimated project impacts as differences in outcomes between the project and comparison villages; target attainment as differences between project outcomes and prespecified targets; and on-site spending as expenditures reported by communities, donors, governments, and the project. We estimated generally favorable impacts—that is, the Millennium Villages showed better outcomes than the comparison villages, with differences larger than would be expected by chance given the variation in the data across the 10 locations:

“The MVP had favourable impacts on outcomes in all MDG [Millennium Development Goal] areas, consistent with an integrated rural development approach. The greatest effects were in agriculture and health, suggesting support for the project’s emphasis on agriculture and health systems strengthening.”

Two years later, a different research group, unaffiliated with the project, performed a separate

<sup>59</sup>See Andrew Gelman, Shira Mitchell, Jeffrey Sachs, and Sonia Sachs (2022), Reconciling evaluations of the Millennium Villages Project, *Statistics and Public Policy* 9, 1–9, and Shira Mitchell et al. (2018), The Millennium Villages Project: A retrospective, observational, endline evaluation, *Lancet Global Health* 6, e500–e513.

analysis of a single MVP site in northern Ghana. They did a prospective controlled comparison, also not randomly assigned but with the control villages selected at the beginning of the project. They reported mostly small or null results:<sup>60</sup>

“Our study finds that the impact of MVP on the MDGs was limited, and that core welfare indicators such as monetary poverty, child mortality and under-nutrition were not affected. . . . despite some positive impacts, we found mostly null results, suggesting that the intervention was ineffective.”

We looked into the details to try to understand how these two studies came to such different conclusions. We concluded that the key differences were: (1) the full MVP study went for 10 years, whereas the later study was conducted for less than five and hence would be expected to have smaller effects, and (2) effects vary by location, and with just one site it is difficult to get a precise estimate. Given the inherent noisiness in estimates for a single site over a short time period, we feel it was a mistake for the report of the second study to summarize in terms of statistical significance (for example, “the count of statistically significant impacts is low”) and to report non-significant comparisons as if they were zero (for example, “we found mostly null results, suggesting that the intervention was ineffective”).

Beyond these technical issues, much of the difference in the conclusions of the two reports can be explained by differences in framing. Is it a plus that “the project conclusively met one third of its targets,” or is it a disappointment that, although effects were positive, they mostly did not reach the original stated goals? The two apparently contradictory evaluations of the Millennium Villages Project are both consistent with a larger picture in which the MVP has positive average effects (compared to untreated villages) across a broad range of outcomes, but with effects that are variable across sites and that require several years to take effect.

This story is relevant to the week’s reading as an example of causal inference for an observational study using matching. In discussing the applied problem, we have considered some of the practical reasons why controlled experimentation is not always done, and the controversies that can result from this decision. The example is relevant to the course as a whole as an example of a policy dispute that turns on statistical analysis.

## 2. Postal Service cost modeling: What are inferences used for?

Several years ago, we worked with some statisticians and economists on a project with the U.S. Postal Service. The goal was to estimate the volume of all sorts of mail (first-class mail of different weights, second-class mail, packages, etc.) handled by the Postal Service, along with the costs of processing these. The reason was that the Postal Service was required by law to charge an appropriate fee for each of its services, neither overcharging any of its customers by taking advantage of its monopoly position nor undercutting any of its competitors by cross-subsidizing any of its services. To this end, the Postal Service had several large ongoing sampling projects, including a survey of mail pieces based on a sample of locations around the country, a time-use survey of employees, and various smaller studies. These surveys were in continuous operation and cost millions of dollars per year. Our team was tasked to evaluate the system that existed at the time and make suggestions for improvement.<sup>61</sup>

After immersing ourselves in the details of the sampling schemes and cost estimates, we decided to set up a simulation model to propagate the estimates and uncertainties in all stages of the calculations. In doing this, we learned that many of the largest contributors to the uncertainties came from small offline studies that had been performed to estimate elasticities.

<sup>60</sup>Edoardo Masset, Jorge Hombrados, and Arnab Acharya (2020), Aiming high and falling low: The SADA-Northern Ghana Millennium Village Project, *Journal of Development Economics* 143, 102427.

<sup>61</sup>Richard Waterman, Donald Rubin, Neal Thomas, and Andrew Gelman (2000), Simulation modeling for cost estimation, in *Current Directions in Postal Reform*, edited by Michael Crew and Paul Kleindorfer, 171–193, <http://www.stat.columbia.edu/~gelman/research/published/postal.pdf>.

To simplify slightly, it went like this: to get a simple estimate of the cost per unit of handling mail type  $X$ , you can take the total cost spent on handling  $X$  during the past year, divided by the total number of pieces of type  $X$  mailed during the year. The numerator is estimated from the time use survey and the denominator is estimated using the survey of mail pieces. A lot of effort had been placed into minimizing the biases in these surveys and making sure the samples were large enough that variances of the relevant summaries were small.

But this simple ratio does not answer the question of interest, first because of various complexities (for example, an employee when contacted by the time use survey might be handling many pieces of mail at once or might be doing a part of the job that cannot be simply allocated across types of mail), and also because there are other costs beyond employee hours, for example cost of driving and repairing the trucks and planes used to transport the mail. To estimate the allocation these costs, the Post Office had performed a series of studies estimating “elasticities” such as the extra amount of fuel consumed per weight of mail. Each of these elasticities was estimated from its own little regression analysis that would result in an estimate and standard error, and in our big simulation model we propagated all these uncertainties. As noted above, it turned out that the standard errors in these little elasticity studies were driving the final uncertainties in the cost estimates. So, even though we’d started by focusing on the design of the big surveys, our final recommendations were to redo the elasticity studies, and we recommended an ongoing program where these elasticities were periodically re-estimated.

This story relates to the week’s reading in that a problem which appeared to be all about survey sampling turned out ultimately to center on observational studies, using available data to estimate things such as the effect on fuel costs of mailing one more 1-ounce letter. The story relates to the course as a whole as an example of a problem where the goal of inference was not a particular parameter or coefficient in a model but rather a combination of unknowns that were estimated from many different sources. We ask the students in pairs if they can think of other problems where this is the case.

### Class-participation activities

#### 1. Imbalance and lack of overlap

The activity here is to consider some example of real-world imbalance and lack of overlap in observational studies. Students should first divide into groups of four to consider possible examples on topics of interest to them. The class as a whole can then discuss several of these examples, one at a time, going over imbalance and lack of overlap for each pre-treatment variable of interest and discussing how these problems could be addressed. This is relevant to the week’s reading in elaborating on these important concepts in causal inference, and it relates to the course as a whole in being an opportunity to connect a statistical challenge to topics of interest to students.

#### 2. Observational study on students: Crime victimization and policy views

There is a saying that “a conservative is a liberal who has been mugged.” What effect does being a crime victim have on attitudes toward crime and criminal justice?<sup>62</sup>

We address this question with a class-participation demonstration gathering responses from students on crime victimization and attitudes on justice policies. Two challenges arise: first, deciding what questions to ask; second, the study is observational so it will be necessary to adjust for differences between the treated and control groups, however they are defined.

<sup>62</sup>There is some literature on this topic, for example Daniel J. Koenig (1980), The effects of criminal victimization and judicial or police contacts on public attitudes toward local police, *Journal of Criminal Justice* 8, 243–249, R. Thomas Dull and Arthur Wint (1997), Criminal victimization and its effect on fear of crime and justice attitudes, *Journal of Interpersonal Violence* 12, 748–758, James Unnever, Francis Cullen, and Bonnie Fisher (2007), “A liberal is someone who has not been mugged”: Criminal victimization and political beliefs, *Justice Quarterly* 24, 309–334, and Anna King and Shadd Maruna (2009), Is a conservative just a liberal who has been mugged?: Exploring the origins of punitive views, *Punishment & Society* 11, 147–169.

- Pre-treatment predictors  $x$ : Background variables
- Treatment (or exposure)  $z$ : Victim of a crime
- Outcome  $y$ : Tough-on-crime attitudes

**Figure 112** *Setup for the in-class crime victimization survey. The instructor can display this on the screen to guide a class discussion of what questions to include on their crime victimization and attitudes survey.*

The questions to ask will depend on the local political context. In our class at Columbia University, we asked a set of questions regarding a policy issue that was covered in a local news story:<sup>63</sup>

“The poll question asked: ‘Three years ago, New York passed a law eliminating monetary bail for people facing misdemeanor and non-violent felony charges. Which of the following two views is closest to yours? “The so-called bail reform law should be amended to give judges more discretion to keep dangerous criminals off the streets.” — or “The law should not be amended to give discretion on bail back to judges because it could once again lead to people of color being disproportionately denied bail.”’

A total of 65 percent of the 803 respondents said the bail law should be amended to take into account a defendant’s prior violent record, and only 27 percent of voters said the law should not be changed, while the rest were undecided.”

After sharing this news story, we displayed Figure 112 and used this to organize a discussion on what questions should be asked. We ended up with the following questions:

- Do you favor or oppose the death penalty for persons convicted of murder? (five options: “strongly favor,” “favor,” “no opinion,” “oppose,” “strongly oppose”)
- Do you agree or disagree with the statement that crime is a serious problem in New York? (five options, from “strongly agree” to “strongly disagree”)
- Do you agree or disagree with the statement that criminal justice should be more severe? (five options)
- Do you favor or oppose the law eliminating monetary bail for people facing misdemeanor and non-violent felony charges? (five options)
- Have you been a victim of a crime? (three options: “No,” “Yes, a minor crime,” “Yes, a major crime”)
- Demographic items: age, sex, and highest education level of parents (five levels, from no high school degree to postgraduate degree)

We typed these into a Google form, which the students then filled out using their laptops or phones, then we downloaded the data and performed a simple analysis. First we combined the attitude questions into a single “tough on crime” measure (being careful to scale each item in the appropriate regression). Next, we compared the average values of this score for students in each of the three victimization categories. We then fit a regression predicting the score given victimization level (coded as 0, 1, or 2), also adjusting for the pre-treatment predictors. The resulting estimate had a high standard error—no surprise given the small sample size. For the purpose of discussion we asked students to ignore the uncertainties, pretend that the point estimate of the coefficient represented a population value, and then discuss potential problems with the observational study—differences between the different groups that were not fully addressed by the pre-treatment predictors—and how it could be improved.

<sup>63</sup>Carl Campanile (2022), 65% of New York voters back stricter bail law, *New York Post*, 22 Feb, <https://nypost.com/2022/02/22/65-percent-of-new-york-voters-back-stricter-bail-law-poll/>.

This activity relates to the week's reading by bringing to life the challenges of causal inference from an observational study. It is relevant for the course as a whole in linking data collection to analysis and real-world conclusions.

### Computer demonstrations

#### 1. Simulated-data example of poststratification for causal inference

Consider a hypothetical experiment that evaluates a yoga program that is intended to reduce stress. The challenge is that retirees are overrepresented in the experiment, and we are interested in the average treatment effect in the population. In this demonstration, we simulate pre-test, treatment, and post-test data for the sample and then adjust for the mismatch between sample and population. For simplicity we assume that stress is measured on a standardized continuous scale so we do not need to worry about discrete-data issues.

In general we can estimate population average causal effects by regression and poststratification.<sup>64</sup>

```
# Simulate the sample data
n <- 2000
retiree <- rbinom(n, 1, 0.4) # Approx 40% of sample are retirees
pre_stress <- rnorm(n, ifelse(retiree==1, 90, 100), 20) # Retirees have less stress
z <- sample(c(0,1), n, replace=TRUE) # Randomized treatment assignment

# Simulate from a regression model on pre_stress, where the treatment effect
# is 10 for non-retirees and 5 for retirees
post_stress <- rnorm(n, 100 + 0.7*(pre_stress - 100) - 10*z + 5*z*retiree, 15)

# Fit a regression to sample data
sample_data <- data.frame(retiree, pre_stress, z, post_stress)
fit <- stan_glm(post_stress ~ pre_stress + retiree + z + retiree:z, data=sample_data)
print(fit)

# Posterior simulations of sample average treatment effect
sims <- as.matrix(fit)
sate <- sims[, "z"] + sims[, "retiree:z"] * mean(sample_data$retiree)
cat("Estimate and s.e. of sample avg treatment effect:",
  c(mean(sate), sd(sate)), "\n")

# Posterior simulations of population average treatment effect,
# assuming 20% of population are retirees
pate <- sims[, "z"] + sims[, "retiree:z"] * 0.20
cat("Estimate and s.e. of population avg treatment effect:",
  c(mean(pate), sd(pate)), "\n")
```

The problem would become more challenging if the treatment effect varies by pre-test, because then it is necessary to poststratify over pre-test, which requires some knowledge or assumptions about pre-treatment stress levels in the population. Here we simulated the simplest case of poststratification on a binary variable that is assumed completely known in the population.

#### 2. Measurement-error models

Section 22.2 of *Regression and Other Stories* mentions that in a regression of  $y$  on  $x$ , if  $x$  is

<sup>64</sup>See, for example, Jennifer Hill (2011), Bayesian nonparametric modeling for causal inference, *Journal of Computational and Graphical Statistics* 20, 217–240, and Lauren Kennedy and Andrew Gelman (2021), Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample, *Psychological Methods* 26, 547–558.

measured with error, this will induce bias in the estimated coefficients. If  $y$  is measured with error, this induces variance but not bias. We demonstrate this here with a simulation, where there is a true relation,  $y = a + bx + \text{error}$ , with  $x^*$  and  $y^*$  defined as  $x$  and  $y$  measured with error:

```
n <- 1000
x <- rnorm(n, 50, 10)
a <- 20
b <- 0.8
y <- a + b*x + rnorm(n, 0, 10)
plot(x, y)

y_star <- y + rnorm(n, 0, 10)
x_star <- x + rnorm(n, 0, 10)
data <- data.frame(x, y, x_star, y_star)
```

First fit a regression of the latent data to confirm that you reproduce the underlying relation of interest:

```
fit_1 <- stan_glm(y ~ x, data=data, refresh=0)
print(fit_1, digits=2)
```

Next try the regression with the observed data,  $x^*$  and  $y^*$ , which will give a different result, as the measurement error induces bias in the estimate:

```
fit_2 <- stan_glm(y_star ~ x_star, data=data, refresh=0)
print(fit_2, digits=2)
```

To understand what is going on, next try regressing the observed outcome on the latent predictor, and then the latent outcome on the observed predictor:

```
fit_3 <- stan_glm(y_star ~ x, data=data, refresh=0)
print(fit_3, digits=2)
```

```
fit_4 <- stan_glm(y ~ x_star, data=data, refresh=0)
print(fit_4, digits=2)
```

If there is time, you can then try this more elaborate example:

```
n <- 1000
x <- rnorm(n, 50, 10)
z <- rbinom(n, 1, 0.5)
a <- 20
b <- 0.8
theta <- 5
y <- a + b*x + theta*z + rnorm(n, 0, 10)
data_new <- data.frame(x, y, z)
fit_new <- stan_glm(y ~ x + z, data=data_new, refresh=0)
print(fit_new)

error <- rbinom(n, 1, 0.25)
z_star <- ifelse(error==1, 1-z, z)
data_new$z_star <- z_star
fit_new_2 <- stan_glm(y ~ x + z_star, data=data_new, refresh=0)
print(fit_new_2)

data_new$x_star <- data_new$x + rnorm(n, 0, 10)
fit_new_3 <- stan_glm(y ~ x_star + z, data=data_new, refresh=0)
print(fit_new_3)
```

## Drills

### 1. Ignorability of treatment assignment

In the following cases of observational studies, discuss possible problems with the assumption of ignorability of the treatment assignment, conditional on the pre-treatment predictors and with respect to the potential outcomes.

- (a) Regression of health outcomes on a treatment indicator, with the treatment chosen by patients' doctors.

*Solution:* The ignorability assumption is implausible here unless the predictors include the information that doctors used to choose the treatment for each patient. If the treatment assignment was based only on information in the patients' written charts, then those data could be included as regression predictors. If the treatment assignment was also based on unrecorded information, for example doctors' impressions after observing and talking with the patients, then ignorability would be more difficult to attain.

- (b) Regression of income at age 40 on an indicator of having been admitted to university, including high school grade point average as a pre-treatment predictor, in a country where there is only one university, admission depends only on high-school grade point average, and there are so many applicants that many equally capable applicants get rejected.
- (c) Regression across a number of countries of voter turnout on perceived levels of corruption, as well as on predictors measuring the vote share of the incumbent party in the previous election, per-capita GDP, and indicators for different types of colonial history.

### 2. Imbalance and lack of complete overlap

In the following examples of observational studies, discuss possible concerns about imbalance or lack of complete overlap.

- (a) Regression of income at age 40 on an indicator of having been admitted to university, including high school grade point average as a pre-treatment predictor, in a country where there is only one university, admission is only dependent on high-school grade point average, and there are so many applicants that many equally capable applicants get rejected.

*Solution:* There should be overlap on high-school grade-point average in the middle-to-high end of the range, where admission is competitive so some students enter university and others do not. We might see lack of overlap at the low end of the range, as only very few students with low grades may be admitted to the university.

- (b) Regression of farm yields on the administration of fertilizer as well as a covariate indicating precipitation levels, if all treated units are located in Ireland and all control units are located in Morocco.
- (c) Regression of sentencing outcomes on a variable indicating the time of day of the sentencing, when the analysis is limited to one courthouse and time assignment is random.
- (d) Regression of personal consumption on an indicator whether a person lives in New York or in Pennsylvania, including predictors for income, age, sex, and number of children.

## Discussion problems

### 1. Causal challenges in estimating the effects of campaign contributions on politicians' behaviors

The first step here is to consider possible treatments or interventions. Be specific: the treatment is not just "campaign contributions"; it must be some specific activity that could be done, for example anonymously donating \$100 to a particular campaign. Even that would not be specific enough, as you would want to define exactly when the money would be given. Next you need to define a population—which candidates and which elections—and outcomes of interest. Only at that point does it make sense to specify a hypothetical experiment involving randomized treatment

assignment. The next step is to consider why it could be difficult to do this, along with limitations of such an experiment—for one thing, it could only be done in the future so could not address questions of past effects.

The usual approach to causal inference is to start with the general question (What is the effect of X on Y?, with X and Y only vaguely defined) and available data and work back from there. The point of this discussion problem is that being specific about treatments, units, pre-treatment measurements, and post-treatment outcomes can be a good way to think about a problem.

## 2. Variation in social science patterns

A study was done concluding,<sup>65</sup>

“Sports participation causes women to be less religious, more likely to have children, and, if they do have children, more likely to be single mothers.”

The authors of that study continued:

“It is true that many successful women with professional careers, such as Sheryl Sandberg and Brandi Chastain, are married. This fact, however, is not necessarily opposed to our hypothesis. Women who participate in sports may ‘reject marriage’ by getting divorces when they find themselves in unhappy marriages. Indeed, Sheryl Sandberg married and divorced before marrying her current husband.”

Accepting the claim in the first quote as stated (actually the statistical evidence for the claim is not so clear), in what way does the second quote betray a misunderstanding?

---

<sup>65</sup>See the story on page 47 of this book.

## 4.24 Additional topics in causal inference

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Deterrent effect of death penalty	Two measures of the same quantity	Instrumental variables	Assumptions for instrumental	Effects of masks
Regression discontinuity mishaps	“Why” questions and causal inference	Adjustment in regression discontinuity	Regression discontinuity	Coaching for admissions tests

### Reading

Chapter 21 of *Regression and Other Stories*: Additional topics in causal inference

### Pre-class warmup assignments

#### 1. Instrumental variables

- An instrumental variables analysis is done in which the estimated average intent-to-treat effect is 0.5 and the estimated average effect of the instrument on the treatment indicator is 0.7. What is the estimated treatment effect?
- A medical experiment is conducted in which patients are randomly assigned to one of two groups: a group where an experimental blood-pressure-lowering drug is offered, and a group where the drug is not offered. It is recorded who in the first group actually takes the drug; nobody in the second group takes the drug. Everyone's blood pressure is measured before and after the experiment, and an instrumental variables analysis is done to estimate the effect of the drug on blood pressure.

This estimates the average treatment effect among what group of patients?

#### 2. Adjust for pre-treatment variables

- A medical experiment is conducted in which patients are randomly assigned to one of two groups: a group where an experimental blood-pressure-lowering drug is offered, and a group where the drug is not offered. It is recorded who in the first group actually takes the drug; nobody in the second group takes the drug. Everyone's blood pressure is measured before and after the experiment, and an instrumental variables analysis is done to estimate the effect of the drug on blood pressure.

How should this analysis make use of the pre-treatment measurement?

- An experiment is performed in which patients' blood pressure is measured, and then the patients whose blood pressure is higher than some preset threshold are given an experimental blood-pressure-lowering drug. Everyone who is offered this drug takes it, and the drug is not available to those who are not offered it. After the experiment, everyone's blood pressure is measured.

A regression discontinuity analysis is then performed. Suppose that you would like to also adjust for patient age in the analysis. How would you do it?

### Homework assignments

- (a) Instrumental variables (Exercise 21.1 of *Regression and Other Stories*)

The following study is performed at a university. Students are sent emails encouraging them to click on a university website. Each student is randomly assigned to one of two sites: a site with encouragement to vote in the upcoming student government election, and a neutral site with study tips. The students are then followed up to see if they voted. Define  $y = 1$  if the student voted or 0 otherwise; define  $u = 1$  if the student was assigned to the encouragement site or 0 if assigned to the neutral site; define  $v = 1$  if the student actually accessed the site (which can be checked using unique identifiers) or 0 if the student never clicked on the link.

- i. From which of the following regressions or pair of regressions can we compute the instrumental variables estimate of the effect of accessing the site on voting?

- Regression of  $y$  on  $u$ .
- Regression of  $y$  on  $v$ .
- Regression of  $y$  on  $u$  and  $v$ .
- Regression of  $y$  on  $u$ , and the regression of  $v$  on  $u$ .
- Regression of  $y$  on  $v$ , and the regression of  $v$  on  $u$ .

- ii. What assumptions are required for the instrumental variables estimate to be reasonable in this case? Do these assumptions seem plausible here?

(b) Intermediate outcomes (Exercise 21.15 of *Regression and Other Stories*)

In Exercise 19.10 of *Regression and Other Stories*, you estimated the effect of incumbency on votes for Congress. Now consider an additional variable: money raised by the congressional candidates. Assume that this variable has been coded in some reasonable way to be positive in districts where the Democrat has raised more money and negative in districts where the Republican has raised more.

- i. Explain why it is inappropriate to include money as an additional input variable to “improve” the estimate of incumbency advantage in the regression in Exercise 19.10 of *Regression and Other Stories*.
- ii. Suppose you are interested in estimating the effect of money on the election outcome. Set this up as a causal inference problem (that is, define the treatments and potential outcomes).
- iii. Explain why it is inappropriate to simply estimate the effect of money using instrumental variables, with incumbency as the instrument. Which of the instrumental variables assumptions would be reasonable in this example and which would be implausible?
- iv. How could you estimate the effect of money on congressional election outcomes?

(c) Causes of effects and effects of causes (Exercise 21.17 of *Regression and Other Stories*)

Apply the ideas of Section 21.5 of *Regression and Other Stories* to an applied problem of interest for you. Consider reverse causal questions and then potential forward causal questions, along with associated experiments or experiments that could be used to estimate relevant causal effects.

2. (a) Regression discontinuity (Exercise 21.9 of *Regression and Other Stories*)

Take the Chile schools example from Section 21.3 and perform a series of analyses using different subsetting ranges. Plot the estimate  $\pm$  standard error as a function of the subsetting range. As the width of the range increases, the standard error should go down, but there should be more of a concern about the use of the estimate for causal inference, given the lack of overlap.

(b) Regression discontinuity (Exercise 21.10 of *Regression and Other Stories*)

Suppose you are trying to evaluate the effect of a new procedure for coronary bypass surgery that is supposed to help with the postoperative healing process. The new procedure is risky, however, and is rarely performed in patients who are over 80 years old. Data for this (hypothetical) example are displayed in Figure 113.

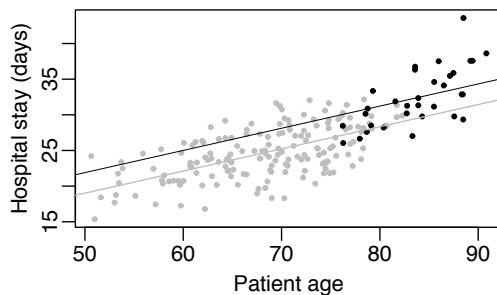


Figure 113 Hypothetical data of length of hospital stay and age of patients, with separate points and regression lines plotted for each treatment condition: the new procedure in gray and the old procedure in black.

- i. Does this seem like an appropriate setting in which to implement a regression discontinuity analysis?
  - ii. The folder Bypass contains data for this example: stay is the length of hospital stay after surgery, age is the age of the patient, and new is the indicator for whether the new surgical procedure was used. Preoperative disease severity (severity) was unobserved by the researchers, but we have access to it for illustrative purposes. Can you find any evidence using these data that the regression discontinuity design is inappropriate?
  - iii. Estimate the treatment effect using a regression discontinuity estimate (ignoring) severity. Estimate the treatment effect in any way you like, taking advantage of the information in severity. Explain the discrepancy between these estimates.
- (c) *In pairs:* Working through your own example (Exercise 20.12 of *Regression and Other Stories*)  
Continuing the example from the final exercises of the earlier chapters, consider a treatment effect that can only be estimated using an observational study. Using your data, assess issues of imbalance and lack of overlap. Use matching if necessary to get comparable treatment and control groups, then perform a regression analysis adjusting for matching variables or the propensity score and estimate the treatment effect. Graph the data and fitted model and assess your assumptions.

## Stories

### 1. Deterrent effect of the death penalty

In 2006, law professor John Donohue and economist Justin Wolfers summarized some of the research on the effects of capital punishment:<sup>66</sup>

“Over much of the last half-century, the legal and political history of the death penalty in the United States has closely paralleled the debate within social science about its efficacy as a deterrent. Sociologist Thorsten Sellin’s careful comparisons of the evolution of homicide rates in contiguous states from 1920 to 1963 led to doubts about the existence of a deterrent effect caused by the imposition of the death penalty. This work likely contributed to the waning reliance on capital punishment, and executions virtually ceased in the late 1960s. In 1975, Isaac Ehrlich’s analysis of national time-series data led him to claim that each execution saved eight lives. Solicitor General Robert Bork cited Ehrlich’s work to the Supreme Court a year later, and the Court, while claiming not to have relied on the empirical evidence, ended the death penalty moratorium when it upheld various capital punishment statutes . . .”

<sup>66</sup>John Donohue and Justin Wolfers (2006), Uses and abuses of empirical evidence in the death penalty debate, *Stanford Law Review* 58, 791–845. The discussion here is taken from Andrew Gelman (2006), The deterrent effect of the death penalty, [https://statmodeling.stat.columbia.edu/2006/01/18/decision\\_analys\\_2/](https://statmodeling.stat.columbia.edu/2006/01/18/decision_analys_2/).

Donohue and Wolfers review the data on death sentences and homicide:

"No clear correlation between homicides and executions emerges from this long time series. In the first decade of the twentieth century, execution and homicide rates seemed roughly uncorrelated, followed by a decade of divergence as executions fell sharply and homicides trended up. Then for the next forty years, execution and homicide rates again tended to move together—first rising together during the 1920s and 1930s, and then falling together in the 1940s and 1950s. As the death penalty fell into disuse in the 1960s, the homicide rate rose sharply. The death penalty moratorium that began . . . in 1972 and ended . . . in 1976 appears to have been a period in which the homicide rate rose. The homicide rate then remained high and variable through the 1980s while the rate of executions rose. Finally, homicides dropped dramatically during the 1990s."

They then review some of the literature between 1975 and the publication of their article in 2006 on the statistical evidence regarding a possible deterrent effect of the death penalty, in particular discussing problems with an article from 2003 claiming that each death sentence has the effect of deterring, on average, 18 homicides.

They summarize:

"We have surveyed data on the time series of executions and homicides in the United States, compared the United States with Canada, compared non-death penalty states with executing states, analyzed the effects of the judicial experiments provided by the Furman and Gregg decisions comparing affected states with unaffected states, surveyed the state panel data since 1934, assessed a range of instrumental variables approaches, and analyzed two recent state-specific execution moratoria. None of these approaches suggested that the death penalty has large effects on the murder rate. Year-to-year movements in homicide rates are large, and the effects of even major changes in execution policy are barely detectable. Inferences of substantial deterrent effects made by authors examining specific samples appear not to be robust in larger samples; inferences based on specific functional forms appear not to be robust to alternative functional forms; inferences made without reference to a comparison group appear only to reflect broader societal trends and do not hold up when compared with appropriate control groups; inferences based on specific sets of controls turn out not to be robust to alternative sets of controls; and inferences of robust effects based on either faulty instruments or underestimated standard errors are also found wanting."

Death-penalty deterrence is a difficult topic to study. The treatment is observational, the data and the effect itself are aggregate, and changes in death-penalty policies are associated with other policy changes. Much of the discussion of the deterrence studies reminds us of a little-known statistical principle, which is that statisticians (or, more generally, data analysts) look best when they are studying large, clear effects. This is a messy problem, and nobody is going to come out of it looking so great.

More specifically, a quick analysis of the data, at least since 1960, finds that homicide rates went up when the death penalty went away, and then homicide rates declined when the death penalty was re-instituted, and similar patterns have happened within states. So it's not a surprise that regression analyses have found a deterrent effect. But, as noted, the difficulties arise because of the observational nature of the treatment, and that other policies are changed along with the death penalty. There are also various technical concerns that Donohue and Wolfers discuss.

Policy questions about the death penalty have sometimes been expressed in terms of the number of lives lost or saved by a given sentencing policy. But this direction of thinking might be a dead end. First, as noted above, it may very well be essentially impossible to statistically estimate the net deterrent effect of death sentencing—what seem like hard numbers or “careful econometric analysis” aren't so clear at all.

More generally, though, it's not clear how one would balance the chance of deterring murders with the chance of executing an innocent person. What if each death sentence deterred 0.1 murder, and 5% of people executed were actually innocent? That's still a 2-to-1 ratio (assuming that you're fine with executing the guilty people), so the death penalty is saving lives. Then again, maybe these innocent people who were executed weren't so innocent after all. But then again, not every murder victim is innocent either. Conversely, suppose that executing an innocent person were to deter 2 murders (or, conversely, that freeing an innocently-convicted man were to un-deter 2 murders). Then the utility calculus would suggest that it's actually fine to execute the innocent. In general we are sympathetic to probabilistic cost-benefit analyses, but here we don't see it working out. The main worries—on one hand, concern about out-of-control crime, and on the other hand, concern about executing innocents—seem too difficult to put on the same scale.

Finally, regarding decision analysis, incentives, and so forth: much of the discussion (not in the Donohue and Wolfers paper, but elsewhere) seems to go to the incentives of potential murderers. But the death penalty also affects the incentives of judges, juries, prosecutors, and so forth. One of the arguments in favor of the death penalty is that it sends a message that the justice system is serious about prosecuting murders. This message is sent to the population at large, not just to deter potential murderers but to make clear that the system works. Conversely, one argument against the death penalty is that it motivates prosecutors to go after innocent people, and to hide or deny exculpatory evidence. Lots of incentives out there.

This story is relevant to the week's reading as a demonstration of the instability of instrumental variables estimates and an example where a causal effect of interest cannot be estimated in any good way from available data. It relates to the course as a whole with its opaque connection between statistical modeling and social science theory. As Donohue and Wolfers put it,

“By any measure, the resumption of the death penalty in recent decades has been fairly minor, and both the level of the execution rate and its year-to-year changes are tiny: since 1960 the proportion of homicides resulting in execution ranged from 0% to 3%. By contrast, there was much greater variation in execution rates over the previous sixty years, when the execution rate ranged from 2.5% to 18%. This immediately hints that—even with modern econometric methods—it is unlikely that the last few decades generated enough variation in execution rates to overturn earlier conclusions about the deterrent effect of capital punishment.”

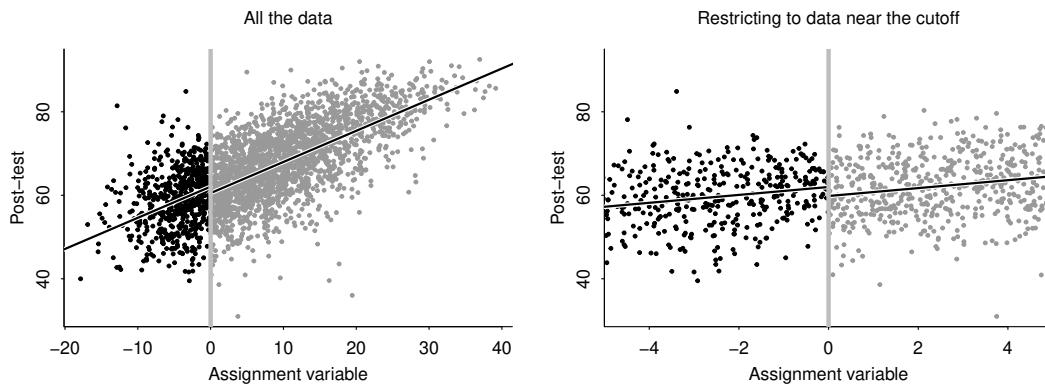
This story relates to the week's reading as an example of the difficulty of causal inference from observational data and, in particular, the instability of estimates from instrumental variables.

## 2. Regression discontinuity mishaps

Section 21.3 of *Regression and Other Stories* discusses regression discontinuity, an approach to estimating a causal effect in a setting where the treatment  $z$  is assigned based on a single predictor  $x$ , with  $z = 1$  if  $x$  exceeds some threshold (the “discontinuity”) or  $z = 0$  otherwise. The idea is to fit a regression of  $y$  given  $z$ ,  $x$ , and any other important pre-treatment predictors, just using the data  $x$  in some zone near the discontinuity. The discontinuity design has the drawback of imbalance and zero overlap on  $x$  but the advantage that the assignment method is known, and sometimes it can be reasonable to consider the assignment as essentially random in the neighborhood of the discontinuity.

Different statistical methods have different characteristic modes of failure. For regression discontinuity, failure often arises when an unreasonable curve is fit to predict  $y$  from  $x$  and when other important pre-treatment predictors are not included in the model, so that the regression is not doing a good job of comparing similar units for treatment and controls.

Figure 114 from *Regression and Other Stories* shows an example of regression discontinuity analysis that we find reasonable. The model makes sense to us for theoretical reasons—it makes



**Figure 114** A regression discontinuity analysis that we find plausible, from Section 21.3 of Regression and Other Stories. Each dot on the graph is a school in Chile, and the assignment variable is school-level average score on a pre-test. (a) School-level data and regression lines predicting the outcome (reading test score in 1992) given the assignment variable. (b) Subset of data close to the cutoff, representing schools that had a reasonable chance of receiving or not receiving the assignment. For both graphs, black dots represent schools that were below the cutoff and gray dots indicate schools above the cutoff. Lines show the fitted regression model in each case with a discontinuity at the threshold.

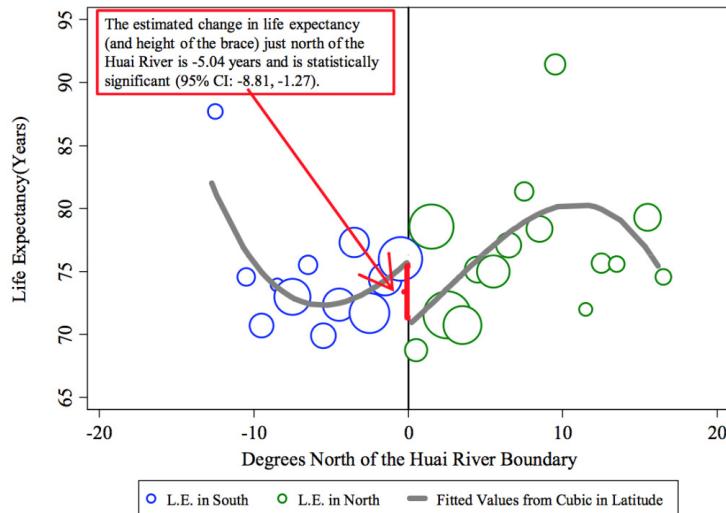
sense to compare schools that are similar in pre-test scores—and for the empirical reason that the estimated curve predicting  $E(y)$  given  $x$  seems plausible. The average treatment effect is small but can be estimated using a large number of schools.

Here we share various well-publicized examples from social and health sciences of discontinuity regressions that we did not find convincing. There are too many examples here to fit into the time allocated for discussion in one class period, so an instructor should feel free to choose just one or two to focus on.

Figure 115 shows an estimate of the effect of indoor air pollution on life expectancy.<sup>67</sup> Each circle in the graph represents a different location in China, and after the cubic curve is fit to the data there is a large discontinuity at zero, with lower life expectancies in the areas north of the river, where there had been a policy offering free indoor coal heating. This was taken as evidence of a causal effect.

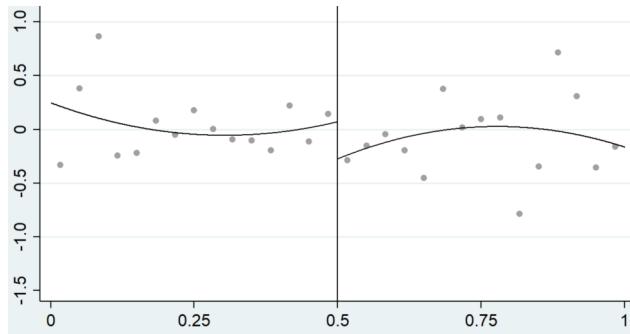
There are two clear problems with this conclusion. First, the drop at the discontinuity point arises only to counterbalance the steep increase in the fitted curve in that region of the graph. If you look at the data without the fitted curve, there is no evidence of a difference in life expectancy on the left and right sides of the boundary. And there is no scientific reason to think that this cubic polynomial makes sense: what is happening is that a curve is being overfit to the data and then overinterpreted. The second problem with this analysis is the data. Look at the green dot on the upper right portion of the graph. Is it really plausible that there is an area with a life expectancy of 92 years? Indeed, if you believe the fitted model, the life expectancy there would have been 97 in the absence of the coal heating policy. That high point, along with the implausible life expectancy of 88 for the blue circle on the upper left of the graph, appears to be driving the fitted curve. A conclusion is only as good as the data that go into it, and in this case we doubt the data.

<sup>67</sup>Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li (2013), Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy, *Proceedings of the National Academy of Sciences* 110, 12936–12941. The discussion here is taken from Andrew Gelman and Adam Zelizer (2015), Evidence on the deleterious impact of sustained use of polynomial regression on causal inference, *Research and Politics* 2, 1–7. See also the followup study, Avraham Ebenstein, Maoyong Fan, Michael Greenstone, and Maigeng Zhou (2017), New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy, *Proceedings of the National Academy of Sciences*, 114, 10384–10389.



**Fig. 3.** The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

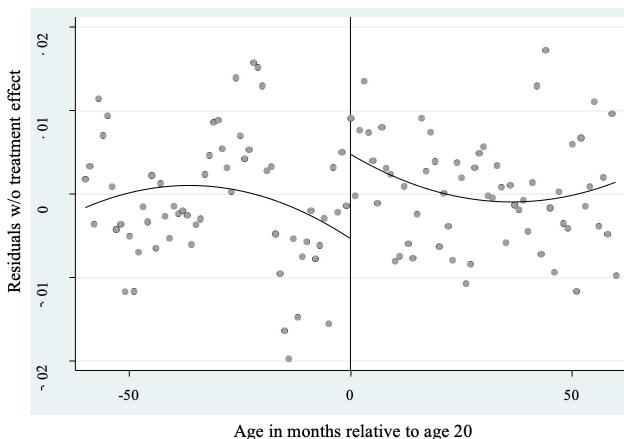
Figure 115 *Example of a regression discontinuity analysis we do not trust, purporting to show a large effect on life expectancy from an indoor coal-heating policy that was available in China only in areas north of the Huai River. Look at the dots without the curve, and you will see no evidence of a discontinuity across the border.*



**Figure 116** *Example of a regression discontinuity analysis we do not trust, purporting to show that unionization reduces the risk of a company's stock crashing. Data come from a set of companies that had union elections. The dots show average values of a standardized outcome relevant to the risk of a stock-price crash, plotted vs. the proportion of votes favoring unionization, and the curve shows a fitted quadratic curve with discontinuity. Look at the dots without the curve and you will see no evidence of a discontinuity across the border.*

For another example, Figure 116 summarizes an analysis purporting to find that adoption of labor unions reduces the risk of a company's stock price crashing.<sup>68</sup> As in the previous example, the apparent discontinuity cancels a fitted curve trending in the other direction, and if you look at

<sup>68</sup>Jeong-Bon Kim, Eliza Xia Zhang, and Kai Zhong (2021), Does unionization affect the manager-shareholder conflict? Evidence from firm-specific stock price crash risk, *Journal of Corporate Finance* 69, 101991. The discussion here is taken from Andrew Gelman (2019), Another regression discontinuity disaster and what can we learn from it, <https://statmodeling.stat.columbia.edu/2019/06/25/another-regression-discontinuity-disaster-and/>.



**Figure 117** Example of a regression discontinuity analysis we do not trust, purporting to show that the use of health care rises immediately before elections. The noisy fitted curves create the conditions for the discontinuity which otherwise does not appear in the data.

the data without the curve, no discontinuity appears at all. In addition, the curve itself makes no sense. One could fit such a curve and find a discontinuity just about anywhere in such data.<sup>69</sup>

Figure 117 shows another example, this time a study from Taiwan claiming to show that “elections increased health care use and expense only during legally specified campaign periods by as much as 19%.”<sup>70</sup> A curve is fit with a discontinuity at age 20, corresponding to the age at which people were eligible to vote. The curve on the left side of the discontinuity drops just enough to allow a big jump right next to it, but that did not stop the claim from being reported entirely uncritically in some economics and policy-related outlets.<sup>71</sup> There’s no discontinuity in the data, but it’s possible to see a discontinuity appear by fitting enough things to the two sides of the border.

For our next example, we point to a paper claiming to show a large benefit of air filters on student performance within a school district.<sup>72</sup> This purported effect was found by fitting a piecewise linear regression to data from 28 schools, predicting growth in average test scores given distance from gas leak, with a discontinuity at 5 miles, because air filters were installed only for schools within 5 miles of the site of a gas leak. Figure 2 of the research article shows the fitted piecewise linear model: it is approximately  $y = 0.10 + 0.11(x - 5)$  for schools with new air filters (those less than 5 miles from the leak:  $x < 5$ ) and  $y = -0.10 + 0.09(x - 5)$  for schools without the new air filters (further than 5 miles from the leak:  $x > 5$ ); thus the estimate of the effect at the discontinuity is 0.2, implying that the air filters caused an increase of test scores by an average of 0.2 standard deviations.

<sup>69</sup>For more systematic discussion of this point, see Andrew Gelman and Guido Imbens (2019), Why high-order polynomials should not be used in regression discontinuity designs, *Journal of Business and Economic Statistics* 37, 447–456. The title of that article notwithstanding, these problems also arise when fitting low-order polynomials or nonparametric curves.

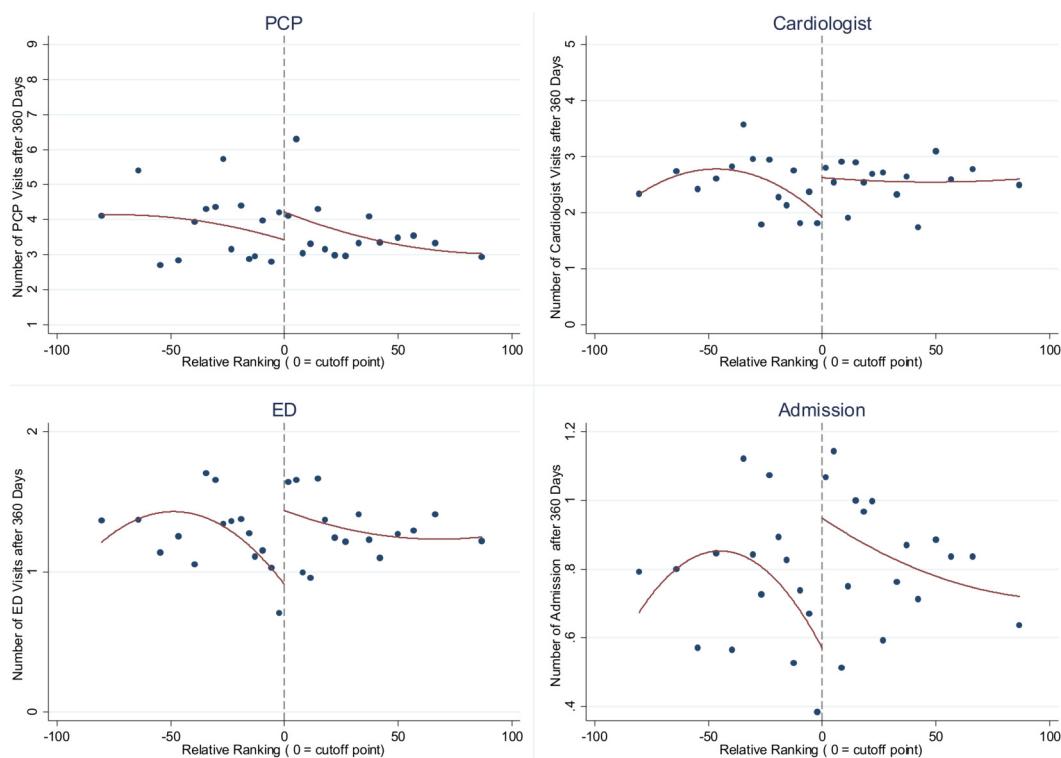
<sup>70</sup>Hung-Hao Chang and Chad Meyerhoefer (2023), Do elections make you sick? Evidence from first-time voters, *Health Economics* 32, 1064–1083. The discussion here is taken from Andrew Gelman (2020), Does regression discontinuity (or, more generally, causal identification + statistical significance) make you gullible?, <https://statmodeling.stat.columbia.edu/2020/12/27/rd-gullible/>.

<sup>71</sup>See, for example, Liz Scott (2021), Do elections make you sick?, *Chicago Policy Review*, <https://chicagopolicyreview.org/2021/03/15/do-elections-make-you-sick/>, Stephen Dubner (2020), Does the president matter as much as you think?, *Freakonomics Radio*, <https://freakonomics.com/podcast/does-the-president-matter-as-much-as-you-think-ep-404/>, Tyler Cowen (2020), Do elections make you sick?, <https://marginalrevolution.com/marginalrevolution/2020/01/do-elections-make-you-sick.html>, and Russell Lynch (2020), It’s official: elections are bad for your health, *Telegraph*, <https://www.telegraph.co.uk/business/2020/01/28/official-elections-bad-health/>.

<sup>72</sup>Michael Gilraine (2023), Air filters, pollution and student achievement, *Journal of Human Resources*. The discussion here is taken from Andrew Gelman (2020). No, I don’t think that this study offers good evidence that installing air filters in classrooms has surprisingly large educational benefits, <https://statmodeling.stat.columbia.edu/2020/01/09/filters-benefits/>.

#### 4.24. ADDITIONAL TOPICS IN CAUSAL INFERENCE

293



**Figure 118** Example of a regression discontinuity analysis we do not trust, estimating the effect of a care management intervention for high risk patients. The data show no clear patterns, but the fitted curves induce trends that are then countered by the fitted jumps.

We do not believe this result. Our problem is not with the general idea—yes, we’re skeptical of claims of large effects, but really we have no idea about what’s going on with air pollution and the brain—but, rather, what we don’t believe is that the study in question provides the claimed evidence.

The whole thing is driven by one data point and a linear trend of sharply increasing test scores as a function of distance from the leak, an effect that makes no theoretical sense in the context of the paper (from the abstract: “Air testing conducted inside schools during the leak (but before air filters were installed) showed no presence of natural gas pollutants, implying that the effectiveness of air filters came from removing common air pollutants”) but does serve to create a background trend to allow a big discontinuity with some statistical significance. As before, the steep slope of the fitted curve in the neighborhood of the discontinuity induces an opposite estimated effect at the threshold.

Given what we know about statistics, how should we think about this problem?

It goes like this: The installation of air filters can be considered as a natural experiment, so the first step is to compare outcomes in schools in the area with and without air filters: in statistics terms, a regression with one data point per school, with the outcome being average post-test score and the predictors being average pre-test score and an indicator for air filters. Make a scatterplot of post-test vs. pre-test with one point per school, displaying treated schools as open circles and control schools as dots. There’s nothing wrong with looking at the distance variable, but in this case the discontinuity analysis seems more like a distraction. There will be other differences between schools. In addition, it is difficult to interpret the reported statistical significance, given the many researcher degrees of freedom involved in the analysis, including what other variables

to adjust for and the functional forms used in the adjustments. At the end of the day, it's a cool natural experiment, but it might be just too small to give any useful information.

For a final example, see Figure 118, which shows various outcomes plotted vs. a healthcare utilization risk score.<sup>73</sup> The large discontinuities are artifacts of the wildly varying underlying curves being fit. Remove the curves and the discontinuities go away.

The point of all these examples is not that regression discontinuity analyses are always wrong or even that they deserve special scrutiny. Rather, the issue is that these sorts of observational studies or natural experiments require careful comparison of treated and control units that are similar in their pre-treatment characteristics. Pushing the regression-discontinuity button will not in general work, except in those cases where the selection variable is strongly predictive of the outcome and where the underlying fitted model makes sense.

This example relates to the week's reading in supplying examples of a method for causal inference. It is relevant to the course as a whole in demonstrating some ways that statistical analysis can go wrong.

### Class-participation activities

1. Gather, plot, and discuss two measurements of the same underlying quantity from students

Together, the class should come up with two ways of measuring a single characteristic (for example, political ideology, social class, interest in sports, . . . ). These will be displayed in scatterplots, so it is best if the measurements are continuous or approximately continuous. For example, a question on party identification (Democrat, Independent, Republican) will take on only three values, and this would be hard to display a scatterplot, so it would be better in that case to ask students for a continuous measure on a 0–100 scale, in which case it would be necessary to calibrate the scale in some way. One can also get a more continuous response by recording the sum of responses to a few questions.

Once the class has settled on the specific questions to ask, the instructor can put them on a Google form, and the students can use their phones or laptops to go to the form and enter their responses. Then the students should discuss what the scatterplot might look like. In pairs, students sketch guesses of the scatterplot. The instructor should go around the room to see what the students have done and then get a couple of students to go up to the board to sketch their guessed graphs.

Once the data have been gathered, the instructor can download the data from the form, and make a scatterplot in R that displays on the screen. Students should compare this plot of actual data to the guessed plots they sketched on the board and discuss differences.

This activity relates to the week's reading because questions about interpretation of the results inevitably lead into questions of the form, "Why" does the graph look like this? As discussed in Section 21.5 of *Regression and Other Stories*, these sorts of Why questions can be viewed as explorations of implicit models. For example, if students report political ideology and they are mostly on the liberal or left side of the spectrum, then the question, "Why is this?", is a comparison to an implicit model in which the distribution of ideology is closer to uniform, and we would seek an explanation for this difference. Or if you expect to see a strong relation between age and ideology, and you see only a weak pattern or none at all, then this again reflects a departure from your implicit model, which requires an explanation. The activity relates to the course as a whole as an example of the open-ended nature of empirical research; also it connects back to ideas of data collection and measurement in Chapter 2 of *Regression and Other Stories*.

<sup>73</sup>Guy David, Aaron Smith-McLallen, and Benjamin Ukert (2019), The effect of predictive analytics-driven interventions on healthcare utilization, *Journal of Health Economics* 64, 68–79, and Andrew Gelman (2021), Just another day at the sausage factory . . . It's just funny how regression discontinuity analyses routinely produce these ridiculous graphs and the authors and journals don't even seem to notice, <https://statmodeling.stat.columbia.edu/2021/11/21/sausage-notice/>.

## 2. “Why” questions and causal inference

Statistical methods focus on estimating treatment effects—what we call “forward causal inference,” but what about the “reverse causal problem” of seeing a pattern and trying to understand its cause? Section 21.5 of *Regression and Other Stories* considers what can be learned from Why questions.<sup>74</sup> A Why question points to an anomaly, some pattern in data that contradicts or cannot be explained by existing theories.

The idea of learning from anomalies is appealing, and it connects to the formulation in Chapter 2 of *Regression and Other Stories* considering statistical graphs as comparisons. Exploratory data analysis is the search for unexplained patterns in data, where “unexplained” is with reference to some set of implicit or explicit models representing one’s prior expectations about the data. Here we use the phrase “prior expectation” not in any formal Bayesian sense but just to acknowledge that exploration and surprise must be defined relative to some baseline of what would *not* be surprising. Every exploratory graph is in some sense a model check, and it can be valuable to consider what are the models being used as comparisons.<sup>75</sup>

To explore these ideas we go online and pull up a link to a graph from the *New York Times* showing county-level vaccine hesitancy estimates that came from a news report during the coronavirus epidemic.<sup>76</sup> We project onto the screen the map from that *New York Times* link. We ask students what they notice. Most clearly, there is variation across the country, with lower rates of willingness to take the vaccine in the southeast and higher rates in the west and northeast.

Also noticeable are several stark boundaries between states, such as Minnesota and its neighbors, or Colorado and Wyoming. Students should discuss in pairs where this pattern comes from. Before going on, we point out that this is an example of the use of statistical graphics to reveal unexpected patterns in data, and this motivates a consideration of what is expected. Is it reasonable to expect to see jumps across states in the estimated share of adults who say they would definitely or probably get the coronavirus vaccine?

One possible explanation for the discontinuities across states in the *New York Times* map is that there are underlying differences between neighboring states, for example having to do with policy or political differences. For example, perhaps people in Minnesota are much more likely than people across the border in Wisconsin to say they would get the vaccine because the vaccine was more readily available in Wisconsin. Or perhaps Wyoming residents are much more politically conservative than Coloradans living across the border.

Statistics student Luke Vrotsos looked into the problem:

“Immediately, it seems really implausible how big some of the state-border discontinuities are (like Colorado-Wyoming). I guess it’s possible that there’s really such a big difference, but if you check the 2020 election results, which are presumably pretty correlated with vaccine hesitancy, it doesn’t seem like there is. For example, estimated vaccine hesitancy for Moffat County, CO is 17% vs. 31% for neighboring Sweetwater County, WY, but Trump’s vote share was actually higher (81%) in Moffat County than in Sweetwater County (74%). . . . It’s strange to see results that seem so unlikely, just by looking at a map, reported so widely.”

<sup>74</sup>See also Andrew Gelman and Guido Imbens (2013), Why ask why? Forward causal inference and reverse causal questions, National Bureau of Economic Research working paper 19614, <https://www.nber.org/papers/w19614>.

<sup>75</sup>See Andrew Gelman (2003), A Bayesian formulation of exploratory data analysis and goodness-of-fit testing, *International Statistical Review* 71, 369–382.

<sup>76</sup>Apoorva Mandavilli (2021), Reaching “herd immunity” is unlikely in the U.S., experts now believe, *New York Times*, 3 May, <https://www.nytimes.com/2021/05/03/health/covid-herd-immunity-vaccine.html>. The discussion here is taken from Andrew Gelman (2021), Whassup with the weird state borders on this vaccine hesitancy map?, <https://statmodeling.stat.columbia.edu/2021/05/04/whassup-borders/>.

What happened? The data come from the U.S. government's Household Pulse Survey, which describes its methodology as follows:<sup>77</sup>

"Our statistical analysis occurred in two steps. First, using the HPS, we used a logistic regression to analyze predictors of vaccine hesitancy using the following sociodemographic and geographic information: age, gender, race/ethnicity, education, marital status, health insurance status, household income, state of residence, and interaction terms between race/ethnicity and having a college degree. Second, we applied the regression coefficients from the HPS analysis to the data from the ACS to predict hesitancy rates for each ACS respondent ages 18 and older. We then averaged the predicted values by the appropriate unit of geography, using the ACS survey weights, to develop area-specific estimates of hesitancy rates."

So they have no county-level data at all—they "averaged the predicted values," not the survey responses, within each geographic unit. The county differences in the map come entirely from demographic differences between counties. This still doesn't explain the large differences between states. We wouldn't be surprised to see some state-level effects, because policies vary by state and the political overtones of vaccines can vary by state, but the border effects just look too large and too consistent here. Perhaps part of the problem here is that they are using health insurance status as a predictor, and maybe that varies a lot from state to state.

The Household Pulse Survey had 80 000 respondents in week 26 (the source of the data used to make the map being discussed).<sup>78</sup> 80 000 is a lot! Not big enough to get good estimates for all the 3000 counties in the U.S., but big enough to get good estimates for subsets of states. For example, if we divide states into chunks of 200 000 people each, then we have, ummm,  $80\,000 * 200\,000 / (330 \text{ million}) = 48$  people per chunk. That would give us a raw standard error of  $0.5 / \sqrt{48} = 0.07$ , or 7 percentage points, per chunk, which is pretty big, but some regression modeling should help with that, and it still should be enough to improve certain things such as the North Dakota / Minnesota border.

On the other hand, the map is not a disaster. The reader of the map can realize that the state borders are artifacts, and that tells us something about the quality of the data and model. Any graph should contain the seeds of its own destruction—that is, provide enough granularity to reveal its flaws—and it's appealing, in a way, that this map shows the seams, in this case at the state borders.

This activity relates to the week's reading as an example of a Why question that pushes us to understand an anomaly in some published data. It relates to the course as a whole in that we are often in the habit of graphing and analyzing data, without always looking carefully for problems.

## Computer demonstrations

### 1. Instrumental variables estimation

Consider expected survival times after a heart attack (in years), depending on hospital quality and the administration of a special treatment. Hospital quality and the probability of offering the treatment can be correlated, and supposing we cannot observe hospital quality directly, a direct estimate of the effects of the treatment suffers from omitted variable bias.

One way to get around this is using an instrument, if some experiment or natural experiment

<sup>77</sup>U.S. Department of Health and Human Services (2021), ASPE predictions of vaccine hesitancy for COVID-19 vaccines by geographic and sociodemographic features, <https://aspe.hhs.gov/system/files/pdf/265341/vaccine-hesitancy-COVID-19-Methodology.pdf>.

<sup>78</sup>U.S. Census Bureau (2020), Source of the data and accuracy of the estimates for the Household Pulse Survey — Phase 3, [https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase3\\_Source\\_and\\_Accuracy\\_Week\\_26.pdf](https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase3_Source_and_Accuracy_Week_26.pdf).

is available. For example, suppose that special funding had been randomly provided to some hospitals, making it more likely that they would administer the treatment. We can then estimate the treatment effect using an instrumental variables analysis. In the code below we simulate the experimental assignment along with a model in which funding increases the probability of receiving the treatment:

```
n <- 1000
quality <- rnorm(n, 0, 1.5) # confounder
funding <- sample(c(0,1), n, replace=TRUE) # instrument
treatment <- rbinom(n, 1, invlogit(0.5 * quality + 1 * funding)) # treatment
survival <- 8 + 0.5 * quality + 3.0 * treatment + rnorm(n, 0, 1.5) # outcome
fake <- data.frame(funding, treatment, survival)
```

In this case, we are using an additive model for survival times, which does not make much sense given that this outcome is restricted to be positive. But that is not the focus of this example so we will continue.

Next you can perform the simple regression, which suffers from omitted variable bias:

```
fit1 <- stan_glm(survival ~ treatment, data=fake, refresh=0)
print(fit1)
```

You can then compare the estimate to the true treatment effect of 3.0 (see above code).

And here is the instrumental variables estimate:

```
itt_zt <- stan_glm(treatment ~ funding, data=fake, refresh=0)
itt_zy <- stan_glm(survival ~ funding, data=fake, refresh=0)
wald_est <- coef(itt_zy)[2] / coef(itt_zt)[2]
print(wald_est)
```

Again, check and compare to the true effect. You can also try embedding the whole thing in a loop to see the sampling distributions of the two estimates.

You can also experiment by re-running with different coefficients for funding in the simulation to make the instrument stronger or weaker.

## 2. Adjustment for pre-test score in a regression discontinuity analysis

Consider a world in which the length of the sentences that children speak (a proxy measure for cognitive development) is influenced by parents' income. To help less well-off families, society offers an intervention below a certain income threshold. You can create this world and then try to recover the effect using regression discontinuity. As the demonstration shows, we get more precise estimates when adjusting for pre-treatment variables.

```
# Pre-intervention world: 4-year-old children, sentences averaging 4.5 words
n <- 1000
income <- exp(rnorm(n, log(50), 0.3)) # in tens of thousands of dollars
words_pre <- 4.5 + 0.005*(income - median(income)) + rnorm(n, 0, 0.3)

# Post-intervention world: 1 year later, sentences averaging 1 word more
# ... for everybody, and 0.25 words more for those with intervention
rule <- income - 25
eligible <- ifelse(rule < 0, 1, 0)
words_post <- words_pre + 1 + 0.25*eligible + rnorm(n, 0, 0.5)

# Regression discontinuity
fake <- data.frame(income, words_pre, words_post, eligible)
fit1 <- stan_glm(words_post ~ eligible, data=fake,
```

```
subset=abs(rule) < 10, refresh=0)
print(fit1)

# Adjust for past word_count
fit2 <- stan_glm(words_post ~ eligible + words_pre, data=fake,
  subset=abs(rule) < 10, refresh=0)
print(fit2)
```

## Drills

### 1. Assumptions for instrumental variables estimation

Remind yourself of the assumptions for instrumental variables estimation, namely (1) ignorability of the instrument, (2) monotonicity and nonzero association between instrument and treatment variable, and (3) exclusion restriction. Then assess the following examples: how well do you think the assumptions hold?

- (a) You want to study the effect of education on earnings, using relative proximity to a four-year college as an instrument.<sup>79</sup>

*Solution:* (1) There could be nonignorability. The outcome is earnings, and it could be that people who live closer to a four-year college have higher earnings on average, for example. (2) Monotonicity and nonzero association seem plausible: we would expect that living closer to a college is associated with a higher probability of attending college. (3) The exclusion restriction seems like a problem. Students who are not affected by the instrument (those who would have gone to college, or not, regardless of their distance from the nearest college when growing up) could still have earnings affected by the instrument.

- (b) You want to study the effect of breast cancer screenings on survival rates, using randomly-sent invitations to breast cancer screenings as an instrument.<sup>80</sup>

- (c) You want to study the effect of the composition of municipal councils (balance of left- and right-leaning representatives) on education policy and total government spending. You use rainfall as an instrument, under the belief that potential left-wing voters are more likely to abstain from voting if it rains on election day.<sup>81</sup>

### 2. Regression discontinuity

Consider the following examples of discontinuity designs. Which would plausibly allow you to estimate a causal effect? What variables would you want to adjust for?

- (a) In France, municipalities with 3500 inhabitants or more have a proportional representation system, while smaller ones use a majoritarian system. A researcher finds a discontinuity in voter turnout at the threshold (higher turnout in proportional representation systems) and suggest this is a causal effect of going from one system to another.<sup>82</sup>

*Solution:* You would want to investigate if other differences occurring at that population threshold could provide alternative explanations for the discontinuity. In addition, in the analysis you would want to adjust for variables that could be predictive of voter turnout, such as region of the country and average socioeconomic status of the residents.

- (b) A special remedial educational program is given only to students for whom the pre-test score

<sup>79</sup>David Card (1993), Using geographic variation in college proximity to estimate the return to schooling, National Bureau of Economic Research working paper 4483, [https://davidcard.berkeley.edu/papers/geo\\_var\\_schooling.pdf](https://davidcard.berkeley.edu/papers/geo_var_schooling.pdf).

<sup>80</sup>Thad Dunning (2009), Instrumental variables, [http://www.thaddunning.com/wp-content/uploads/2009/12/Dunning\\_IEPS\\_InstrumentalVariables2.pdf](http://www.thaddunning.com/wp-content/uploads/2009/12/Dunning_IEPS_InstrumentalVariables2.pdf).

<sup>81</sup>Jo Thuri Lind (2020), Rainy day politics: An instrumental variables approach to the effect of parties on political outcomes, *European Journal of Political Economy* 61, 101821.

<sup>82</sup>Andrew Eggers (2013), Proportionality and turnout: Evidence from French municipalities, <https://www.bi.edu/globalassets/forskning/institutt-for-samfunnsokonomi/seminar-v13/eggers.pdf>.

is below some threshold. Students with and without the remedial program are then compared on scores from a post-test taken a year later.

- (c) A city sets up a historic preservation law protecting buildings that were constructed before 1940. Rents on buildings subject or not subject to the law are then compared, five years after the law has been passed.

### Discussion problems

#### 1. Estimating the effects of masks and social distancing

During the covid epidemic, people were not assigned at random to wear masks or to practice social distancing, but they were indirectly affected by national, state, and local policies mandating these actions.<sup>83</sup> Discuss how you might use instrumental variables to estimate these effects from available data on state-level policies, compliance, and outcomes. Then consider potential objections to such analyses, and finally discuss possible data that could be gathered to better estimate the effects of interest.

#### 2. Coaching for college admissions tests

Companies that provide coaching for college admissions tests advertise large gains from their programs. For example, Kaplan Educational Centers claimed benefits of 120 points on the SAT, and the Princeton Review claimed a score increase of 140 points.<sup>84</sup> Suppose these represent average gains among students who take the test, then do these programs, then take the test again. Explain why the average before-after difference is not a good estimate of the effect of the program. How could you estimate the magnitudes of the different biases involved when considering the before-after difference as an estimated causal effect?

Now suppose you were to estimate the effect of one of these coaching programs using a randomized experiment. How would you have to adjust this to obtain an estimate of the average treatment effect in the real world?

<sup>83</sup>See Andrew Gelman (2020), The point here is not the face masks; it's the impossibility of assumption-free causal inference when the different treatments are entangled in this way, <https://statmodeling.stat.columbia.edu/2020/06/17/face-masks/>.

<sup>84</sup>See Donald Powers and Donald Rock (1998), Effects of coaching on SAT I: Reasoning scores. College Board report 98-6, <https://files.eric.ed.gov/fulltext/ED562638.pdf>.

## 4.25 Advanced regression and multilevel models

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Nonlinearity in leafout dates	Nonlinear treatment effect	Modeling golf putting in Stan	Nonlinear models	20 data points and 16 predictors
Governors' elections and lifespans	When do students stop coming to class?	Opinions on same-sex marriage	Error terms for nonlinear models	Noisy time series

### Reading

Chapter 22 of *Regression and Other Stories*: Advanced regression and multilevel models

### Pre-class warmup assignments

1. Express models in a common framework
  - (a) Explain why  $y = \beta_0 + \beta_1 e^{-x} + \beta_2 e^{-2x} + \text{error}$  can be considered a linear model but  $y = \beta_0 + \beta_1 e^{-\beta_2 x} + \text{error}$  cannot.
  - (b) Explain what happens if you fit a regression model,  $y = a + bx + \text{error}$ , and  $y$  is itself measured with error.
  - (c) Explain what happens if you fit a regression model,  $y = a + bx + \text{error}$ , and  $x$  is itself measured with error.
2. Graph a nonlinear model
  - (a) Sketch the curve  $y = 1 - e^{-x}$  for positive values of  $x$  using pen on paper, without using the computer. Label the axes.
  - (b) Use R to graph the curve  $y = 1 - e^{-x}$  using a range of positive values of  $x$  that shows the interesting behavior of the curve.

### Homework assignments

1. (a) Measurement error in  $y$  (Exercise 22.1 of *Regression and Other Stories*)  
Simulate data  $(x, y)_i, i = 1, \dots, n$  from a linear regression model,  $y = a + bx + \text{error}$ , but suppose that the outcome  $y$  is not observed directly, but instead we observe  $v = y + \text{error}$ , with independent measurement errors with mean zero. Use simulations to understand the statistical properties of the observed-data regression of  $v$  on  $x$ , compared to the desired regression of  $y$  on  $x$ .  
(b) Measurement error in  $x$  (Exercise 22.2 of *Regression and Other Stories*)  
Simulate data  $(x, y)_i, i = 1, \dots, n$  from a linear regression model,  $y = a + bx + \text{error}$ , but suppose that the predictor  $x$  is not observed directly, but instead we observe  $u = x + \text{error}$ , with independent measurement errors with mean zero. Use simulations to understand the statistical properties of the observed-data regression of  $y$  on  $u$ , compared to the desired regression of  $y$  on  $x$ .
2. (a) Understanding nonlinear models
  - i. Sketch the curve,  $y = a(1 - e^{-bx})$  for positive values of  $x$  on a sheet of paper, without using the computer. Label  $a$  and  $b$  on the axes.

- ii. Sketch the curve,  $y = a + b \frac{1}{1+x/c}$  for positive values of  $x$  on a sheet of paper, without using the computer. Label the axes in terms of  $a$ ,  $b$ , and  $c$ .

(b) *In pairs:* Working through your own example (Exercise 21.18 of *Regression and Other Stories*)

Continuing the example from the final exercises of the earlier chapters, consider a causal problem that it would make sense to estimate using instrumental variables. Perform the instrumental variables estimate, compare to the estimated causal effect using direct regression on the treatment variable (in both cases including relevant pre-treatment predictors in your regressions), and discuss the different assumptions of these different approaches.

## Stories

### 1. Nonlinear patterns in leafout dates

In a collaboration with a group of biologists, we wrote,<sup>85</sup>

“Temperature sensitivity—the magnitude of a biological response per °C—is a fundamental concept across scientific disciplines, especially biology, where temperature determines the rate of many plant, animal and ecosystem processes. Recently, a growing body of literature in global change biology has found temperature sensitivities decline as temperatures rise. Such observations have been used to suggest climate change is reshaping biological processes, with major implications for forecasts of future change.”

Figure 119 shows an example of this literature, plotting trends over time in the temperature sensitivity of leaf unfolding for several species of trees.<sup>86</sup> For each species and for a series of 10-year moving windows, a linear regression was fit predicting the time of leaf unfolding given average temperature, and what is plotted in Figure 119 are the slopes from those regressions. These slopes show a general decline over time.

The decline in the regression slopes—the decrease in temperature sensitivity—has been found in many empirical studies, and there has been speculation of why this decline is happening and what it implies for the environment. It turns out, though, that the decline in slopes can be explained as an artifact of fitting linear models to nonlinear data.

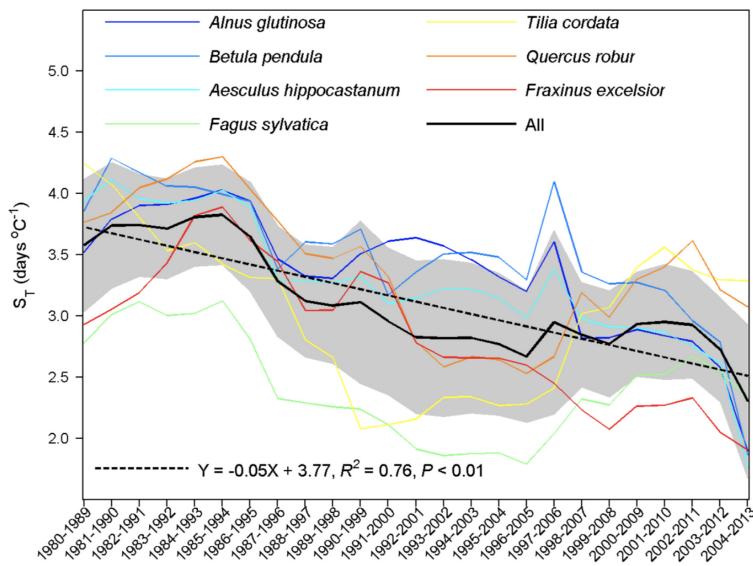
Figure 120a illustrates with a time series of budding of leaves as a function of temperature, for several different varieties of walnut trees.<sup>87</sup> Suppose these curves represented a persistent pattern over a long period but with average temperature gradually increasing over a period of decades. Then in the early part of the time series, you would see data with temperatures in the 5°–10° range, hence a high temperature sensitivity—a steep slope of budding time vs. temperature—and for the later part of the series, you would see data in the 20°–25° range, hence a low temperature sensitivity. Even if the underlying nonlinear curve is not changing at all over time, there will appear to be a regime change, because temperature sensitivity was defined based on a linear model.

We also demonstrated this behavior using a theoretical model, as shown in Figure 120b, where simulated data are divided into temperature windows to make it clear that, even when the big picture is nonlinear, data can look locally linear. This example also shows how the problem would not necessarily be resolved by checking the fit of the individual linear models. In this case, the temperature sensitivity—the local regression line—really is changing over time; the mistake is just to consider this as evidence for a change in the underlying curve.

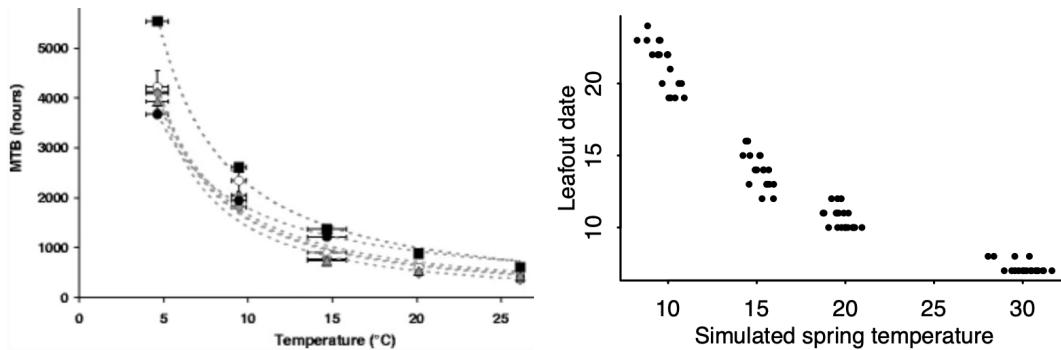
<sup>85</sup>Elizabeth Wolkovich, Jonathan Auerbach, Catherine Chamberlain, Daniel Buonaiuto, Ailene Ettinger, Ignacio Morales-Castilla, and Andrew Gelman (2021), A simple explanation for declining temperature sensitivity with warming, *Global Change Biology* 27, 4947–4949.

<sup>86</sup>From Yongshuo Fu et al. (2015), Declining global warming effects on the phenology of spring leaf unfolding, *Nature* 526, 104–107.

<sup>87</sup>From Guillaume Charrier, Marc Bonhomme, André Lacointe, and Thierry Améglio (2011), Are budburst dates, dormancy and cold acclimation in walnut trees (*Juglans regia* L.) under mainly genotypic or environmental control?, *International*



**Figure 119** For several species of trees, time trends of temperature sensitivities of leaf unfolding. This sort of pattern can be understood as an artifact of fitting a series of linear models to nonlinear data; see Figure 120.



**Figure 120** (a) Mean time until budbreak for six sorts of walnut trees, as a function of temperature. (b) Simulated data showing a nonlinear relation of leafout with temperature, showing data from different temperature windows.

This story is relevant to the week's reading as an example of the limitations of linear models. It relates to the course as a whole by demonstrating the sometimes subtle connections between the models we fit to data and the scientific conclusions that we draw.

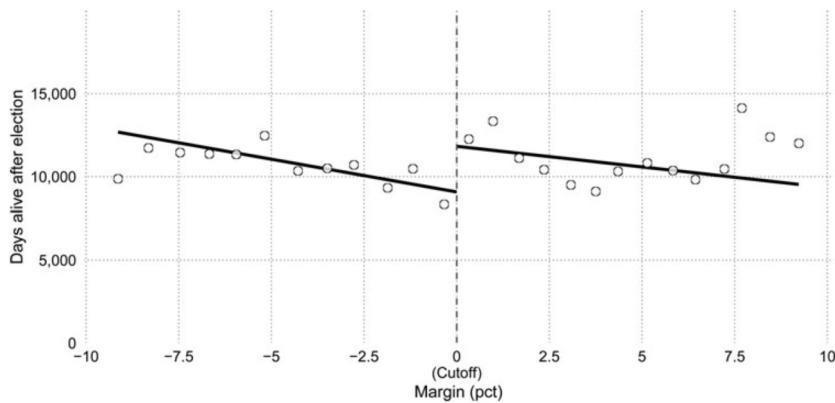
## 2. Nonlinear modeling for data exploration

Beginning on page 289, we discussed several examples of regression discontinuity analyses that we do not trust. Here we present another, this time using loess regression to explore the data.

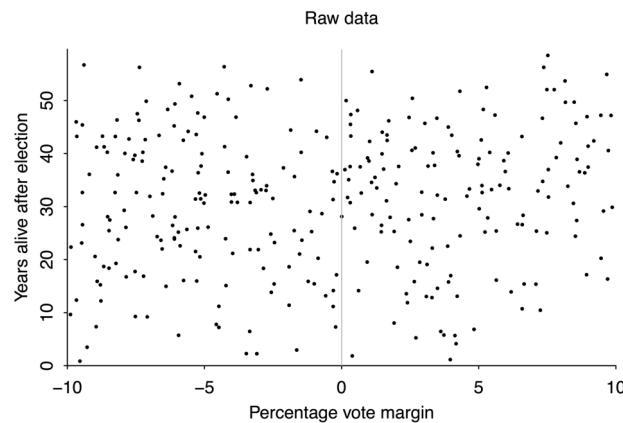
Figure 121 is from a research article claiming that winning an election causes an increase of 5 to 10 years of life.<sup>88</sup> The analysis was performed on a historical database including the years of birth and death of the winning and losing candidates in elections for governors of U.S. states. The graph, plotting average lifespan after the election vs. margin of victory, in binned averages, shows

*Journal of Biometeorology* 55, 763–774.

<sup>88</sup>Sebastian Barfort, Robert Klemmensen, and Erik Gahner Larsen (2021), Longevity returns to political office, *Political Science Research and Methods* 9, 658–664. The discussion here is taken from Andrew Gelman (2022), Criticism as asynchronous collaboration: An example from social science research, *Stat* 11, e464.



**Figure 121** Example of a regression discontinuity analysis we do not trust, purporting to show a large effect on life expectancy from winning or losing an election for governor. Each dot represents an average of data points. The pattern looks impressive, but upon careful reflection the fitted model makes no sense. After removal of duplicate cases from the data, the estimate can be explained as an artifact of variation and is consistent with a null effect.



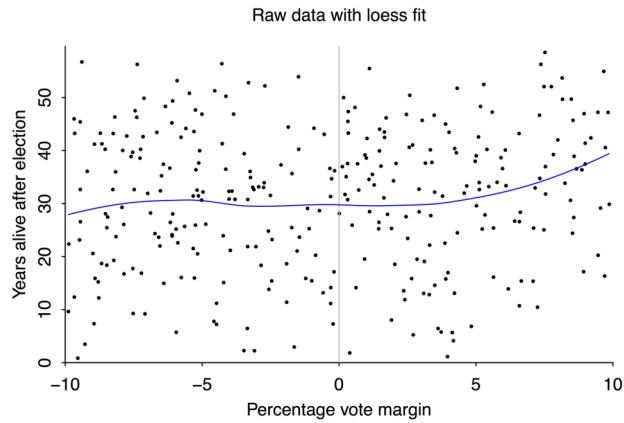
**Figure 122** Raw data from the governors' elections and lifespans analysis. Compare to the binned averages in Figure 121.

a fitted line with a discontinuity at zero. However, that jump at zero can be seen as an artifact of the negative slope of the line on each side of the cutoff.

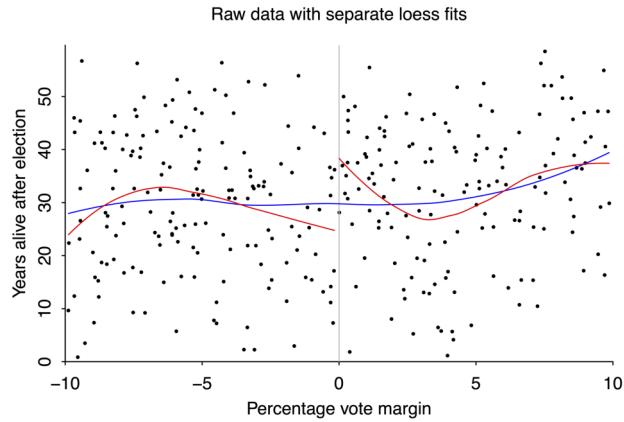
However, that line itself does not make sense, as it implies that a candidate's life expectancy is:

- 30 years if he loses an election by 5 percentage points
- 25 years if he loses narrowly
- 35 years if he wins narrowly
- 30 years if he wins by 5 percentage points.

The problem here is not just with the discontinuity but with the entire model—and, crucially, the discontinuity can only be interpreted in the context of the fitted line. It's a lot more believable that this variation is just noise, some artifact of the few hundred cases in this dataset, than that it represents some general truth about elections, or even about elections for governor.



**Figure 123** Data from the governors’ elections and lifespans analysis along with a fitted local linear smoother.



**Figure 124** Data from the governors’ elections and lifespans analysis along with the fitted local linear smoother (continuous curve) and two separate locally linear smoothers fit separately to the data where the candidates lost and won the election (broken curves). A discontinuity at  $x = 0$  counteracts a strong and implausible negative relation between vote margin and lifespan near the boundary.

Figure 121 might at first look compelling, but now consider Figure 122, which shows the raw, unbinned data. For ease of interpretation we have plotted the data in years rather than days.

Next we throw a local linear smoother (loess) on there and see what we get; this is shown in Figure 123. Figure 124 shows two additional loess fits, one for negative  $x$  and one for positive  $x$ , at which point a discontinuity opens up. We’ve seen this before with discontinuity regressions: when the data are relatively flat near the break point, but the fitted continuous curve from the regression discontinuity model happens to have a steep negative slope right near the discontinuity, so that a positive jump is needed to counter an essentially random pattern in the overfitted curves.

Further analysis finds the discontinuity estimate to be very sensitive to the details of the model that is fit. In particular, the graphs show remaining years of life, which has a very high negative correlation with the candidate’s age at the time of the election.

The purpose of this particular story is not to rehash all the problems with sloppy statistical models; rather, the relevance to the week’s reading is that fitting a nonlinear model can give us intuition about problems with a different fit—even if the nonlinear model does not make sense either. The example relates to the course as a whole by illustrating the importance of interpreting models, not just fitting them.

### Class-participation activities

#### 1. Nonlinear treatment effect

It is usual in regression models to assume a constant treatment effect (that is, a coefficient for the treatment indicator) or an effect that varies linearly with some predictors (that is, coefficients for interactions of the treatment indicator and other predictors in the model). But real effects can show strong nonlinear dependence on predictors.

For example, consider a medical treatment that has no effect for the healthiest patients (because they do not need the treatment) or for the sickest (for whom the treatment is too late). This corresponds to a nonlinear treatment effect that starts at zero, rises in a “sweet spot” somewhere in the middle of the range, and declines back to zero. A similar example would be an educational innovation that benefits students in the middle but not the least-prepared students (who do not know enough to make use of the new teaching approach) and not the best-prepared students (because they already understand the material well enough not to need the help).

In this demonstration, students are set up to simulate this scenario. First, the class should together come up with a story, similar to the medical or educational example above, with a pre-test measurement  $x$ , a treatment effect that is a non-monotonic function of  $x$ , and potential outcomes  $y^0, y^1$  under the control and treatment. Students should work in pairs to sketch the expected value of  $y^0$  and  $y^1$  given  $x$  and try to come up with mathematical formulas to approximate these curves. The instructor can walk around the room and monitor progress. After a few minutes the instructor can graph formulas that show the desired behavior and project these onto the screen.

The next step is to assign data to each student. The pre-treatment measurement  $x$  can be drawn from a uniform distribution representing the range of interest of this variable. The treatment  $z$  can be assigned completely at random, or using some more complicated approach such as people with higher values of  $x$  being more likely to get the treatment. Finally, the outcome  $y$  is computed using the given formula. For simplicity, we recommend using the deterministic model here, that is adding no random error to  $y$ .

At this point, we demonstrate the challenge of nonlinear effects by estimating the average treatment effect using a linear model fit to different subsets of the data: just students with low values of  $x$ , just students with high values of  $x$ , students with intermediate values, all students. Fitting to different subsets will give different results. With the correct nonlinear model, all should be fine, but in general such a model can be difficult to construct.

This activity connects to the week’s reading by being an example of nonlinear modeling. It relates to the course as a whole in connecting modeling errors to estimation problems.

#### 2. When do students quit a class?

It is usual for some students to decide to drop a course as the semester goes on. You can study this with your own statistics class with an activity that starts at the beginning of the semester, follows up a few weeks later, and ends with a final analysis.

The first step is for students in pairs to predict attendance for the next class, or the percentage of homework assignments that will be turned in on time. Students can first write down their guesses and then discuss how to evaluate these with future data. If we assume that there will be no new students joining the class, this can be framed as a binomial model with some probability that a student who had registered for the course will attend also the next class or return a homework assignment. If the activity is continued over the course, the previous weeks’ predictions can be compared to actual events, and students can make new predictions each week. Students can discuss in pairs how they can improve their predictions given information from the previous weeks, and whether the binomial model assumption makes sense.

At the end of the course, the instructor can graph trends in past retention and students can make

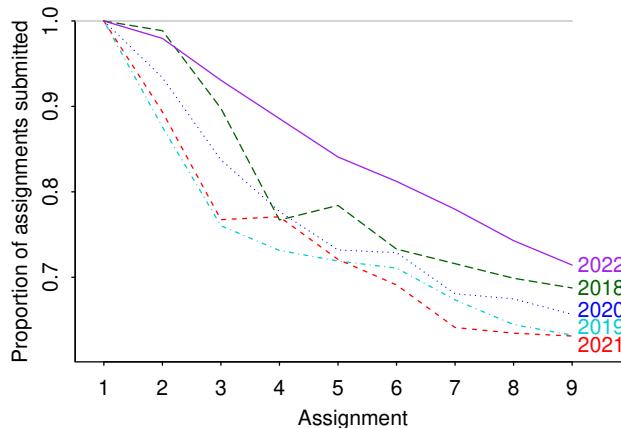


Figure 125 *Proportion of each assignment handed in as the semester went on, for five different years that a course was offered. In each year, participation declined, but in the most recent year the decline in participation had reduced. This is an example of data that would be best fit by a flexible nonlinear model.*

their guesses given those trends and then discuss how to fit a nonlinear model to make a prediction with uncertainty.

For example, Figure 125 shows data from a class we taught at Aalto University with nine homework assignments. The course was not required, and it is common practice for students to sign up for extra courses and then later in the semester decide which to stick with. Some of the students who drop out eventually finish the course in a later semester, but that is not shown here. We are interested in whether student retention is similar in different years, as we may hope that changes in the course arrangements would improve the retention.

This activity relates to the week's reading as a motivation for flexible nonlinear modeling. Here is an example of code to fit a spline model to the data in R and Stan:

```
fit6 <- brm(y | trials(n) ~ s(t, k=4) + (t | year),
  family=binomial(), data=retention)
```

Here, the data frame `retention` has 45 rows corresponding to 9 assignments in each of 5 years, `y` is the number of students who turned in assignments in a given assignment number and year, `n` is the number of students who began the class that year, `t` is the assignment number, `year` is the year number, and we are fitting a spline with 4 tuning parameters (that is the `k=4` in the code), along with a linear term that varies by year.

## Computer demonstrations

### 1. Fitting golf putting models in Stan

Section 22.6 of *Regression and Other Stories* presents models for the probability of getting a golf ball into a hole as a function of distance. Here we demonstrate how to fit a logistic model using `rstanarm` and use Stan to fit a geometry-based model. The model can be improved further, beyond what we show here, also accounting for how hard the ball is hit.<sup>89</sup>

```
# Setup
library("rstanarm")
```

<sup>89</sup>See Section 10 of Andrew Gelman et al. (2020), Bayesian workflow, [http://www.stat.columbia.edu/~gelman/research/unpublished/Bayesian\\_Workflow\\_article.pdf](http://www.stat.columbia.edu/~gelman/research/unpublished/Bayesian_Workflow_article.pdf), and Andrew Gelman (2022), Hierarchical model golf putting success!, <https://statmodeling.stat.columbia.edu/2022/02/20/golf-success/>.

```
library("cmdstanr")
golf <- read.table(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Golf/data/",
  "golf.txt"),
  header=TRUE, skip=2)

# Logistic regression
fit1 <- stan_glm(cbind(y, n-y) ~ x, family=binomial(link="logit"), data=golf,
  refresh=0)
plot(golf$x, golf$y/golf$n, ylim=c(0,1))
curve(invlogit(coef(fit1)[1] + coef(fit1)[2] * x), add=TRUE)

# Fit logistic regression using Stan
golf_data <- list(x=golf$x, y=y, n=n, J=J)
golf_logistic <- cmdstan_model("golf_logistic.stan")
fit_logistic <- golf_logistic$sample(data=golf_data, refresh=0)
print(fit_logistic)

# Geometry-based nonlinear model using Stan
r <- (1.68/2)/12
R <- (4.25/2)/12
golf_angle<- cmdstan_model("golf_angle.stan")
fit_angle <- golf_angle$sample(data=golf_data, refresh=0)
print(fit_angle)

# Plot data and both fits

draws_angle <- fit_angle$draws(format="df")
sigma_sim <- draws_angle$sigma
sigma_hat <- median(sigma_sim)
par(mar=c(3,3,2,1), mgp=c(1.7,.5,0), tck=-.02)
plot(0, 0, xlim=c(0, 1.1*max(x)), ylim=c(0, 1.02), xaxs="i", yaxs="i", bty="l",
  xlab="Distance from hole (feet)", ylab="Probability of success",
  main="Two models fit to the golf putting data", type="n")
segments(x, y/n + se, x, y/n-se, lwd=.5)
curve(invlogit(a_hat + b_hat*x), from=0, to=1.1*max(x), add=TRUE)
x_grid <- seq(R-r, 1.1*max(x), .01)
p_grid <- 2*pnorm(asin((R-r)/x_grid) / sigma_hat) - 1
lines(c(0, R-r, x_grid), c(1, 1, p_grid), col="blue")
points(x, y/n, pch=20, col="blue")
```

## 2. Nonlinear regression for opinions on same-sex marriage

Section 22.7 of *Regression and Other Stories* presents several nonlinear models fit to data on the support for same-sex marriage given age. This demonstration shows calculations for one of the two questions demonstrated in the textbook. We fit and display two different nonlinear models, locally weighted regression using the loess function in R, and spline regression using stan\_gamm4. We collapse the data, which is not essential for the estimation but facilitates plotting the data along with the fitted models.

```
# Setup
library("rstanarm")
marr <- read.csv(paste0(
  "https://raw.githubusercontent.com/avehtari/",
  "ROS-Examples/master/Gay/data/",
  "naes04.csv")
```

```
)
marr <- marr[!is.na(marr[, "age"]) & !is.na(marr[, "gayFavorStateMarriage"])]
marr$age[marr$age>90] <- 91
y <- as.character(marr[, "gayFavorStateMarriage"])
marr$y <- ifelse(y=="Yes", 1, ifelse(y=="No", 0, NA))

# Collapse data
uniq_age <- sort(unique(marr$age))
n_age <- length(uniq_age)
tab <- table(marr$age, marr$y)
y_sum <- as.vector(tab[,2])
n_sum <- as.vector(tab[,1] + tab[,2])
marr_sum <- data.frame(n=n_sum, y=y_sum, age=uniq_age)

# Loess
marr_loess <- loess(y ~ age, data=marr)
marr_loess_fit <- predict(marr_loess, data.frame(age=marr_sum$age))

# Spline
marr_spline <- stan_gamm4(I(y/n) ~ s(age), data=marr_sum)
marr_spline_fit <- posterior_linpred(marr_spline, data.frame(age=marr_sum$age))

# Plot
uniq_age <- sort(unique(marr$age))
plot(marr_sum$age, marr_sum$y/marr_sum$n, ylim=c(0.1, 0.6)) # data
lines(uniq_age, marr_loess_fit, col="blue", lwd=3) # loess
n_sims <- nrow(marr_spline_fit) # spline
for (i in sample(n_sims, 20)){
  lines(marr_sum$age, marr_spline_fit[i,], lwd=.5, col="gray50")
}
lines(marr_sum$age, colMeans(marr_spline_fit), lwd=2)
```

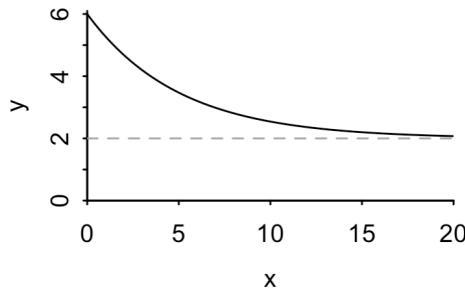
## Drills

### 1. Nonlinear models

Each of these functions is defined for positive values of  $x$ . Sketch each function using pen on paper, labeling the axes appropriately.

(a)  $y = 2 + 4e^{-0.2x}$

*Solution:*



(b)  $y = 4 \frac{x+2}{x+10}$

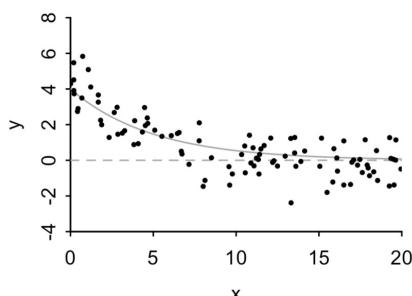
(c)  $y = -2x^{-1.5}$

### 2. Error terms for nonlinear models

Using pen on paper, sketch scatterplots of each of the following models, plotting roughly 100 points with  $x$  uniformly distributed between 0 and 20.

- (a)  $y = 4e^{-0.2x} + \text{error}$ , with independent errors normally distributed with mean 0 and standard deviation 1

*Solution:*



- (b)  $y = 4e^{-0.2x} + \text{error}$ , with independent normally-distributed errors with mean 0 and standard deviation 0.1  
(c)  $y = 4e^{-0.2x} * \text{error}$ , with independent *lognormally*-distributed errors with mean 0 and standard deviation 1 on the log scale  
(d)  $y = 4e^{-0.2x} * \text{error}$ , with independent *lognormally*-distributed errors with mean 0 and standard deviation 0.1 on the log scale

### Discussion problems

#### 1. Regression with 21 data points and 16 predictors

Many years ago we taught a course in statistical consulting. The consulting was done by graduate students in pairs: each pair had open office hours once a week, clients would come in to discuss their problems and then were told to return in a week, then each week we would have a meeting with all the students where we would go over the consulting problems that had come in, which would prepare them for their followup meetings.

Lots of interesting problems would come in. One week, a pair of students reported that someone had shown up who was studying the efficiency of industrial plants. The researcher had data on 21 factories, and for each of them she had a measure of efficiency and 16 predictors—different variables that might be predictive of that outcome. She wanted to use these data to see which of these factors was most important. We’re sorry, but we have no records from this class, so many details are missing—we’re reconstructing this from memory.

But one thing we do remember are the numbers: 21 data points, 16 predictors.

The first problem here is to ask what can be done with these data. It’s not an easy question. Indeed, it might seem ridiculous to suppose that you could tease out a regression relationship among so many predictors with so few observations. And this is without even getting into potential interactions (16 \* 15/2 two-way interactions and so forth) or the difficulties of causal identification from observational data. If students cannot come up with any ideas, the instructor should push them in another way, by asking what decisions they might make based on these data, if they were designing this sort of industrial plant.

When this example came up in our consulting class years ago, one of the other students said that he remembered that researcher from the previous semester: she’d come by with 15 data points and 16 predictors, and he and his partner had told her that, with fewer data points than predictors, they couldn’t help her. In the meantime this researcher had gathered data from 6 more plants and was emboldened to return.

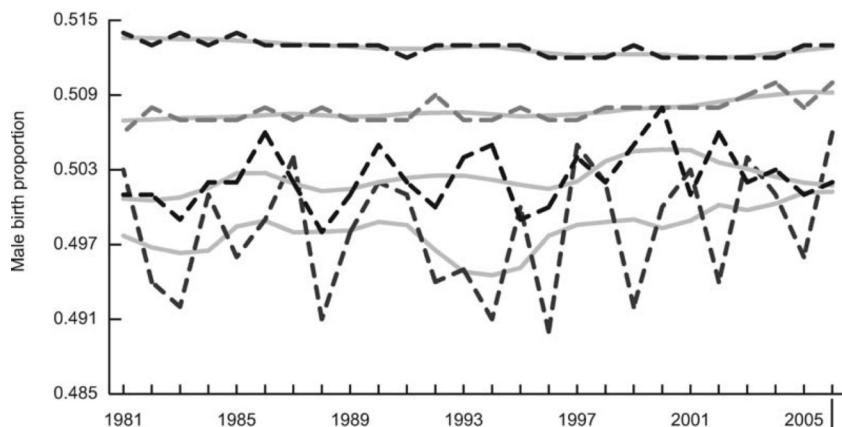


Figure 126 *Proportion of boy births over time in the United States, among white single births, white multiple births (twins, etc.), black single births, and black multiple births. The solid curves are smoothed versions for each group. The first challenge is to identify which of the four lines is which. The class can then discuss how to model and assess the visible trends and fluctuations.*

Fine. Laugh all you want. But . . . there are things that can be done even using this small dataset. Think about it this way: suppose the researcher had come in with 21 factories and just one predictor. Then you could do something, right? You can make a scatterplot of the outcome vs. the predictor, you could run a regression predicting the outcome from this one variable. You can potentially learn a lot from 21 data points, or even from 15. Even if all you learn is that none of the 16 available predictors is by itself a strong indicator of the outcome, that still is relevant information.

This example relates to the week’s reading because it is a problem with a large number of predictors compared to the number of data points, which is a setting where classical least-squares regression will not work, and some sort of regularization would be necessary, as is done in various Bayesian or machine learning approaches to statistics. It is relevant to the course as a whole as an example where if we think carefully about our inferential goals, we realize we can learn something useful from our data—just not the “statistically significant” comparison that we’re used to looking for.

## 2. Noisy time series

Figure 126 shows trends in the sex ratio for single and multiple births among white and black mothers.<sup>90</sup> The first step of this discussion problem is to figure out which of the four series is which. The solution is to look at variability: the larger the sample size, the less noisy the time series should look. More white babies than black babies are born each year, and there are many more singletons than twin births. The ranking of the four time series from least to most noisy should thus be white singletons, black singletons, white multiples, and black multiples, which happens to be the ordering of the four series in the graph, from upper to lower.

The next step is to discuss how these time series could be modeled, and to what extent you think some of the fluctuations in the curves are meaningful. The changes from year to year give some sense of what to believe. The long-term trend of increasing proportion of boy births among black mothers seems like a consistent pattern. At the other extreme, the big year-to-year fluctuations in sex ratio among black multiple births look like noise. A careful statistical analysis of these series would have to use the  $0.5/\sqrt{n}$  standard deviation of proportions that would arise from pure randomness, as empirical studies have found that the sexes of births are statistically independent (with the exception that identical twins must be of the same sex).

<sup>90</sup>Amy Branum, Jennifer Parker, and Kenneth Schoendorf (2009), Trends in US sex ratio by plurality, gestational age and race/ethnicity, *Human Reproduction* 24, 2936–2944.

## 4.26 Review of the course

### Plan for two classes

Stories	Activities	Computer demonstrations	Drills	Discussion problems
Randomized trials in international development	Designing a paper helicopter	Quadratic regression and regression	Basic statistics	Design using simulation
Is North Carolina less democratic than North Korea?	Review in groups	Bias and unmodeled uncertainty	Logistic regression and causality	Electoral integrity index

### Reading

Review all of *Regression and Other Stories*

### Pre-class warmup assignments

1. Basic statistics and linear regression
  - (a) Give the standard error corresponding to an estimated proportion of 0.3 from a simple random sample of 1000 people.
  - (b) You are doing an experiment feeding rats, and in the range of your data you find out that if you give a rat 1% more food, its weight will increase by 0.8% on average. Write the mathematical form of this model, where  $x$  is the amount of food given to the rat and  $y$  is its weight.
2. From logistic regression through causal inference
  - (a) A logistic regression model is fit,  $\Pr(y = 1) = \text{logit}^{-1}(a + 0.2x)$ , predicting whether a survey respondent plans to vote for the conservative party in the upcoming election, given an issue attitude  $x$  that is on a 0–10 scale. What is a reasonable value for the intercept,  $a$ ?
  - (b) Suppose a fitted regression model is  $y = 0.1 + 0.2x + 0.3z + 0.4xz + \text{error}$ , where  $x$  is a pre-treatment measurement,  $z$  is the treatment, and  $y$  is the outcome of interest. What is the estimated population average causal effect?

### Homework assignments

1. *In pairs:* Working through your own example

Write a few paragraphs summarizing what you have learned about your example from all the analyses you have done on it during the semester.

2. *No assignment due last day of class.*

### Stories

1. The rise and fall and rise of randomized controlled trials in international development

*Regression and Other Stories* follows the standard presentation of causal inference in which randomized experimentation is the baseline. In many areas of social science, though, observational

<sup>9</sup>Luciana de Souza Leão and Gil Eyal (2019), The rise of randomized controlled trials (RCTs) in international development in historical perspective, *Theory and Society* 48, 383–418. The discussion here is taken from Andrew Gelman (2020), The rise and fall and rise of randomized controlled trials (RCTs) in international development, <https://statmodeling.stat.columbia.edu/2020/11/14/rise-fall-rcts/>.

studies are the norm. Sociologists Luciana de Souza Leão and Gil Eyal discuss the history of randomized experiments in studies of international development:<sup>91</sup>

“Although the buzz around RCT [randomized controlled trial] evaluations dates from the 2000s, . . . what we are witnessing now is a second wave of RCTs, while a first wave began in the 1960s and ended by the early 1980s.”

What were the key differences between the two waves? Leão and Eyal start with the most available explanation:

“What could explain the rise of RCTs in international development? Randomistas tend to present it as due to the intrinsic merits of their method, its ability to produce ‘hard’ evidence as compared with the ‘softer’ evidence provided by case studies or regressions. They compare development RCTs to clinical trials in medicine, implying that their success is due to the same ‘gold standard’ status in the hierarchy of evidence.”

They quote an economist who wrote, “It’s not the Middle Ages anymore, it’s the 21st century . . . RCTs have revolutionized medicine by allowing us to distinguish between drugs that work and drugs that don’t work. And you can do the same randomized controlled trial for social policy.”

Leão and Eyal offer a slightly different perspective:

“While the buzz around RCTs certainly dates from the 2000s, the assumption—implicit in both the randomistas’ and their critics’ accounts—that the experimental approach is new to the field of international development—is wrong. In reality, we are witnessing now a second wave of RCTs in international development, while a first wave of experiments in family planning, public health, and education in developing countries began in the 1960s and ended by the early 1980s. In between the two periods, development programs were evaluated by other means.”

They then set up the puzzle:

“Instead of asking, ‘why are RCTs increasing now?’ we ask, ‘why didn’t RCTs spread to the same extent in the 1970s, and why were they discontinued?’ In other words, how we explain the success of the second wave must be consistent with how we explain the failure of the first.”

Good question, illustrating an interaction between historical facts and social science theorizing. Leão and Eyal continue:

“The comparison demonstrates that the recent widespread adoption of RCTs is not due to their inherent technical merits nor to rhetorical and organizational strategies. Instead, it reflects the ability of actors in the second wave to overcome the political resistance to randomized assignment, which has bedeviled the first wave, and to forge an enduring link between the fields of development aid and academic economics.”

This historical story is relevant to the course because it gets us thinking about the reasons that certain designs and analyses get used. To loop back to the beginning of the course, the goals of statistics are modeling, prediction, and causal inference—and, as discussed in Chapter 18 of *Regression and Other Stories*, causal inference can be thought of as a special case of prediction. Meanwhile, fundamental challenges of statistics are generalizing from sample to population, from treatment to control group, and from observed data to underlying constructs of interest. Many different approaches to design, data collection, measurement, and analysis are available, but they all go back to addressing these fundamental challenges to answer these fundamental questions.

2. The Harvard study claiming North Carolina is less democratic than North Korea

The story starts with this news article from 2017 in which a political science professor at the University of North Carolina wrote,<sup>92</sup>

“In 2005, in the midst of a career of traveling around the world to help set up elections in some of the most challenging places on earth . . . my Danish colleague, Jorgen Elklit, and I designed the first comprehensive method for evaluating the quality of elections around the world . . . In 2012 Elklit and I worked with Pippa Norris of Harvard University, who used the system as the cornerstone of the Electoral Integrity Project. Since then the EIP has measured 213 elections in 153 countries and is widely agreed to be the most accurate method for evaluating how free and fair and democratic elections are across time and place . . .”

In the just released EIP report, North Carolina’s overall electoral integrity score of 58/100 for the 2016 election places us alongside authoritarian states and pseudo-democracies like Cuba, Indonesia and Sierra Leone. If it were a nation state, North Carolina would rank right in the middle of the global league table—a deeply flawed, partly free democracy that is only slightly ahead of the failed democracies that constitute much of the developing world.”

This got us curious, so we searched on the web and found *The Year in Elections* by Norris et al., which included a map from 2014 with North Korea colored as one of the countries with “moderate” electoral integrity, in 65th place out of 127 countries. The poor saps in Bulgaria and Romania were ranked 90 and 92, respectively. Clearly what they needed was a dose of Kim Jong-il.

Let’s see what this measure actually is. From the report:<sup>93</sup>

“The survey asks experts to evaluate elections using 49 indicators, grouped into eleven categories reflecting the whole electoral cycle. Using a comprehensive instrument, listed at the end of the report, experts assess whether each national parliamentary and presidential contest meets international standards during the pre-election period, the campaign, polling day and its aftermath. The overall PEI index is constructed by summing the 49 separate indicators for each election and for each country. . . . Around forty domestic and international experts were consulted about each election, with requests to participate sent to a total of 4970 experts, producing an overall mean response rate of 29%. The rolling survey results presented in this report are drawn from the views of 1429 election experts.”

Now let’s check what the experts said about North Korea; it’s on page 9 of the report:

Electoral laws	53
Electoral procedures	73
District boundaries	73
Voter registration	83
Party and candidate registration	54
Media coverage	78
Campaign finance	84
Voting process	53
Vote count	74
Results	80
Electoral authorities	60

Each of these is on a 0–100 scale with 100 being the best. North Korea is above 50 in *every category* on the scale. What’s going on?

<sup>92</sup>Andrew Reynolds (2016), North Carolina is no longer classified as a democracy, *Raleigh News & Observer*, 22 Dec. The discussion here is taken from Andrew Gelman (2017), About that bogus claim that North Carolina is no longer a democracy, <https://statmodeling.stat.columbia.edu/2017/01/02/bogus-north-korea/>, with followup, Andrew Gelman (2017), “Constructing expert indices measuring electoral integrity”—reply from Pippa Norris, <https://statmodeling.stat.columbia.edu/2017/01/02/reply-pippa-norris/>. For more recent material, see the Electoral Integrity Project at <http://www.electoralintegrityproject.com/>.

<sup>93</sup>Pippa Norris, Ferran Martinez i Coma, and Max Grömping (2015), The year in elections, 2014, Harvard Kennedy School working paper RWP15-008, [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2567075](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2567075).

Let's look more carefully. From one of the tables in the report, the response rate for North Korea is given as 6%. And it said they consulted about 40 "domestic and international experts" for each election. 6% of 40 is 2.4, so maybe they got 3 respondents for North Korea, 2 of whom were Stalinists. Or, more realistically, maybe they were scoring North Korea not in comparison to other countries but in relation to their expectation, in which case it would be inappropriate to use these numbers to directly compare countries, as was done in the Electoral Integrity Index report, let alone to compare to U.S. states.

That report gave North Korea a rating of 65.3 out of 100 and Cuba a rating of 65.6. Both these numbers are higher than at least 27 of the 50 U.S. states in 2016, according to the ratings from the Electoral Integrity Project. And these claims were picked up in the news media, including the *New York Times*.<sup>94</sup>

In 2015, the director of the Electoral Integrity Project wrote,<sup>95</sup>

"The map identifies North Korea and Cuba as having moderate quality elections. The full report online gives details on how to interpret this. It does not mean that these countries are electoral or liberal democracies. The indicators measure expert perceptions of the quality of an election based on multiple criteria derived from international standards."

Later that year, they dropped North Korea from their reports, which is a start but we don't think is enough. First, it's disturbing that they didn't see a problem back in 2014 when they rated that dictatorship as over 50 on all dimensions of electoral integrity. Second, the North Korea numbers were created using the same approach used for all other countries. Given that we can all agree that North Korea's numbers were problematic, this calls into question the method more generally. On the plus side, it is good that the Harvard team is open about its data and methods, as this allows us to better assess what they have been doing wrong and gives them an opportunity to improve.

This example is relevant to the course in reinforcing fundamental issues of data collection and data quality, discussed in Chapter 2 of *Regression and Other Stories* but often forgotten in quantitative research. We also like that the last story of the semester is an example for which the data analysis is simple and the challenge comes in data collection and the definition of the problem.

### Class-participation activities

#### 1. Designing a paper helicopter

On page 31 of this book there is a homework problem in which students are asked to build a series of paper "helicopters," take measurements on their flight times, and use these data to design a helicopter to maximize time in the air.<sup>96</sup>

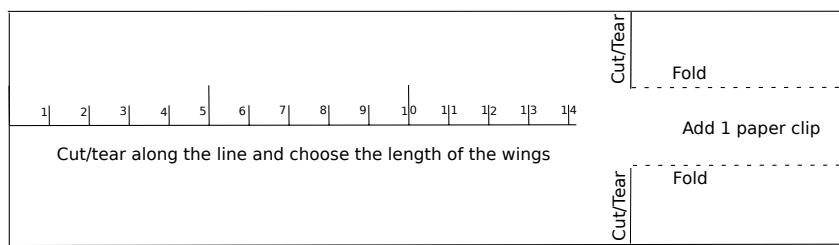
This exercise can be expanded into a classroom activity that touches on several aspects of design, data collection, and statistical modeling. At each step, students will only be given a few sheets of paper, so they will need to consider carefully how to design the helicopters they will test.

We start by simplifying and considering just one factor to modify, for example wing length. Figure 127 shows a template that can be printed out and handed to students. As usual, students work in pairs. In the first version of the activity, we give each pair of students three copies of the template and ask how they would choose the values for the one factor to build three helicopters so that they would expect to learn as much as possible.

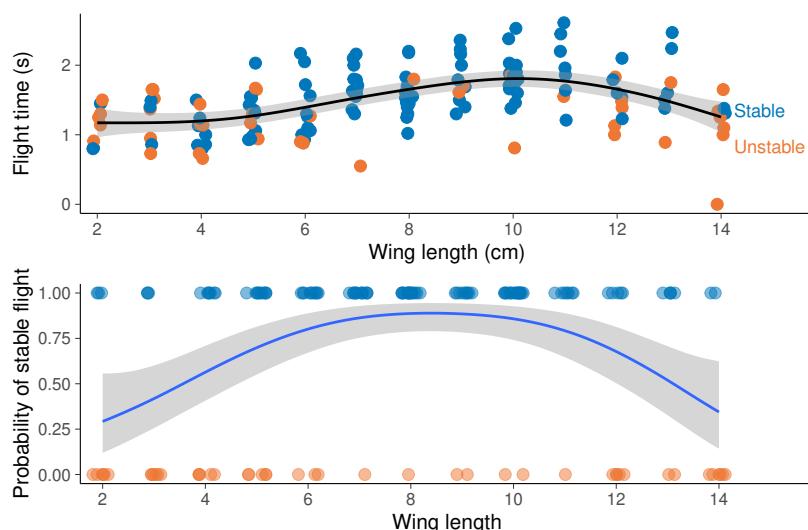
<sup>94</sup>Eduardo Porter (2017), A threat to U.S. democracy: Political dysfunction, *New York Times*, 3 Jan, <http://www.nytimes.com/2017/01/03/business/economy/trump-election-democracy.html>.

<sup>95</sup>See Pippa Norris (2015), The best and worst elections of 2014, *Washington Post*, 16 Feb, <https://www.washingtonpost.com/news/monkey-cage/wp/2015/02/16/the-best-and-worst-elections-of-2014/>.

<sup>96</sup>For further discussion and a more elaborate version of this activity, see George Box (1992), Teaching engineers experimental design with a paper helicopter, *Quality Engineering* 4, 453–459, <https://williamghunter.net/george-box-articles/teaching-engineers-experimental-design-with-a-paper-helicopter>.



**Figure 127** Template for the one-parameter helicopter activity, printed here at half scale so it can fit on the page. The instructor should print out many copies of this at full scale to hand to students, who can then cut out and construct their own helicopters.



**Figure 128** Results from the helicopter activity performed by students in a class. Expected flight time is a non-monotonic function of wing length, implying that a linear model would not be appropriate for modeling these data with a goal of finding a wing length to optimize flight time.

We then tell them they can have six helicopters, and that they can choose to design the experiment for all six at once, or they can first make three and then design sequentially new experiments. Which approach is better? This is the difference between batch and sequential designs.

If in the three first experiments the wing lengths are ordered  $x_1 < x_2 < x_3$  but the flight time  $t_2$  is the longest, we ask them to think how they could fit some simple nonlinear model such as a quadratic,  $y = b_0 + b_1x + b_2x^2 + \text{error}$ , to predict flight times for other values of  $x$ . Given this model, how should they choose the next values of  $x$ ?

Figure 128 shows results from a class where we performed this activity and gathered data by assigning a wing length to each pair of students. In these graphs, each dot represents a single drop of a helicopter from a fixed height, and the students were also asked to characterize each flight as being stable or unstable. The longest and stablest flights came with wings that were not too short or too long. A linear model would not be appropriate for optimizing this function.

If students start with 25 sheets of paper but get a time bonus for each non-used sheet, how should they decide when to stop making additional experiments (time is money)? This can be made into a competition. We can then ask students how they might proceed with additional design factors.

This activity is relevant to the week's reading in providing a motivation for nonlinear modeling

1. Choose a method you have learned during the semester
2. Review the method
3. Discuss where the method works and where it fails
4. Discuss relevance to your applied interests
5. Points of confusion and open questions

**Figure 129** Steps for the semester review activity. This can be displayed on the screen. Students should divide into groups of four, and then each group can go through the above steps.

amid variation, along with some data to try out such models. It relates to the course as a whole by connecting regression modeling to larger questions about experimental design, data collection, and decision making.

## 2. Semester review

For the last class, it makes sense to do an activity that looks back on the semester. The class should divide into groups of four, with each group having an in-depth discussion of a single topic from the course, with the discussion following the steps listed in Figure 129. When each group chooses a topic, they should announce it aloud and the instructor can write it on the board; that way the different groups of students will discuss different topics, and among them they should cover the entire course.

Having chosen their topics, each group should go through the further steps in Figure 129, reviewing the topic they have chosen, considering its relevance to applied work including their own outside interests, and listing and trying to resolve any confusions they have. The class as a whole can then rejoin and discuss various open questions and confusing points that have arisen from the separate group conversations.

## Computer demonstrations

### 1. Quadratic regression

We can demonstrate problems with the “too much talent” study (see page 166 of this book) by simulating data with diminishing returns and fitting a quadratic:

```
n <- 200
x <- runif(n, 0, 200)
y <- 500 + 1000*(1 - exp(-x/50)) + rnorm(n, 0, 100)
plot(x, y, pch=20, cex=.3)
fake <- data.frame(x, y)

fit_1 <- stan_glm(y ~ x, data=fake, refresh=0)
print(fit_1)
curve(coef(fit_1)[1] + coef(fit_1)[2]*x, add=TRUE, col="blue")
fit_2 <- stan_glm(y ~ x + I(x^2), data=fake, refresh=0)
print(fit_2)
curve(coef(fit_2)[1] + coef(fit_2)[2]*x + coef(fit_2)[3]*x^2, add=TRUE, col="red")
```

This idea can be expanded by simulation of a nonlinear model of a causal effect with nonrandom sampling and treatment assignment, following by a comparison of linear and nonlinear models.

### 2. Bias and unmodeled uncertainty

The first step of this demonstration is to show that a random sample of 10 000 people allows us to estimate the vote share in the population very precisely. The next step is to consider the case of a

biased sample (for example, due to response rates that differ by sex): here, the estimate remains precise but is systematically biased.

```
# Population characteristics
p_male <- 0.45 # voter support among males
p_female <- 0.60 # voter support among females
share_male <- 0.49 # population share of males
share_female <- 1 - share_male # population share of females
true_support <- p_male * share_male + p_female * share_female
print(round(true_support, 4))

# Estimate of voter support based on unbiased sample (with n = 10000)
n <- 10000
survey <- rbinom(n, size=1, prob=true_support)
p_hat <- mean(survey)
se <- sqrt(p_hat * (1 - p_hat) / n)
ci <- p_hat + c(-2, 2) * se
round(ci, 4)

# Phone response behavior and resulting biased measure of support
p_male_answering <- 0.05      # response probability for men
p_female_answering <- 0.10    # response probability for women
phone_male <- share_male * p_male_answering
phone_female <- share_female * p_female_answering
measured_support <- (phone_male * p_male + phone_female * p_female) /
  (phone_male + phone_female)
round(measured_support, 4)

# Estimate of voter support based on biased sample (also with n = 10000)
survey <- rbinom(n, size=1, prob=measured_support)
p_hat <- mean(survey)
se <- sqrt(p_hat * (1-p_hat) / n)
ci <- p_hat + c(-2, 2) * se
round(ci, 4)
```

## Drills

### 1. Basic statistics and linear regression

- Consider the following model of price  $x$  and sales  $y$ : when the price is \$20, sales are 2000 units, and for every 1% increase in price, sales decrease by 0.8%. Write this as a formula.  
*Solution:*  $y = 2000(x/20)^{-0.8}$  or  $\log y = \log(2000) - 0.8 \log(x/20)$
- You are planning to sample  $n$  people in a country in which 80% of the population are native born and 20% are immigrants. As part of the analysis you will compare these two groups according to the percentage who support more restrictive immigration laws. You want to estimate this difference to within a standard error of 5 percentage points. How large should  $n$  be?
- Write an R function to compute the average and standard deviation of 1000 random draws from a Poisson distribution with parameter  $\theta$ .
- Write R code to fit a linear regression with predictors  $x_1, x_2, x_3$ , and all their two-way interactions.
- List at least four of the assumptions of linear regression, in decreasing order of importance.

### 2. From logistic regression through causal inference

- Here is the result from a fitted logistic regression:

```
family:      binomial [logit]
formula:     y ~ x
observations: 100
predictors:  2
-----
          Median MAD_SD
(Intercept) 1.0    0.5
x           -0.3   0.1
```

Suppose you define  $z = 20 + 10 * x$ . What would be the estimated coefficients of the logistic regression of  $y$  on  $z$ ?

*Solution:*  $1.0 - 0.3x = 1.0 - 0.3(z - 20)/10 = 1.6 - 0.03z$ , so the estimated coefficients are 1.6 and  $-0.03$ .

- (b) Give R code for fitting an ordered logistic regression predicting an outcome  $y$  that can take on the values 1, 2, 3, 4, 5 from predictors  $x_1$  and  $x_2$ .
- (c) You are planning to conduct a randomized experiment with 100 people in the treatment group and 100 controls. The outcome is test scores, in a population where scores have a mean of 60 and standard deviation 15. You have a pre-test measurement, and you expect that the model fit to estimate the treatment effect will have an  $R^2$  of 50%. Approximately what will be the standard error of the estimated treatment effect?
- (d) In an experiment you have outcome  $y$ , treatment indicator  $z$ , and a pre-test variables  $x$  in a data frame called `sample`. You also have  $x$  for a population of interest in a data frame called `pop`. Give R code to estimate the average causal effect in the population, allowing for the treatment effect to vary with  $x$ .

### Discussion problems

#### 1. Designing an experiment using simulation

Suppose you want to design an experiment to estimate the effects of canvassing (face-to-face conversations) on voter turnout in an upcoming election.<sup>97</sup> You have records on a large number of registered voters, with data on their past voter turnout, and your plan is to randomly select  $n$  people from this database and randomly chose  $n/2$  to be contacted and encouraged to vote. You will then follow up after the election to see who actually voted. How someone votes is a secret, but whether a person votes in a particular election is public record.

The problem here is to choose a reasonable  $n$ , and you should do this by simulation, as follows. First start with a tentative guess, for example  $n = 1000$ . Then simulate data. The challenge here is not so much the programming as it is making reasonable guesses for all the components of the model. You need to make some assumption about past voter turnout (for example, the number of times a person has voted in the six previous elections) and then a model for the probability of voting in the current election, conditional on past voter turnout and the treatment indicator. At this point you should be able to simulate an experiment and fit a model to estimate the treatment effect from the simulated data, and then loop this to get a sense of the uncertainty of your estimate. If this uncertainty is large compared to the assumed true effect size, you can scale up  $n$  accordingly.

#### 2. Creating a better electoral integrity index

Suppose you wanted to create a better “electoral integrity index,” improving on the work of the Harvard team (see page 312). How would you do it? How could you put North Carolina and North Korea on the same scale? Consider issues of definition, measurement, and validation.

<sup>97</sup>For background, see Alan Gerber and Donald Green (2000), The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment, *American Political Science Review* 94, 653–663.

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

---

## Appendices

---

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

---

## Appendix A

# Pre-test questions

---

### A.1 First semester

We give a short three-part quiz at the beginning of the semester for students to answer at home on their own time, along with the following instructions:

This pre-test includes questions on statistics, mathematics, programming, and social science modeling, a mixture of background knowledge that would be useful to learn the course material and some topics we will cover in the course but you might have already seen in an earlier statistics course. *Please spend no more than 45 minutes on it.*

*This pre-test will not count toward your course grade.* We are giving it to assess your current state of knowledge, as this will allow us to better structure the course for all of you this semester. So you will directly benefit by doing your best on this exam and reporting your answers honestly. You can use pencil and paper but *no book, notes, or computer* and do not work with other students on this or get any outside help. You can discuss this pre-test with other students after you've turned it in.

Some of the problems might cover topics you know nothing about. That's OK. Again, we just want to get a sense of where everyone is right now, at the start of the semester.

Others of you might find most of these questions to be easy. That's OK too; this just tells us that you're well prepared.

**Math:**

1. Suppose a variable  $y$  takes on the value 200 in the year  $t = 1900$  and 240 in the year  $t = 2000$ . If this variable follows a linear time trend, what are the values of  $a$  and  $b$  in the equation of the line,  $y = a + bt$ ?
  - (a)  $a = -560, b = 0.4$
  - (b)  $a = 200, b = 0.4$
  - (c)  $a = 160, b = 4$
  - (d)  $a = -1700, b = 40$
2. What is the logarithm of 0? (We use natural log, not log base 10.)
  - (a)  $-1$
  - (b)  $1/e$
  - (c)  $0$
  - (d)  $\text{undefined}$
3. What is the average of the numbers  $1, 2, 3, \dots, 100$ ?
  - (a) 49.5
  - (b) 50
  - (c) 50.5
  - (d) 51

4. Which of these functions equals 0 when  $x = 0$  and asymptotes to the value 10 as  $x \rightarrow \infty$ ?

- (a)  $y = 10x/\infty$
- (b)  $y = 10 \exp(-x)$
- (c)  $y = 10(1 - \exp(-x))$
- (d)  $y = 10 - 0.1x^2$

5. If the vector  $\hat{y}$  is defined as  $\hat{y} = X\hat{\beta}$ , with  $X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$  and  $\hat{\beta} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ , what is  $\hat{y}$ ?

- (a)  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- (b)  $\begin{pmatrix} 2 \\ 4 \end{pmatrix}$
- (c)  $\begin{pmatrix} 3 \\ 7 \\ 11 \end{pmatrix}$
- (d)  $\begin{pmatrix} 2 \\ 6 \\ 10 \end{pmatrix}$

**Computing:**

1. What is one way to make a scatterplot of two vectors,  $x$  and  $y$ , in R?

- (a) `plot(x, y)`
- (b) `plot x, y`
- (c) `scatter(x, y)`
- (d) `scatter x, y`

2. Which of these lines of R code generates ten random numbers between  $-2$  and  $+2$ ?

- (a) `sample(c(-2,2), 10)`
- (b) `seq(-2, 2, length=10)`
- (c) `runif(10, -2, 2)`
- (d) `rnorm(10, -2, 2)`

3. Which of these lines of R code fits a regression of  $y$  on  $x$ , with data set `data0`, using the `lm` command, storing the result in object `fit`?

- (a) `lm$fit(y ~ x, data=data0)`
- (b) `fit(y ~ x, data=data0)`
- (c) `fit(y ~ x, data=data0, model=lm)`
- (d) `fit <- lm(y ~ x, data=data0)`

4. You want to write a line of R code that regresses `income` on `age` and `male` as well as the interaction of both variables. The data is stored in the data frame `data0`. Which of the following does *not* work?

- (a) `lm(income ~ age + sex + age:male, data=data0)`
- (b) `lm(income ~ 0 + age + male + age:male, data=data0)`
- (c) `lm(income ~ age + male + age*male, data=data0)`
- (d) `lm(income ~ (age + male)^2, data=data0)`

5. Which of these lines of R code takes a vector of numbers,  $a$ , and centers the observations at their mean?
- (a)  $a - \text{mean}(a)$
  - (b)  $a/\text{mean}(a)$
  - (c)  $\text{mean}(a)$
  - (d)  $a - \text{mean}(a - \text{mean}(a))$

**Statistics:**

1. What is the approximate interval containing 95% of observations, if your data are normally distributed with mean 5 and standard deviation 2?
  - (a)  $(3, 7)$
  - (b)  $(1, 9)$
  - (c)  $(5 - \sqrt{2}, 5 + \sqrt{2})$
  - (d)  $(5 - 2\sqrt{2}, 5 + 2\sqrt{2})$
2. Consider the regression line  $y = 0.2 + 0.3x + \text{error}$ , with errors normally distributed with mean 0 and standard deviation 2. You observe a new data point at  $x = 10$ . Which of these is a 68% predictive interval for the corresponding observation  $y$ ?
  - (a)  $(1.2, 5.2)$
  - (b)  $(-0.8, 7.2)$
  - (c)  $(3.2 - 2/\sqrt{n}, 3.2 + 2/\sqrt{n})$
  - (d)  $(\bar{y} - 2/\sqrt{n}, \bar{y} + 2/\sqrt{n})$
3. You record the median income in each state in 2013 and in 2015, and then you run a linear regression on these 50 data points, predicting 2015 median income from 2013 median income. What will be the approximate slope of this regression?
  - (a) 0
  - (b) 0.5
  - (c) 1
  - (d) 2
4. You gather data on a random sample of 500 American adults and record their age and their feeling thermometer score (a 0–100 measure) toward Donald Trump. You fit a linear regression predicting feeling thermometer score from age. The result is the line  $y = a + bx + \text{error}$ , with errors normally distributed with mean 0 and standard deviation  $\sigma$ . What is a plausible value for  $b$ ?
  - (a) -0.20
  - (b) -0.02
  - (c) 0.02
  - (d) 0.20
5. You have a data set with the variables `income`, `age`, `male`. What regression could you fit to estimate the difference between the average incomes of men and women?
  - (a) `lm(income ~ male)`
  - (b) `lm(income ~ age)`
  - (c) `lm(income ~ male + age)`
  - (d) `lm(income ~ male*age)`

## A.2 Second semester

At the beginning of the second semester, we give students the first-semester final exam (two questions from each chapter; see Section B.1) for students to answer at home on their own time, along with the following instructions:

This is the final exam from the first semester of this course. *Please spend no more than 90 minutes on it.*

*This pre-test will not count toward your course grade.* We are giving it to you to assess your current state of knowledge, as this will allow us to better structure the course for all of you this semester. So you will directly benefit by doing your best on this exam and reporting your answers honestly. You can use pencil and paper but *no book, notes, or computer* and do not work with other students on this or get any outside help. You can discuss this pre-test with other students after you've turned it in.

Some of the problems might stump you. That's OK. If so, please bring up your questions in class right away so we can all start on the same page.

Others of you might find most of these questions to be easy. That's OK too; this just tells us that you're well prepared.

## Appendix B

# Final exam questions

---

### B.1 Multiple-choice questions for the first semester

Here are a few multiple-choice questions we prepared for each chapter in the first half of *Regression and Other Stories*, representing a mix of conceptual, programming, and methods problems. As discussed in Section 1.5 of this book, we sampled one question for each chapter to construct a practice final exam, then sampled two questions for each chapter to form the exam itself.

Section B.2 gives questions for the second semester. For a faster-moving course that covers all this material in one semester, we just combine the multiple-choice questions in Sections B.1 and B.2, sampling one question for each chapter to construct a practice final exam, then sampling a different question for each chapter to form a 24-question exam.

#### 1. Questions for Chapter 1 of *Regression and Other Stories*

- (a) Here is a regression predicting incumbent party's vote percentage from economic growth, fit to several recent elections:

	Median	MAD_SD
(Intercept)	46.7	1.4
growth	2.8	0.6

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 3.7 0.7

Someone summarizes the model as follows: “A 1% increase in economic growth is predicted to lead to a 2.8% share in incumbent party’s vote share.” What is wrong with that statement?

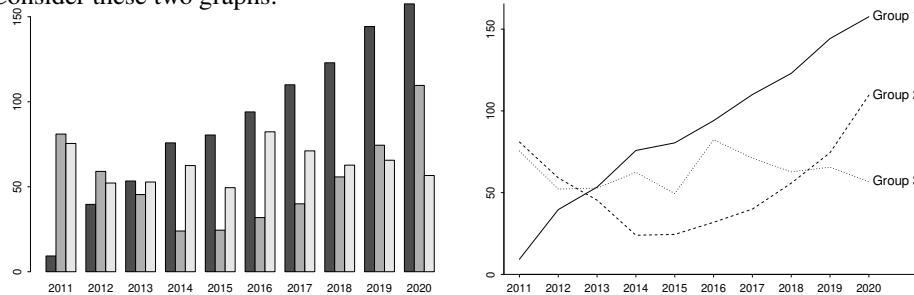
- i. The estimate of 2.8 is not statistically significant so it is misleading to present it in this way.
  - ii. With these observational data we can only make descriptive, not causal, claims.
  - iii. People are interested in who wins the election, not in vote share.
  - iv. The model accounts for the economy but leaves out predictors for other important factors such as foreign policy and social issues.
- (b) A group of students are being trained to do a particular skill, and for each student we record the number of hours they are trained and their score on a later test. We see a pattern of diminishing returns: at zero hours of training, the test scores are zero; for low values of training, there is an approximately positive and linear relation between hours of training and test score; for high values of training, the relationship levels off and becomes flat. We fit a linear regression of the form,  $\text{score} = a + b * \text{hours} + \text{error}$ . What can we say about the estimates of  $a$  and  $b$ ?
- i.  $a$  will be negative,  $b$  will be positive.
  - ii.  $a$  will be approximately zero,  $b$  will be positive.
  - iii.  $a$  will be positive,  $b$  will be positive.
  - iv. It will not be possible to fit the regression because the data show a nonlinear relationship.

- (c) Suppose a variable  $y$  takes on the value 200 in the year  $t = 1900$  and 240 in the year  $t = 2000$ . If this variable follows a linear time trend, what are the values of  $a$  and  $b$  in equation of the line,  $y = a + bt$ ?
- $a = -560, b = 0.4$
  - $a = 200, b = 0.4$
  - $a = 160, b = 4$
  - $a = -1700, b = 40$
- (d) Which of the following statements is correct?
- “Generalizing from sample to population” is a concern for sample surveys but not for causal inference from randomized experiments.
  - “Generalizing from treatment to control group” is a concern for randomized experiments but not for observational studies.
  - “Generalizing from observed measurements to the underlying constructs of interest” refers to the problem of extrapolation to people who are systematically different from those in the study.
  - Regression can be used to generalize from sample to population and to generalize from treatment to control group.

2. Questions for Chapter 2 of *Regression and Other Stories*

- (a) What is one way to make a scatterplot of two vectors,  $x$  and  $y$ , in R?
- `plot(x, y)`
  - `plot x, y`
  - `scatter(x, y)`
  - `scatter x, y`
- (b) Which of these lines of R code graphs the function  $y = 2 + 3x$ ?
- `abline(2, 3)`
  - `abline(2 + 3*x)`
  - `curve(2 + 3x)`
  - `curve(2, 3)`
- (c) Which of these lines of R code generates 10 random numbers between  $-2$  and  $+2$ ?
- `runif(10, -2, 2)`
  - `rnorm(10, -2, 2)`
  - `sample(c(-2,2), 10)`
  - `seq(-2, 2, length=10)`

- (d) Consider these two graphs:



We typically prefer the line plot to the bar plot. Why?

- Both plots look good on paper but the line plot will be easier to view on a computer or mobile device.
- The line plot shows more information than the bar plot.
- It is easier to read the numbers off the line plot than the bar plot.
- The line plot allows a more direct comparison of the time patterns in the three groups.

B.1. MULTIPLE-CHOICE QUESTIONS FOR THE FIRST SEMESTER

327

3. Questions for Chapter 3 of *Regression and Other Stories*

- (a) A survey is conducted on 100 Americans, who are asked whether they approve of the president's job performance. What is the approximate standard error for the resulting estimate?
- i. 0.01
  - ii. 0.03
  - iii. 0.05
  - iv. 0.10
- (b) A population of heights is normally distributed with mean 65 inches and standard deviation 3.2 inches. Which of the following lines of R code gives the proportion of people who are taller than 69.5 inches?
- i. `dnorm(65, 69.5, 3.2)`
  - ii. `1 - dnorm(65, 69.5, 3.2)`
  - iii. `1 - pnorm(69.5, 65, 3.2)`
  - iv. `rnorm(69.6, 65, 3.2)`
- (c) You perform a survey that estimates the average monthly spending on a particular consumption category to be \$50 for men and \$55 for women. The standard errors of these estimates are \$3 and \$4, respectively. What do you estimate the difference in spending to be, and what is your estimate's standard error?
- i.  $5 \pm 1$
  - ii.  $5 \pm 5$
  - iii.  $5 \pm 7$
  - iv.  $5 \pm 7/\sqrt{n}$
- (d) A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 51% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 51%, sometimes lower.
- For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?
- i. The larger hospital
  - ii. The smaller hospital
  - iii. About the same (that is, within 5% of each other)

4. Questions for Chapter 4 of *Regression and Other Stories*

- (a) You perform an experiment on  $n$  people. You get an estimated treatment effect of 0.3 with a standard error of 0.2. Is this considered "statistically significant"?
- i. Yes
  - ii. No
  - iii. Yes, if  $n > 30$
  - iv. It depends on whether the data are normally distributed
- (b) A basketball player takes 20 free throws and makes 14 of them. What is the 95% interval for her success probability?
- i.  $0.7 \pm 2\sqrt{0.7 * 0.3 / 20}$
  - ii.  $0.7 \pm 2\sqrt{0.7 * 0.3 * 20}$
  - iii.  $14 \pm 2\sqrt{0.7 * 0.3 / 20}$
  - iv.  $14 \pm 2\sqrt{0.7 * 0.3 * 20}$
- (c) An intervention has a true effect size of 0.2. You design an experiment that will produce an unbiased estimate with standard error 0.1. If you run the experiment, what is the probability that your estimate will be positive and "statistically significant"?

- i. 16%
  - ii. 50%
  - iii. 95%
  - iv. 97.5%
- (d) Out of a random sample of 100 Americans, zero report having ever been robbed. From this information, you compute a 95% confidence interval for the proportion of Americans who have ever been robbed. What should your interval be?
- i. (0, 0) using the standard error  $\sqrt{\hat{p}(1 - \hat{p})/n}$  with estimate  $\hat{p} = 0$
  - ii. (0, 0.05) using the standard error  $\sqrt{\hat{p}^*(1 - \hat{p}^*)/n^*}$  with  $p^* = \frac{y+2}{n+2}$  and  $n^* = n + 4$
  - iii. (0, 0.10) using the standard error  $0.5/\sqrt{n}$
  - iv. (0, 1) because with 0 successes you have no information about  $p$
- (e) What is a  $p$ -value?
- i. The width of a 95% confidence interval
  - ii. The probability of observing something at least as extreme as the observed test statistic
  - iii. The probability of estimating the correct sign
  - iv. The probability that the null hypothesis is true
- (f) A study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to Group A (where they were asked to write daily about positive things they were blessed with) or Group B (where they were asked to write daily about misfortunes that others had to endure). Participants were then tracked until all had died. Participants who wrote about the positive things they were blessed with lived, on average, 8.2 months after diagnosis whereas participants who wrote about others' misfortunes lived, on average, 7.5 months after diagnosis. The standard error of the difference was 1.0, and the  $p$ -value (compared to the hypothesis of no effect) was 0.27. Which statement is the most accurate summary of the results?
- i. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was *greater* than that lived by the participants who were in Group B.
  - ii. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was *less* than that lived by the participants who were in Group B.
  - iii. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was *no different* than that lived by the participants who were in Group B.
  - iv. Speaking only of the subjects who took part in this particular study, it *cannot be determined* whether the average number of post-diagnosis months lived by the participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

##### 5. Questions for Chapter 5 of *Regression and Other Stories*

- (a) A player takes 10 basketball shots, with a 40% probability of making each shot. Assume the outcomes of the shots are independent. Which of the following lines of code does *not* create a random variable  $y$  representing the number of shots that go in?
- i.  $y <- rbinom(1, 10, 0.4)$
  - ii.  $y <- sum(rbinom(10, 1, 0.4))$
  - iii.  $y <- sum(rbinom(10, 10, 0.4))$
  - iv.  $y <- 0; for (i in 1:10) y <- y + rbinom(1, 1, 0.4)$
- (b) Two players each take 50 basketball shots. Player A has a 40% chance of making each shot and player B has a 30% chance of making each shot. Assume the outcomes of the shots are independent. You write the following code to simulate this contest 1000 times to approximately compute the probability that Player A makes more shots than player B:

```
n_sims <- 1000  
y_A <- rbinom(n_sims, 50, 0.4)  
y_B <- rbinom(n_sims, 50, 0.3)  
print(y_A > y_B)
```

This code is not right. What is the problem?

- i. The code has no loop. To simulate the contest 1000 times you need a loop.
  - ii. Your goal is to estimate a continuous probability so you should be using `rnorm`, not `rbinom`.
  - iii. To simulate the contest you need to simulate the outcomes for both players at once, so it's wrong to simulate `y_A` and `y_B` using two separate lines of code.
  - iv. `y_A - y_B` is a vector so you need to take its mean before printing it at the end.
- (c) Which of these lines of R code prints, successively, the rows of matrix `a` (with dimensions  $5 \times 3$ )?
- i. `for (i in 1:3) print(a[i,])`
  - ii. `for (i in 1:5) print(a[i,])`
  - iii. `for (i in 1:3) print(a[,i])`
  - iv. `for (i in 1:5) print(a[,i])`
- (d) You want to write an R function to simulate  $n$  data points from a regression model, with data points  $x$  randomly sampled from the range  $(0, 100)$ , then fit the regression model to the simulated data, print the fitted model, and return the simulated data. Here is the function you write:

```
sim <- function(n, a, b, sigma){  
  library("rstanarm")  
  x <- runif(n, 0, 100)  
  y <- rnorm(a + b*x, sigma)  
  data <- data.frame(x, y)  
  fit <- lm(y ~ x, data=data)  
  print(fit)  
  data.frame(x, y)}
```

One of the lines in this code is wrong. What is the mistake?

- i. You can't put a `library()` call inside a function.
- ii. The specification of `y` is wrong because the error term was not added.
- iii. The `rnorm` function simulating `y` did not specify `n`.
- iv. The data frame cannot be called `data`.

## 6. Questions for Chapter 6 of *Regression and Other Stories*

- (a) Which of these lines of R code fits a regression of `y` on `x`, with data set `data0`, using the `lm` command, storing the result in object `fit`?
- i. `lm$fit(y ~ x, data=data0)`
  - ii. `fit(y ~ x, data=data0)`
  - iii. `fit(y ~ x, data=data0, model=lm)`
  - iv. `fit <- lm(y ~ x, data=data0)`
- (b) You record the median income in each state in 2013 and in 2015, and then you run a linear regression on these 50 data points, predicting 2015 median income from 2013 median income. What will be the approximate slope of this regression?
- i. 0
  - ii. 0.5
  - iii. 1
  - iv. 2

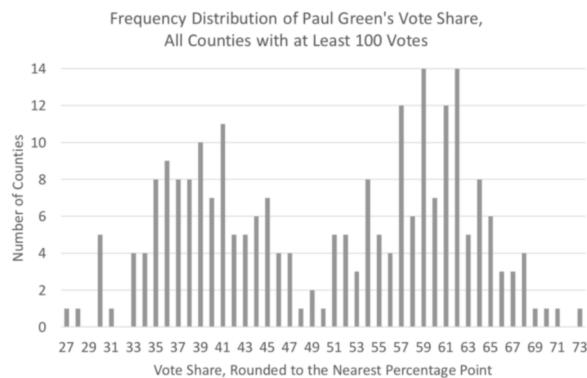
- (c) Consider a pre-test, post-test situation where scores on both tests have mean 50, and the regression of post-test on pre-test has a slope of 0.6. A student scores 70 on the pre-test. What is that student's expected score on the post-test?
- i. 42
  - ii. 62
  - iii. 72
  - iv. 92
- (d) Suppose you have conducted an experiment on 100 people and you run a regression whose estimated slope is 2.5 standard errors from zero. You now have cause to gather data on an additional group of 50 people and fit the regression just to these new data. What is a reasonable guess of the probability that the estimated slope will be "statistically significant" in the replication?
- i. 5%
  - ii. 40%
  - iii. 85%
  - iv. 95%

7. Questions for Chapter 7 of *Regression and Other Stories*

- (a) One of the questions asked in the 2020 American National Election Study was feeling thermometer score (a 0–100 measure) toward Donald Trump. We fit a linear regression predicting feeling thermometer score from age. The result is the line  $y = a + bx + \text{error}$ , with errors normally distributed with mean 0 and standard deviation  $\sigma$ . What is a plausible value for  $b$ ?
- i. -0.20
  - ii. -0.02
  - iii. 0.02
  - iv. 0.20
- (b) You have a data set with the variables `income`, `age`, `male`. What regression could you fit to estimate the difference between the average incomes of men and women?
- i. `lm(income ~ male)`
  - ii. `lm(income ~ age)`
  - iii. `lm(income ~ male + age)`
  - iv. `lm(income ~ male*age)`
- (c) A survey is done of  $n/2$  women and  $n/2$  men, with the goal being to estimate the gender gap: the difference in support for some candidate comparing the two sexes. Assuming the survey is a random sample, approximately how large does  $n$  have to be so the standard error of the estimated gender gap is 5 percentage points?
- i. 100
  - ii. 200
  - iii. 400
  - iv. 800
- (d) The slope from a regression with an outcome variable  $y$  and a single predictor variable  $x$  that is a binary indicator variable is the same as what calculation?
- i. The difference in the mean of  $x$  and the mean of  $y$
  - ii. The difference in the mean of  $x$  when  $y = \text{mean}(y)$  and the mean of  $x$  when  $x = \text{mean}(x)$
  - iii. The difference in the mean of  $x$  when  $y = 0$  and the mean of  $x$  when  $y = 1$
  - iv. The difference in the mean of  $y$  when  $x = 0$  and the mean of  $y$  when  $x = 1$
- (e) The following graph was presented along with this explanation: "The histogram has two peaks, at 40% and 60% of the vote. These correspond to second/first ballot position, and imply a ballot effect of 20 percentage points."

## B.1. MULTIPLE-CHOICE QUESTIONS FOR THE FIRST SEMESTER

331



How can this comparison be expressed as a regression?

- Data points are counties,  $y$  is Paul Green's vote share in the county,  $x$  is Green's vote share in the county in the previous election. Regression of  $y$  on  $x$  has estimated slope of 0.20.
  - Data points are counties,  $y$  is Paul Green's vote share in the county,  $x = 1$  if Green was listed first on the ballot in the county or 0 if he was listed second on the ballot. Regression of  $y$  on  $x$  has estimated slope of 0.20.
  - Data points are counties,  $y$  is Paul Green's vote share in the county,  $x$  is Green's vote share in the county in the previous election,  $z = 1$  if Green was listed first on the ballot in the county or 0 if he was listed second on the ballot. Regression of  $y$  on  $x$  and  $z$  has estimated coefficient of 0.20 for  $x$ .
  - Data points are counties,  $y$  is Paul Green's vote share in the county,  $x$  is Green's vote share in the county in the previous election,  $z = 1$  if Green was listed first on the ballot in the county or 0 if he was listed second on the ballot. Regression of  $y$  on  $x$  and  $z$  has estimated coefficient of 0.20 for  $z$ .
8. Questions for Chapter 8 of *Regression and Other Stories*
- Which of these quantities is *not* minimized in ordinary linear regression?
    - The average of the residuals
    - The sum of the squares of the residuals
    - The residual standard deviation
    - The negative of the likelihood
  - A regression is fit predicting final exam scores from midterm exam scores. The midterm exam scores are approximately normally distributed with mean 50 and standard deviation 15.

	Median	MAD_SD
(Intercept)	24.8	1.4
midterm	0.5	0.1

Auxiliary parameter(s):

	Median	MAD_SD
sigma	11.6	0.3

Suppose a student has a midterm exam score of  $x$ . Which of the following R functions returns the point prediction and approximate predictive standard deviation of this student's final exam score?

- $c(24.8 + 0.5*x, 11.6)$
- $c(24.8 + 0.5*x, sqrt(1.4^2 + 0.1^2))$
- $c(24.8 + 0.5*(x-50), sqrt(1.4^2 + 0.1^2*(x-50)^2))$
- $c(24.8 + 0.5*(x-50), 11.6)$

- (c) Consider the following regression of earnings (in dollars) on height (in inches):

Median MAD\_SD  
(Intercept) -85000 9000  
height 1600 100

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 22000 400

What will the model look like if we instead predict earnings in dollars from height in centimeters?

i. Median MAD\_SD  
(Intercept) -33500 9000  
height\_cm 630 100

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 8700 400

ii. Median MAD\_SD  
(Intercept) -33500 3500  
height\_cm 630 40

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 22000 400

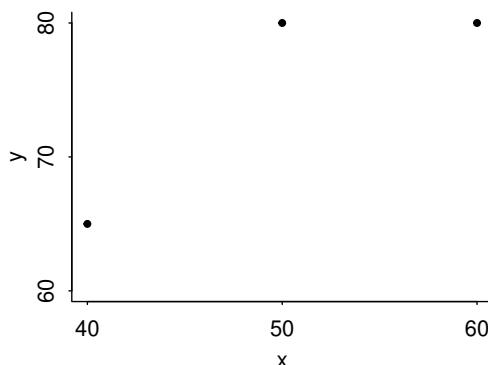
iii. Median MAD\_SD  
(Intercept) -85000 9000  
height\_cm 630 100

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 8700 160

iv. Median MAD\_SD  
(Intercept) -85000 9000  
height\_cm 630 40

Auxiliary parameter(s):  
Median MAD\_SD  
sigma 22000 400

- (d) Given these three data points:



What is the residual sum of squares for the line  $y = 40 + 0.5x$ ?

- i. 125
  - ii. 225
  - iii. 350
  - iv. The points don't fall on a straight line, so the residual sum of squares is undefined.
- (e) Which of the following lines of R code returns a regression  $y$  on  $x$  with the intercept fixed at 0?
- i. stan\_glm(y ~ x, data=data0)
  - ii. stan\_glm(y ~ x, data=data0, intercept=NULL)
  - iii. stan\_glm(y ~ x, data=data0, prior\_intercept=NULL)
  - iv. stan\_glm(y ~ 0 + x, data=data0)

## B.1. MULTIPLE-CHOICE QUESTIONS FOR THE FIRST SEMESTER

333

- (f) A regression of the form  $y = a + bx + \text{error}$  is fit to data from 50 students, yielding an estimated slope  $b$  of 2.4 with standard error 2.0. The residual standard deviation  $\sigma$  is estimated as 1.6. Suppose data were collected from 200 students from the same population and the model was fit to those 200 students. Which of the following would you expect to see?
- A standard error of the slope of approximately 0.5 and a residual standard deviation of approximately 0.4
  - A standard error of the slope of approximately 1.0 and a residual standard deviation of approximately 0.8
  - A standard error of the slope of approximately 2.0 and a residual standard deviation of approximately 0.8
  - A standard error of the slope of approximately 1.0 and a residual standard deviation of approximately 1.6
9. Questions for Chapter 9 of *Regression and Other Stories*
- (a) Consider the regression line  $y = 0.2 + 0.3x + \text{error}$ , with errors normally distributed with mean 0 and standard deviation 2. You observe a new data point at  $x = 10$ . What is a 68% predictive interval for the corresponding observation  $y$ ?
- (1.2, 5.2)
  - (−0.8, 7.2)
  - $(3.2 - 2/\sqrt{n}, 3.2 + 2/\sqrt{n})$
  - $(\bar{y} - 2/\sqrt{n}, \bar{y} + 2/\sqrt{n})$
- (b) What is the approximate interval containing 95% of observations, if your data are normally distributed with mean 5 and standard deviation 2?
- (3, 7)
  - (1, 9)
  - $(5 - \sqrt{2}, 5 + \sqrt{2})$
  - $(5 - 2\sqrt{2}, 5 + 2\sqrt{2})$
- (c) An experiment is performed by Wikipedia to estimate the effect on donations of a certain change in the request page. In the experiment, 10 000 people receive the treatment and 10 000 receive the control. The observed proportion of people who donate is 1.2% under the treatment and 1.4% under the control. The resulting estimated treatment effect is −0.2 percentage points with a standard error of 0.16 percentage points. Your prior is that the treatment effect is as equally likely to be positive as negative, and you have a 68% prior probability that the treatment effect will be between −0.1 and +0.1 percentage points. What is your posterior estimate of the treatment effect?
- −0.20 percentage points
  - −0.14 percentage points
  - −0.10 percentage points
  - −0.06 percentage points
- (d) Why might we use the R function `posterior_linpred()` instead of `posterior_predict()`?
- When we want the best point prediction
  - When we want to incorporate uncertainty into our prediction
  - When we want uncertainty in the predicted average rather than for a single case
  - When we want to use the normal distribution
- (e) Consider this model that was used to predict the yield of mesquite bushes:

```
fit <- stan_glm(log(weight) ~ log(canopy_volume) +
  log(canopy_slope) + group, data=mesquite)
```

We wish to use this model to make inferences about the average mesquite yield in a new set of trees which is summarized by a data frame called `new_trees`. Which of these lines of R code gives an estimate and standard error for this prediction?

- i. 

```
a <- exp(posterior_linpred(fit, newdata=new_trees))
print(c(mean(a), sd(a)))
```
- ii. 

```
a <- exp(posterior_predict(fit, newdata=new_trees))
print(c(mean(a), sd(a)))
```
- iii. 

```
a <- exp(posterior_linpred(fit, newdata=new_trees))
b <- rowMeans(a)
print(c(mean(b), sd(b)))
```
- iv. 

```
a <- exp(posterior_predict(fit, newdata=new_trees))
b <- rowMeans(a)
print(c(mean(b), sd(b)))
```

10. Questions for Chapter 10 of *Regression and Other Stories*

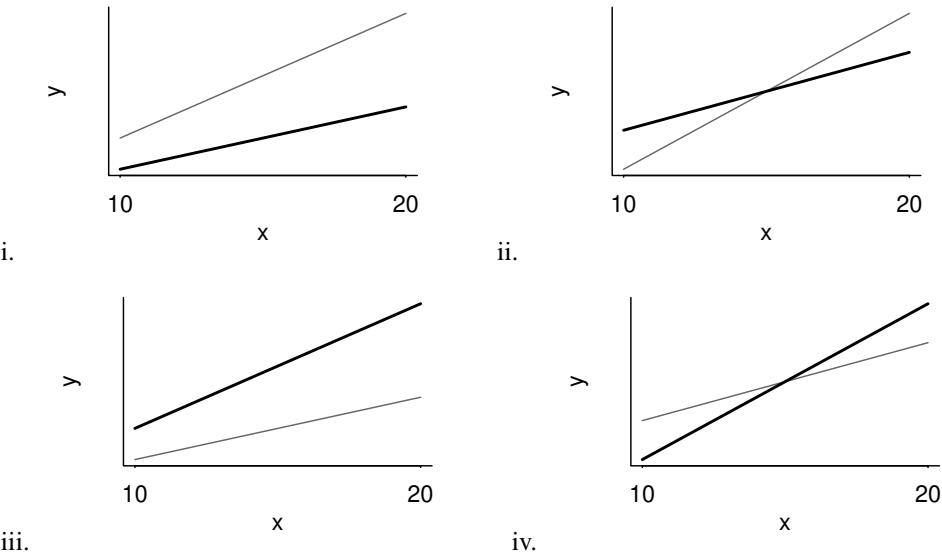
- (a) You want to write a line of R code that regresses income on age and male as well as the interaction of both variables. The data are stored in the data frame `data0`. Which of the following does *not* fit this model?

- i. `lm(income ~ age + male + age:male, data=data0)`
- ii. `lm(income ~ 0 + age + male + age:male, data=data0)`
- iii. `lm(income ~ age + male + age*male, data=data0)`
- iv. `lm(income ~ (age + male)^2, data=data0)`

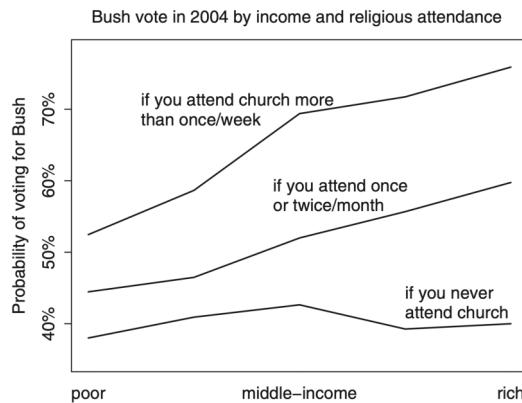
- (b) You are trying to assess the health benefits of walking. Your data variables are called `health` (with 0 = bad health and 10 = good health), `walks` (with 0 = no walking, 1 = infrequent walking, 2 = regular walks), and `smoker` (with 0 = nonsmoker and 1 = smoker). Which of these regression specifications attempts to address the question of whether the benefits of walking are different for smokers than for non-smokers?

- i. `walks ~ health + smoker`
- ii. `walks ~ health + smoker + health:smoker`
- iii. `health ~ walks + smoker`
- iv. `health ~ walks + smoker + walks:smoker`

- (c) Which of the following graphs represents the model,  $y = 15 + 0.4x + z - 0.2xz$ , for  $x$  in the range (10, 20) and the light and dark lines corresponding to  $z = 0$  and 1, respectively?



- (d) Consider the following data:



Suppose this is summarized by a regression predicting probability of supporting Bush given church attendance (coded as 0 if you never attend church, 1 if you attend once or twice a month, 2 if you attend more than once per week) and income (coded on a 1–5 scale, with 1 corresponding to poor and 5 corresponding to rich). The regression includes the intercept, both predictors, and their interaction. What is the approximate value of the coefficient for income?

- i. 0
- ii. 0.1
- iii. 0.2
- iv. 0.3

- (e) A regression was fit to country  $\times$  year data, predicting the rate of civil conflicts given a set of geographic and political predictors. Here are the estimated coefficients and their z-scores (estimate divided by standard error):

	Estimate	(z-score)
(Intercept)	-3.814	(-20.178)
Pre-2000 conflict	0.020	(1.861)
Border distance	0.000	(0.450)
Capital distance	0.000	(1.629)
Population	0.000	(2.482)
Percent mountainous	1.641	(8.518)
Percent irrigation	-0.027	(-1.663)
GDP per capita	-0.000	(-3.589)

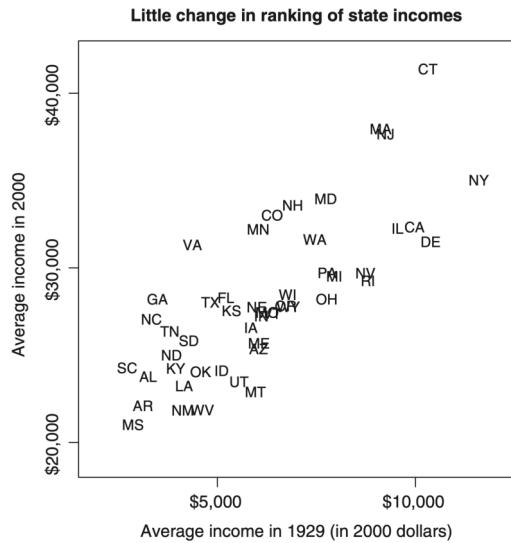
Why are the estimated coefficients for border distance, capital distance, population, and per-capita GDP so small?

- i. When you fit a regression with a large number of predictors, you can expect each individual predictor to have only a small effect.
- ii. The analysis has data on multiple years for each country; this will result in smaller coefficients because of duplication of data.
- iii. The coefficient corresponds to comparing items that differ by 1 unit in the predictor, and a difference of 1 km in distance, one person in population, or one dollar in per-capita GDP is tiny.
- iv. The intercept corresponds to the predicted outcome when all the predictors are set to zero, and it does not make sense to think of border distance, capital distance, population, and per-capita GDP as zero.

- (f) Consider the following interaction model:  $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \text{error}$ . How would we estimate the average difference in  $y$  between when  $x = 1$  and  $z = 4$  and when  $x = 2$  and  $z = 4$ ?

- i. The difference between  $\beta_1$  and  $2\beta_1 + \beta_3$
  - ii. The difference between  $\beta_1 + 4\beta_3$  and  $2\beta_1 + 4\beta_2 + 8\beta_3$
  - iii. The difference between  $\beta_1 + \beta_3$  and  $2\beta_1 + 2\beta_3$
  - iv. The difference between  $\beta_1 + 4\beta_3$  and  $2\beta_1 + 8\beta_3$
11. Questions for Chapter 11 of *Regression and Other Stories*

- (a) Consider the following data:



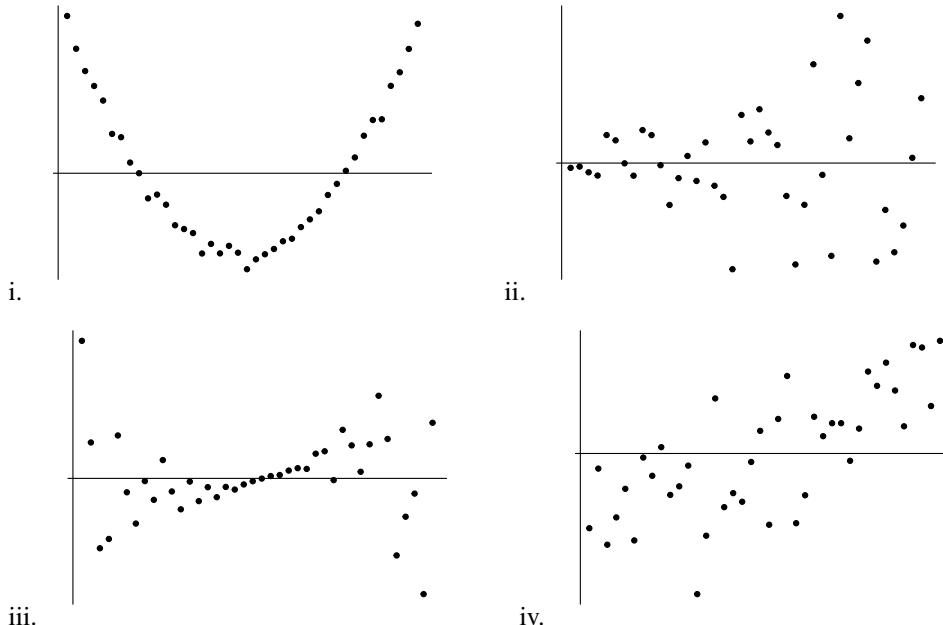
A regression is fit to these data, predicting state-level income in 2000 from state-level income in 1929. Then another regression is fit, just to the 25 states that were poorest in 1929 (so, same regression model, just fit to those 25 data points). What happens to the  $R^2$ , comparing these two fits?

- i. Fitting the model to just the poorer states, there is less unexplained variance compared to the full model, so  $R^2$  clearly goes up.
  - ii. Fitting the model to just the poorer states, the unexplained variance is about the same as in the full model, but the explained variance is lower, so  $R^2$  clearly goes down.
  - iii. The fitted model does not change much, so  $R^2$  stays roughly the same.
  - iv. The model only has state averages, not data on individual people, so  $R^2$  is not defined given the information in the graph.
- (b) A linear regression, `fit <- stan_glm(y ~ pred + z + pred:z, data=data)`, is fit; the data are plotted using `plot(pred, y)`; and the coefficient estimates are saved as `b <- coef(fit)`. Which of the following R code will plot the regression lines corresponding to  $z = 0$  and  $1$  in blue and red, respectively?
- i. `abline(b[1], b[2], col="blue")`  
`abline(b[3], b[4], col="red")`
  - ii. `abline(b[1], b[2], col="blue")`  
`abline(b[1] + b[3], b[2] + b[4], col="red")`
  - iii. `abline(b[1], b[3], col="blue")`  
`abline(b[2], b[4], col="red")`
  - iv. `abline(b[1], b[3], col="blue")`  
`abline(b[1] + b[2], b[3] + b[4], col="red")`
- (c) Which of the following assumptions is typically *least* important for the goal of fitting a regression model?

B.1. MULTIPLE-CHOICE QUESTIONS FOR THE FIRST SEMESTER

337

- i. Equal variance of errors
  - ii. Linearity
  - iii. Additivity
  - iv. Representativeness
- (d) Which of the following is *not* a possible plot of residuals from a fitted regression model?

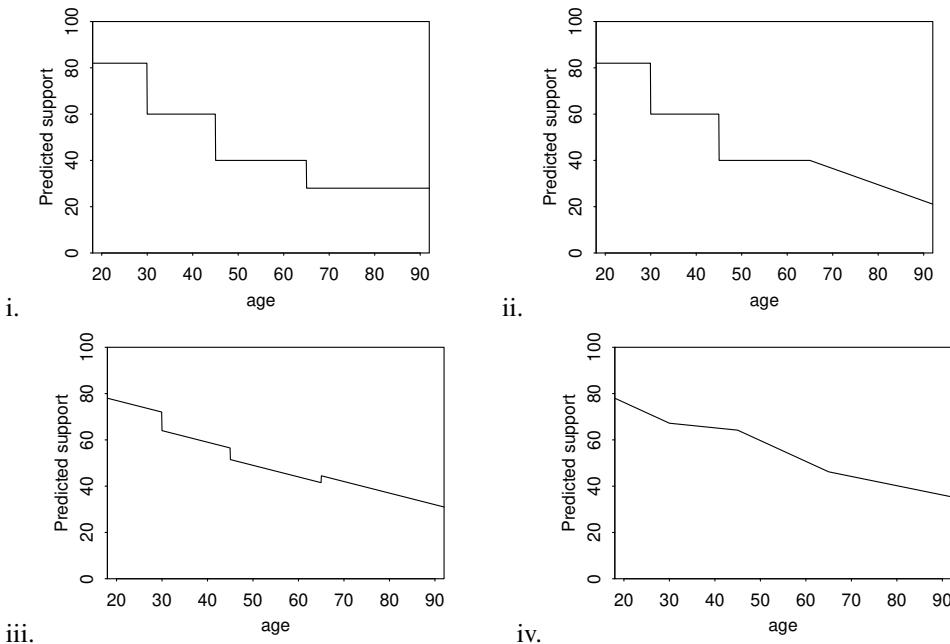


- (e) A researcher plans to fit a linear regression to data from a group of adults, predicting their physical flexibility given age. Flexibility is defined on a 0–30 scale based on measurements from a series of stretching tasks. Which of the following represents a concern with the *validity* assumption?
- i. The stretching measurements do not correspond to aspects of flexibility that are of real-world interest.
  - ii. The people in the study are volunteers and are healthier than the general population.
  - iii. Flexibility declines as a nonlinear function of age.
  - iv. Flexibility measurements are not close to normally distributed.

12. Questions for Chapter 12 of *Regression and Other Stories*

- (a) Data are collected on the size of the military in a large number of countries, and then a regression is fit,  $\log(\text{number of active-duty military (in millions)}) = b_0 + b_1 \log(\text{population (in millions)}) + b_2 x + \text{error}$ , where  $x$  is an indicator that equals 1 if a country is characterized as a democracy and 0 otherwise. Suppose the coefficient of  $b_1$  is 0.6. Which of the following is a correct interpretation of this coefficient?
- i. Increasing the log population of a country by 1 will on average increase its log military size by 0.6.
  - ii. Comparing two countries that differ by 1 million in population size, on average we expect the larger country to have about 600 000 more people in the military.
  - iii. Comparing two countries that differ by 1 million in population size and have the same democracy status, on average we expect the larger country to have about 600 000 more people in the military.
  - iv. Comparing two countries that differ by 10% in population size and are the same democracy status, on average we expect the larger country to have about a 6% larger military.

- (b) A linear regression is fit predicting support for marijuana legalization (on a 0–100 scale) given age. The model includes age in categories (under 30, 30–44, 45–64, 65+) and also age as a linear predictor. The estimated coefficient for age is negative. Which of the following could represent the resulting fitted curve of predicted support given age?



- (c) Data from a survey of adults from 2004 were used to predict support for same-sex marriage (coded as 1 if the respondent supported same-sex marriage or 0 for lack of support) as a function of age and sex. Here is the result of a linear regression using an indicator for each bin of age:

	Median	MAD_SD
(Intercept)	0.51	0.01
factor(age_discrete)(29,39]	-0.10	0.01
factor(age_discrete)(39,49]	-0.14	0.01
factor(age_discrete)(49,59]	-0.14	0.01
factor(age_discrete)(59,69]	-0.25	0.01
factor(age_discrete)(69,79]	-0.28	0.01
factor(age_discrete)(79,100]	-0.32	0.01
male	-0.10	0.01

Auxiliary parameter(s):

Median MAD\_SD  
sigma 0.03 0.00

What is the interpretation of the intercept in this regression?

- The predicted probability of support for same-sex marriage for a hypothetical person with age of 0 is 51%.
- The predicted probability of support for same-sex marriage for a hypothetical woman with age of 0 is 51%.
- The predicted probability proportion of support for same-sex marriage for a hypothetical woman between 18 and 29 is 51%.
- The difference in probability of support for same-sex marriage, comparing two people who differ by 1 on the intercept but are of identical age and sex, is 51%.

B.1. MULTIPLE-CHOICE QUESTIONS FOR THE FIRST SEMESTER

339

- (d) Here is the result of a linear regression predicting log world population given year divided by 1000:

	Median	MAD_SD
(Intercept)	18.3	0.5
year_1000	1.7	0.3

Auxiliary parameter(s):

	Median	MAD_SD
sigma	0.7	0.2

Which of these equations does *not* express the fitted model?

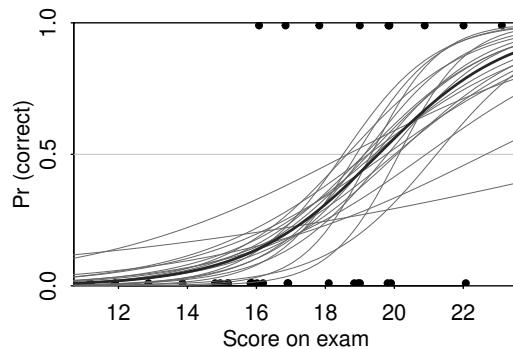
- i.  $\log(\text{population}) = 18.3 + 0.0017 * \text{year} + \text{error}$
  - ii.  $\text{population} = \exp(18.3 + 0.0017 * \text{year}) * \text{error}$
  - iii.  $\text{population} = \exp(18.3) * \text{year}^{0.0017} * \text{error}$
  - iv.  $\text{population} = \exp(18.3) * 1.0017^{\text{year}} * \text{error}$
- (e) The coefficients of a log-log model can also be called elasticities. What does it mean to estimate  $q$  as the elasticity of  $x$  on  $y$ ?
- i. For a 1 unit difference in  $x$ , we predict a  $q\%$  difference in  $y$ .
  - ii. For a 1 unit difference in  $x$ , we predict a  $q$  unit difference in  $y$ .
  - iii. For a 1% difference in  $x$ , we predict a  $q\%$  difference in  $y$ .
  - iv. For a 1% difference in  $x$ , we predict a  $q$  unit difference in  $y$ .
- (f) Here is a log-log model:  $\log y = 2 + 3 \log x + \text{error}$ . How can we represent the model in the untransformed scale?
- i.  $y = \exp(2) + x^3 + \text{error}$
  - ii.  $y = 2x^3 * \text{error}$
  - iii.  $y = \exp(2) * x^3 * \text{error}$
  - iv.  $y = \exp(2) * 3^x * \text{error}$

## B.2 Multiple-choice questions for the second semester

Here are a few multiple-choice questions we prepared for each chapter in the second half of *Regression and Other Stories*, representing a mix of conceptual, programming, and methods problems. As discussed in Section 1.5 of this book, we sample one question for each chapter to construct a practice final exam, then sample two questions for each chapter to form the exam itself.

### 1. Questions for Chapter 13 of *Regression and Other Stories*

- (a) Here are data on success or failure on a particular final exam question plotted vs. total score on the exam, along with a fitted logistic regression line,  $\Pr(y = 1) = \text{logit}^{-1}(a + bx)$ , and curves showing draws of its uncertainty:



What are the approximate intercept and slope of this logistic regression?

- i.  $a = 0, b = 0.05$
- ii.  $a = -2, b = 0.1$
- iii.  $a = -1.5, b = 0.1$
- iv.  $a = -10, b = 0.5$

- (b) Here is a fitted model predicting whether a person with high-arsenic drinking water will switch wells, given the arsenic level in their existing well and the distance to the nearest safe well:

```
stan_glm(formula = switch ~ dist100 + arsenic,
          family=binomial(link="logit"), data=wells)
          Median MAD_SD
(Intercept)    0.00    0.08
dist100       -0.90    0.10
arsenic        0.46    0.04
```

Compare two people who live the same distance from the nearest well but whose arsenic levels differ, with one person having an arsenic level of 0.5 and the other person having a level of 1.0. Give an approximate estimate  $\pm$  standard error for the difference in their probabilities of switching.

- i.  $0.11 \pm 0.01$
- ii.  $0.11 \pm 0.04$
- iii.  $0.05 \pm 0.005$
- iv.  $0.05 \pm 0.02$

- (c) Given the logistic regression model,  $\Pr(y = 1) = \text{logit}^{-1} = (-2.5 + 0.44x)$ , what is the maximum predictive difference corresponding to a unit difference in  $x$ ?

- i. 0.11
- ii. 0.22
- iii. 0.44
- iv. 1.76

- (d) When interpreting logistic regression coefficients, why is the divide-by-4 rule an upper bound approximation for the expected difference in  $y$  associated with a 1-unit difference in  $x$ ?
- It is accurate where  $\text{logit}^{-1}(a + bx) = 0$ , thus where the expected value of  $y$  is *minimized*.
  - It is accurate where  $\text{logit}^{-1}(a + bx) = 1$ , thus where the expected value of  $y$  is *maximized*.
  - It is accurate where  $\text{logit}^{-1}(a + bx) = 0.5$ , thus where the *slope* of the logistic curve is *maximized*.
  - It is accurate where  $\text{logit}^{-1}(a + bx) = 0.5$ , thus where the *intercept* of the logistic curve is *maximized*.

- (e) You first run this:

```
n_loop <- 100
est <- rep(NA, n_loop)
se <- rep(NA, n_loop)
for (loop in 1:n_loop){
  n <- 50
  x <- runif(n, 0, 10)
  y <- rbinom(n, 1, invlogit(-0.2 + 0.3*x))
  fake <- data.frame(x, y)
  fit <- stan_glm(y ~ x, family=binomial(link="logit"), data=fake)
  est[loop] <- coef(fit)[["x"]]
  se[loop] <- se(fit)[["x"]]
}
```

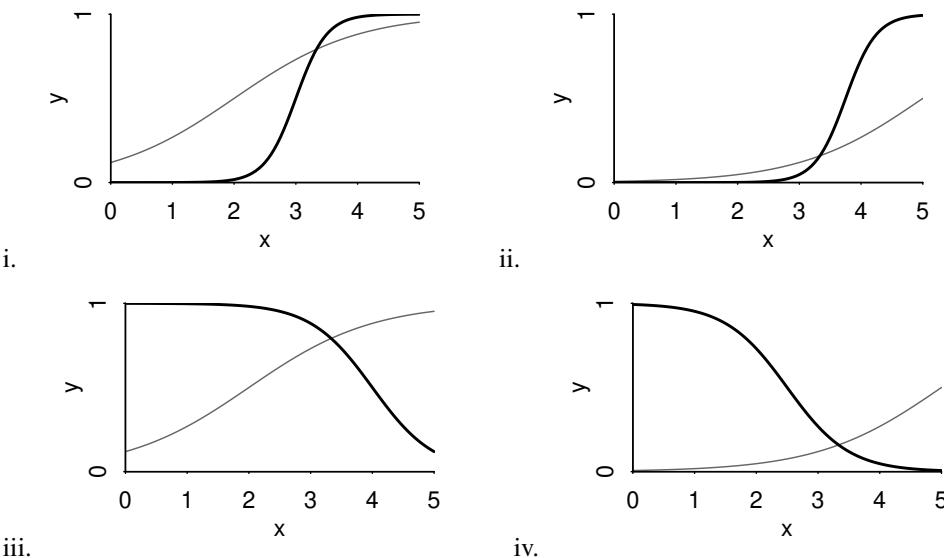
Which of the following lines of code, when run next, will be expected to give a value close to 0.95?

- $\text{mean}((\text{est} + 0.2)/\text{se})$
  - $\text{mean}((\text{est} - 0.3)/\text{se})$
  - $\text{mean}(\text{abs}(\text{est} - 0.3)/\text{se} < 2)$
  - $\text{mean}(\text{est} - 0.3)/\text{sd}(\text{est}) < 2$
- (f) A logistic regression is fit, using grade point average (on a 0–4 scale) to predict whether a student will drop out of college. Which of these is a plausible value for the coefficient of grade point average in this model?
- 1
  - 0
  - 0.1
  - 1

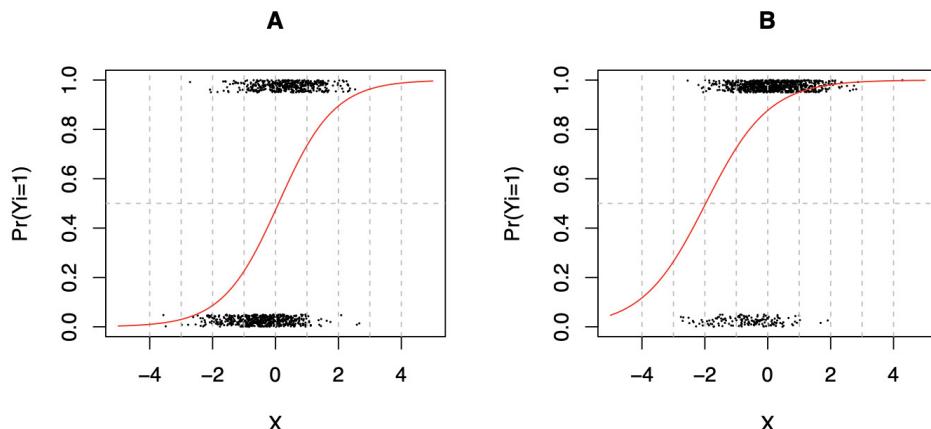
## 2. Questions for Chapter 14 of *Regression and Other Stories*

- (a) Consider the model,  $\Pr(y = 1) = \text{logit}^{-1}(0.1 + 0.2x + 0.3z - 0.4xz)$ , where  $y$  is an outcome variable,  $x$  is a pre-treatment variable, and  $z$  is a randomly assigned treatment. For what value of  $x$  does the treatment have no effect?
- 3/2
  - 1/2
  - 3/4
  - There is no value of  $x$  for which the treatment has no effect.
- (b) 100 people are surveyed and asked about their happiness. 75 say they are happy and 25 say they are not. A logistic regression is then fit, predicting this survey response on a constant term. Approximately what will be the estimated coefficient?
- 0.25
  - 0.8
  - 1.1
  - 3.0

- (c) Which of the following graphs represents the model,  $y = \text{logit}^{-1}(-2 + x - 10z + 3xz)$ , with light and dark lines corresponding to  $z = 0$  and 1, respectively?



- (d) The following two plots show two different logistic regression fits.



Which of the following statements is true?

- i. The intercept of the fitted logistic regression model in plot A is *greater* than the intercept of the fitted logistic regression model in plot B.
  - ii. The intercept of the fitted logistic regression model in plot A is *less* than the intercept of the fitted logistic regression model in plot B.
  - iii. The *slope* of the fitted logistic regression model in plot A is greater than the slope of the fitted logistic regression model in plot B.
  - iv. The data in plot B show imbalance so it is not appropriate to fit a logistic regression there.

(e) You perform the following simulation:

- (e) You perform the following simulation:

```
x <- seq(0, 1, length=50)
y <- c(rep(1,25), rep(0,25))
fake <- data.frame(x, y)
fit <- stan_glm(y ~ x, family=binomial(link="logit"), data=fake)
print(fit)
```

## What will you see?

- i. The estimated coefficient for  $x$  will be  $-\infty$ .
- ii. The estimated coefficient for  $x$  will be a large negative number.
- iii. The estimated coefficient for  $x$  will be a large positive number.
- iv. The estimated coefficient for  $x$  will be  $\infty$ .

3. Questions for Chapter 15 of *Regression and Other Stories*

- (a) Here is the result of a fitted logistic regression:

	Median	MAD_SD
(Intercept)	-0.31	0.13
x	0.44	0.65
z	0.10	0.05

What will approximately be the coefficient and standard error of the coefficient for  $x$  if you fit a probit regression to these data?

- i. The estimate is scaled by a factor of 1.6; thus,  $0.28 \pm 0.65$ .
  - ii. The estimate and standard error are scaled by a factor of 1.6; thus,  $0.28 \pm 0.40$ .
  - iii. There is no simple approximate rule mapping from logit to probit coefficients.
  - iv. It is not possible to fit a probit model to logistic data.
- (b) A negative binomial regression is fit to data from many country-years estimating the number of public protests in the country-year, given many predictors including per-capita GDP in dollars. The model is re-fit using, as a predictor, per-capita GDP in tens of thousands of dollars. What happens to the coefficient estimate and standard error?
- i. The estimate and standard error both are multiplied by 10 000.
  - ii. The estimate and standard error both are divided by 10 000.
  - iii. The estimate is multiplied by 10 000 and the standard error is multiplied by 100.
  - iv. The estimate is divided by 10 000 and the standard error is divided by 100.
- (c) Which of the following statements regarding Poisson and negative binomial regression is *not* correct?
- i. If you were going to fit a Poisson regression, you're generally better off fitting a negative binomial regression.
  - ii. If you include a variable as an offset, you cannot also include it as a predictor in the model.
  - iii. Poisson regression corresponds to a negative binomial regression with a very large value of the reciprocal overdispersion parameter.
  - iv. If you simulate fake data from one of these models and then fit that model and look at a particular coefficient, then there will be an approximate 68% chance that the estimated coefficient will be within 1 standard error of the true value.
- (d) Which of the following conditions must hold for a negative binomial regression to be equivalent to a Poisson regression?
- i. The reciprocal dispersion parameter is equal to 0.
  - ii. The reciprocal dispersion parameter is equal to 1.
  - iii. The predictive standard deviation of a data point is equal to its expected value.
  - iv. The predictive standard deviation of a data point is equal to the square root of its expected value.

4. Questions for Chapter 16 of *Regression and Other Stories*

- (a) A survey experiment is conducted in which each respondent is randomly assigned to read one of two vignettes—a neutral story or an anti-immigration story—and then is asked, “Do you support an increase in the country’s immigration levels?” Suppose the baseline support for this position is 30%. Which of following code computes the power for this experiment, assuming that the result will only be reported if it is statistically significant?

i.

```
power <- function(n, theta){  
  significant <- rep(NA, 1000)  
  for (i in 1:1000){  
    n0 <- rbinom(1, n, 0.5)  
    n1 <- n - n0  
    y0 <- rbinom(1, n0, 0.3)  
    y1 <- rbinom(1, n1, 0.3 + theta)  
    diff <- y1 - y0  
    se <- sqrt((y0/n0)*(1-y0/n0)*n0 + (y1/n1)*(1-y1/n1)*n1)  
    significant[i] <- abs(diff) > 2*se  
  }  
  mean(significant)  
}
```

ii.

```
power <- function(n, theta){  
  correct_sign <- rep(NA, 1000)  
  for (i in 1:1000){  
    n0 <- rbinom(1, n, 0.5)  
    n1 <- n - n0  
    y0 <- rbinom(1, n0, 0.3)  
    y1 <- rbinom(1, n1, 0.3 + theta)  
    diff <- y1 - y0  
    correct_sign[i] <- sign(diff) == sign(theta)  
  }  
  mean(correct_sign)  
}
```

iii.

```
power <- function(n, theta){  
  significant <- rep(NA, 1000)  
  for (i in 1:1000){  
    n0 <- rbinom(1, n, 0.5)  
    n1 <- n - n0  
    y0 <- rbinom(1, n0, 0.3)  
    y1 <- rbinom(1, n1, 0.3 + theta)  
    diff <- y1/n1 - y0/n0  
    se <- sqrt((y0/n0)*(1-y0/n0)/n0 + (y1/n1)*(1-y1/n1)/n1)  
    significant[i] <- abs(diff) > 2*se  
  }  
  mean(significant)  
}
```

iv.

```
power <- function(n, theta){  
  z_score <- rep(NA, 1000)  
  for (i in 1:1000){  
    n0 <- rbinom(1, n, 0.5)  
    n1 <- n - n0  
    y0 <- rbinom(1, n0, 0.3)  
    y1 <- rbinom(1, n1, 0.3 + theta)  
    diff <- y1/n1 - y0/n0  
    se <- sqrt((y0/n0)*(1-y0/n0)/n0 + (y1/n1)*(1-y1/n1)/n1)  
    z_score[i] <- diff/se  
  }  
  mean(abs(z_score))  
}
```

## B.2. MULTIPLE-CHOICE QUESTIONS FOR THE SECOND SEMESTER

345

- (b) You conduct an experiment in which half the people get a special get-out-the-vote message and the others do not. Then you follow up after the election with a new random sample of 500 people to see if they voted. What is the approximate standard error of your estimated effect size?
- 0.015
  - 0.03
  - 0.05
  - 0.07
- (c) A study is designed which would have 80% power if it had 600 participants. But for budgetary reasons, only 300 people could be included in the study. What is the approximate power of this new study?
- 30%
  - 40%
  - 50%
  - 60%
- (d) In a recent survey, 15% of Americans surveyed said they were crime victims in the past year. When the survey is done again in a year, how large a sample would be needed to estimate this proportion to within a standard error of 2 percentage points?
- $0.15 * 0.85 / 0.02^2$
  - $\sqrt{0.15 * 0.85 / 0.02}$
  - $0.15 * 0.85 * 0.02^2$
  - $\sqrt{0.15 * 0.85 * 0.02}$

### 5. Questions for Chapter 17 of *Regression and Other Stories*

- (a) You are creating a data set with two variables,  $x$  and  $y$ , simulated as follows:

```
n <- 100
x <- rnorm(n, 0, 1)
y <- 2 + 3*x + rnorm(n, 0, 2)
```

Which of the following code simulates a pattern of data that are missing at random?

- ```
is_missing_x <- rbinom(n, 1, 0.2) == 1
is_missing_y <- rbinom(n, 1, invlogit(x)) == 1
x[is_missing_x] <- NA
y[is_missing_y] <- NA
```
- ```
is_missing_y <- rbinom(n, 1, invlogit(x)) == 1
y[is_missing_y] <- NA
```
- ```
is_missing_x <- rbinom(n, 1, invlogit(x)) == 1
x[is_missing_x] <- NA
```
- ```
is_missing_y <- rbinom(n, 1, invlogit(y)) == 1
y[is_missing_y] <- NA
```

- (b) Consider the following regression and poststratification code:

```
fit <- stan_glm(vote ~ factor(partyid), data=poll)
poststrat_data <- data.frame(partyid=c("R","D","I"),
  N=c(0.31,0.32, 0.37))
predict <- posterior_linpred(fit, newdata=poststrat_data)
poststrat_est <- predict %*% poststrat_data$N / sum(poststrat_data$N)
print(c(median(poststrat_est), mad(poststrat_est)), digits=2)
```

What is wrong with this code?

- i. You need to regress on other variables—you cannot poststratify on the variable you regress on.
- ii. You need to use predict rather than posterior\_linpred because you need a point estimate without uncertainty.
- iii. You need to flip predict and poststrat\_data\$N/sum(poststrat\_data\$N) for the matrix multiplication %\*% to work.
- iv. Nothing is wrong with the code.

- (c) Here is a logistic regression fit predicting a binary response (“Are you satisfied with your educational experience?”) from a survey of students in a four-year college:

	Median	MAD_SD
(Intercept)	-1.6	0.7
male	-0.1	1.1
factor(year)2	0.5	0.9
factor(year)3	1.4	2.2
factor(year)4	2.6	2.3
male:factor(year)2	-0.3	1.4
male:factor(year)3	-1.5	2.3
male:factor(year)4	-1.8	2.5

You plan to estimate the average for all the students in the college by taking the predictions from the fitted model and poststratifying them on sex and year. To do so, what assumption are you implicitly making about the students in the survey?

- i. That they are a random sample of the students in the college
- ii. That the proportion of men and women and the proportion of students in each year in the sample are close to the corresponding proportions in the college
- iii. That, within each of the categories defined by sex and year, they are a random sample from the corresponding group in the college
- iv. That there are no important predictors of the response that have not been included in this model

- (d) In studying a national survey of smoking among high school students, a researcher writes the following code to estimate the average smoking rate in the population:

```
fit <- stan_glm(smoking ~ female*(factor(ethnicity) + factor(age) +
  parents_SES + parents_smoking), family=binomial(link="logit"),
  data=survey)
epred_fit <- posterior_epred(fit, newdata=population)
avg_pop <- epred_fit %*% population$N / sum(population$N)
print(c(mean(avg_pop), sd(avg_pop)))
```

What is wrong with the above code?

- i. When using logistic regression, you should not poststratify by simply averaging; instead you need to account for the nonlinearity in the prediction.
- ii. Because female is interacted with the other predictors, you should not simply predict and poststratify all at once; instead you need to make separate predictions for male and female students and then average these predictions at the end.
- iii. Because you already are starting from a national survey, you do not need to do any poststratification; you should instead estimate the smoking rate directly from the sample.
- iv. Nothing is wrong with the code.

## 6. Questions for Chapter 18 of *Regression and Other Stories*

- (a) An experiment is performed estimating the effect of a treatment on a sample of people who are 45% women and 55% men, and the goal is to estimate the average effect in a population that is 52% women and 48% men. Which of the following statements is not necessarily true?

## B.2. MULTIPLE-CHOICE QUESTIONS FOR THE SECOND SEMESTER

347

- i. The sample average treatment effect is between the conditional average treatment effect for women and the conditional average treatment effect for men.
  - ii. The population average treatment effect is between the conditional average treatment effect for women and the conditional average treatment effect for men.
  - iii. If the treatment is assigned completely at random, then you can get an unbiased estimate of the sample average treatment effect.
  - iv. If the people in the experiment are a simple random sample from the population, then you can get an unbiased estimate of the population average treatment effect.
- (b) A researcher conducts a randomized experiment and estimates the treatment effect by comparing the average outcome in the treatment group to the average outcome in the control group. The resulting standard error is extremely high. Which of the statement is not necessarily true?
- i. A regression conditioning on pre-treatment variables that are highly predictive of the outcome should reduce the standard error.
  - ii. The estimate is unbiased.
  - iii. Increasing the sample size should reduce the standard error.
  - iv. The randomization was not implemented properly.
- (c) Consider the following small experiment:

Unit $i$	Female, $x_{1i}$	Age, $x_{2i}$	Treatment, $z_i$	Outcome, $y_i$
Audrey	1	40	0	140
Anna	1	40	1	140
Bob	0	50	1	150
Bill	0	50	0	150
Caitlin	1	60	0	155
Cara	1	60	0	155
Dave	0	70	1	160
Doug	0	70	1	160

Suppose the treatment effect is  $-5$  for men and  $-10$  for women. What is the sample average treatment effect for the 8 people in the experiment?

- i.  $-2.5$
  - ii.  $-5$
  - iii.  $-7.5$
  - iv. It cannot be determined from the information given.
- (d) Consider the following small experimental setup including potential outcomes:

Unit $i$	Female, $x_{1i}$	Age, $x_{2i}$	Treatment, $z_i$	Potential outcomes	
				$y_i^0$	$y_i^1$
Audrey	1	40	0	4	4
Anna	1	40	0	4	4
Bob	0	50	0	5	6
Bill	0	50	0	5	6
Caitlin	1	60	1	6	6
Cara	1	60	1	6	6
Dave	0	70	1	7	8
Doug	0	70	1	7	8

What is the estimated treatment effect from these data?

- i.  $0.5$
- ii.  $2.0$
- iii.  $2.5$
- iv. It cannot be determined from the information given.

- (e) A researcher conducts a randomized experiment and estimates the treatment effect by regressing the outcome on a treatment indicator and several pre-treatment predictors. Which of the following is estimated by the coefficient of the treatment indicator?
- The sample average treatment effect (SATE)
  - The population average treatment effect (PATE)
  - The conditional average treatment effect (CATE)
  - The average treatment effect among the treated (ATT)
- (f) Which experiment is likely to violate the stable unit treatment value (SUTVA) assumption?
- A field experiment that randomizes students within a school to test the effect of healthy snacks on health outcomes
  - An online survey experiment that randomizes participants to test the effect of reading fake news on policy support
  - A lab experiment that randomizes participants to test the effect of a new drug
  - None of the above
- (g) Which of the following is a property of a randomized block design?
- The same proportion of units is treated in each block.
  - Each block has the same probability of treatment as all other blocks.
  - Each unit has the same probability of being treated as all other units in its block.
  - Every unit has the same probability of being treated as all other units in the sample.

7. Questions for Chapter 19 of *Regression and Other Stories*

- (a) The following model is fit to data from a randomized experiment on a group of students:

	Median	MAD_SD
(Intercept)	39.41	4.90
treatment	11.61	7.98
pre_test	0.68	0.05
treatment:pre_test	-0.09	0.07

Auxiliary parameter(s):

Median	MAD_SD
sigma	2.17 0.25

Let  $\bar{x}$  be the average value of the pre-test for the students in the experiment and  $\bar{X}$  be the average value of the pre-test in the population. Pre-test scores fall in the range from 0 to 100. Which of the following statements is *incorrect*?

- The treatment is estimated to be more effective for students with lower pre-test scores.
  - The estimated treatment effect is positive for some students and negative for others.
  - The estimated sample average treatment effect is  $11.61 - 0.09\bar{x}$ .
  - The estimated population average treatment effect is  $11.61 - 0.09\bar{X}$ .
- (b) An experiment is performed in which children are randomly assigned at age 10 to a growth-enhancing drug or a placebo. The goal is to estimate the effect of the drug on adult height, and the height of participants is measured at ages 6, 8, 12, and adulthood. The treatment effect is estimated as follows:  $\text{adult\_height} \sim z + \text{height\_6} + \text{height\_8} + \text{height\_12}$ . What is the most important problem with this estimate?
- Height at age 12 is a post-treatment variable so you should not adjust for it.
  - Heights at age 6 and 8 will be nearly collinear, so you should not include both of these in the regression.
  - Given the importance of pre-treatment height, you should include an interaction of this with the treatment indicator.
  - It makes sense to consider the effect as multiplicative so you should fit the model on the log scale.

## B.2. MULTIPLE-CHOICE QUESTIONS FOR THE SECOND SEMESTER

349

- (c) An experiment is performed on a treatment intended to improve college admissions test scores. Every student in the experiment has already taken the test once, and for each student we have this pre-test score, a measure of socioeconomic status, a randomized treatment assignment, the post-test score, and the gain score (post-test minus pre-test). Two models are fit:
- (1)  $\text{post\_test} \sim \text{pre\_test} + \text{SES} + z$ ,
  - (2)  $\text{gain\_score} \sim \text{pre\_test} + \text{SES} + z$
- Which of the following statements is correct about the estimated treatment effect under the two models?
- i. Model 1 is better because it is predicting the outcome directly.
  - ii. Model 2 is better because the goal is to estimate improvement.
  - iii. Which model is better depends on the data.
  - iv. The two models are the same.
- (d) Assume you have an unbiased estimate of the sample average treatment effect (SATE). Which of these would make it an unbiased estimate of the population average treatment effect (PATE)?
- i. Ignorability of treatment assignment
  - ii. Sample is representative of the population
  - iii. Stable unit treatment value assumption
  - iv. Efficiency property

### 8. Questions for Chapter 20 of *Regression and Other Stories*

- (a) An observational study is simulated using the following code for pre-test  $x$ , treatment  $z$ , and post-test  $y$ :

```
n <- 100
x <- runif(n, -1, 1)
z <- rbinom(n, 1, invlogit(x))
y <- 0.2 + 0.3*x + 0.5*z + rnorm(n, 0, 0.4)
fake <- data.frame(x, y, z)
fit <- stan_glm(y ~ x + z, data=fake, refresh=0)
estimate <- coef(fit)["z"]
```

In this simulation, the true treatment effect is 0.5. Which of the following statements is correct?

- i. The estimate will probably be less than 0.5 because the model also adjusts for  $x$ , which is positively correlated with  $z$ , and this adjustment for  $x$  will suck up some of the explanatory power of  $x$ .
- ii. The estimate will probably be greater than 0.5 because there is imbalance in the treatment assignment: the treatment is more likely to be assigned to people with higher pre-test scores, which will artificially make the treatment look more effective.
- iii. The estimate will probably be greater than 0.5 because, given the finite sample size, the adjustment for  $x$  is noisy and is likely to undercorrect for imbalance in the design.
- iv. The estimate is unbiased because the model correctly adjusts for differences in pre-test between treatment and control groups.

- (b) Consider the following simulation of pre-test  $x$  and treatment  $z$ :

```
n <- 100
x <- runif(n, -2, 2)
z <- rbinom(n, 1, invlogit(2 + 3*x))
```

How would you characterize this assignment rule?

- i. No problem with balance or overlap
- ii. Problem with balance; no problem with overlap
- iii. No problem with balance; problem with overlap
- iv. Problem with balance; problem with overlap

- (c) Which of the following models, predicting outcome  $y$  from pre-treatment predictor  $x$  and treatment indicator  $z$ , corresponds to a constant treatment effect?
- $y \sim x + z$
  - $y \sim x + z + x:z$
  - $y \sim x + z$ , family=binomial(link="logit")
  - $y \sim x + z + x:z$ , family=binomial(link="logit")
- (d) The following model is fit to data from an observational study with pre-treatment variables  $u$  and  $v$ , treatment indicator  $z$ , and outcome  $y$ :  $y = b_0 + b_1u + b_2v + b_3z + b_4uv + b_5uz + \text{error}$ . Assuming ignorability of treatment assignment, what is the estimated treatment effect?
- $b_3$
  - $b_3 + b_5u$
  - $b_3z + b_5uz$
  - $b_3z + b_4uv + b_5uz$
- (e) Consider a propensity score used for matching. What does this propensity score represent?
- The estimated probability that a unit receives treatment, given pre-treatment predictors
  - The estimated probability that a unit is in an area of overlap, given pre-treatment predictors
  - The estimated probability that a unit is in an area of balance, given pre-treatment predictors
  - The estimated probability that a unit is representative, given pre-treatment predictors

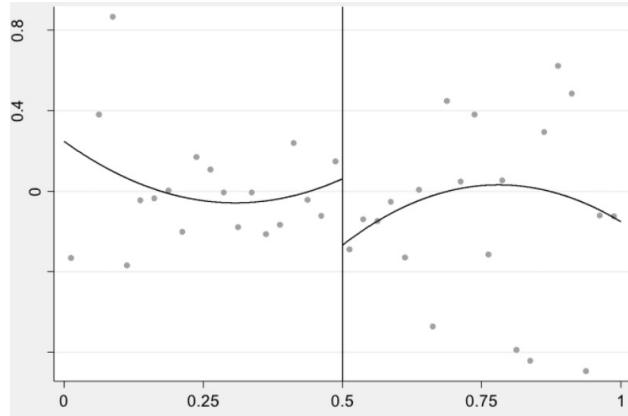
9. Questions for Chapter 21 of *Regression and Other Stories*

- (a) In a simple instrumental variables problem, there are four compliance types (Compliers, Always Takers, Never Takers, and Defiers). If the standard assumptions for instrumental variables estimation hold, which of the following statements must be true?
- People are randomly assigned to the Always Takers and Never Takers groups.
  - The expected number of Compliers equals the expected number of Defiers.
  - Compliers and Always Takers will be approximately balanced in any pre-treatment predictors.
  - There are no Defiers.
- (b) An experiment is performed in which students are randomly assigned to participate or not participate in a study program. For each student, the following variables are recorded: a pre-test score  $x$ , an encouragement indicator  $z$ , a participation indicator  $T$ , and a post-test score  $y$ . Of the following assumptions, which is *not* required for instrumental variables estimation?
- The effect of encouragement on participation for individual students is sometimes positive.
  - The effect of encouragement on participation for individual students is never negative.
  - Encouragement can affect post-test scores only for those students where it has a nonzero effect on participation.
  - The effect of the encouragement treatment cannot depend on the pre-test score.
- (c) The following code is intended to simulate a scenario where an instrumental variables analysis is appropriate; we simulate an observed pre-treatment predictor  $x$ , an unobserved pre-treatment predictor  $u$ , an encouragement  $z$ , a treatment  $T$ , and an outcome  $y$ :

```
n <- 1000
x <- runif(n, -2, 2)
u <- runif(n, -1, 1)
z <- rbinom(n, 1, 0.5)
T <- rbinom(n, 1, invlogit(u + z))
y <- 0.2 + 0.3*x + 0.4*u + 0.5*T + rnorm(n, 0, 0.2)
data <- data.frame(x, z, T, y)
fit_1 <- stan_glm(T ~ x + z, data=data, refresh=0)
fit_2 <- stan_glm(y ~ x + T + z, data=data, refresh=0)
iv_estimate <- coef(fit_2)[["z"]] / coef(fit_1)[["z"]]
```

What is wrong with the above code?

- i. The models adjust for  $x$ ; it is not appropriate in instrumental variable estimation to adjust for a pre-treatment predictor.
  - ii. The simulation violates the exclusion restriction.
  - iii. We have included  $u$  in the simulation but not in the regression so this cannot work.
  - iv. The model `fit_2` should not include  $T$  as a predictor.
- (d) A researcher is interested in estimating the effect of campaign spending following a new policy that has been enacted in some states but not others. The researcher gathers data on campaign spending and voter turnout by state for each of several elections before and after the policy has been implemented. Which of the following analyses would *not* be a reasonable way of studying this policy?
- i. A regression on outcomes at the state-year level, including state-level policy and adjusting for state-level indicators (fixed effects)
  - ii. An instrumental variables analysis, considering state-level policy change as the instrument and the amount of spending in the state post-policy as the treatment
  - iii. A regression discontinuity analysis, considering time as the forcing variable
  - iv. A difference-in-difference analysis, comparing states with and without the policy change, before and after the year the policy was implemented
- (e) A team of researchers performed a regression discontinuity (RD) study purporting to demonstrate that unionization reduces the probability of a company's stock price crashing. The analysis looked at a number of companies that had experienced labor union elections which followed the rule that unionization happened if the union received at least half the vote, and then used this variable to predict certain outcomes related to stock crashes. Here is one of the fitted models showing binned data averages:



The estimated treatment effect was negative and statistically significant, leading the researchers to claim that unionization leads to a decline in crash risk. Which of the following statements is *not* correct?

- i. The difference in the  $y$ -values between the two curves at  $x = 0.5$  is the estimated treatment effect from the RD analysis.
  - ii. A problem with this RD analysis is that the fitted quadratic curves are not plausible.
  - iii. This fails as an RD design because there is no overlap between the treatment and control groups.
  - iv. The analysis could be more trustworthy if it were performed on a narrow band of  $x$  near 0.5, but then the standard error would go up.
- (f) When interpreting an analysis using instrumental variables, when would the intent to treat effect (ITT) be a more useful estimand than the conditional average treatment effect (CATE)?

- i. When you want to know the aggregate effect of a treatment across compliers and non-compliers
  - ii. When you want to isolate the causal effect for compliers
  - iii. When the monotonicity assumption is violated
  - iv. When the exclusion restriction is violated
10. Questions for Chapter 22 of *Regression and Other Stories*
- (a) A researcher would like to measure the effect of a drug that is intended to reduce blood pressure. Let  $x$  be a patient's pre-treatment blood pressure,  $u$  be the concentration of the drug in the patient's blood, and  $y$  be post-treatment blood pressure. Suppose these data, if available, could be well fit by the model  $y = a_0 + a_1x + a_2u + \text{error}$ . However, in this experiment,  $u$  is not known, it is only measured with error. So instead the researcher fits the model,  $y = b_0 + b_1x + b_2v + \text{error}$ , where  $v$  is the noisy measure of drug concentration. What can we typically say about the treatment effect?
    - i.  $a_2$  and  $b_2$  will be about the same.
    - ii.  $a_2$  will be higher in absolute value than  $b_2$ .
    - iii.  $a_2$  will be lower in absolute value than  $b_2$ .
    - iv. At least one of  $a_2$  or  $b_2$  will be approximately zero.
  - (b) In which of the following models is  $y$  an increasing function of  $x$  with an asymptotic threshold?
    - i.  $y = a - bx - cx^2$ , where  $b$  and  $c$  are positive
    - ii.  $y = 1/(a + bx)$ , where  $b$  is positive
    - iii.  $y = a * \log(-bx)$ , where  $b$  is positive
    - iv.  $y = a - b * \exp(-cx)$ , where  $b$  and  $c$  are positive
11. Review questions for the first half of *Regression and Other Stories*
- (a) Data are collected on each of 50 states for 10 years, and a regression is fit predicting crime rate given a continuous measure of state policy. The model looks like this:

```
crime ~ policy + year + factor(state) + year*factor(state)
```

How many coefficients are in this model?
    - i. 53
    - ii. 101
    - iii. 493
    - iv. 562
  - (b) 200 women and 200 men are interviewed in a survey and asked their opinion on the death penalty. 60% of the women surveyed and 70% of the men surveyed support the death penalty. A linear regression is fit, predicting the survey response given a predictor, `sex`, that equals 1 for women and 2 for men. Approximately what will be the estimate and standard error for the coefficient of `sex`?
    - i.  $0.05 \pm 0.05$
    - ii.  $0.05 \pm 0.10$
    - iii.  $0.10 \pm 0.05$
    - iv.  $0.10 \pm 0.10$
12. Review questions for the second half of *Regression and Other Stories*
- (a) In order to estimate the effect of an intervention designed to increase the rate at which its employees use its health plan's counseling services, a company performs an experiment on 500 of its employees, in which 250 are assigned the treatment and 250 are not. They use the following code to estimate the treatment effect:

```
fit <- stan_glm(counseling_use ~ z + previous_counseling_use +
  female + age + employee_rank + z*previous_counseling_use, data=expt)
```

B.2. MULTIPLE-CHOICE QUESTIONS FOR THE SECOND SEMESTER

353

Which of the following code gives the estimate and standard error of the population average treatment effect (PATE)?

- i.  

```
epred_fit <- posterior_epred(fit, newdata=company)
avg_pop <- epred_fit %*% company$N / sum(company$N)
print(c(mean(avg_pop), sd(avg_pop)))
```
  - ii.  

```
effect_pop <- posterior_epred(fit, newdata=company)
print(c(mean(effect_pop), sd(effect_pop)))
```
  - iii.  

```
effect_pop <- posterior_epred(fit, newdata=data.frame(company, z=1) -
posterior_epred(fit, newdata=data.frame(company, z=0)
c(mean(effect_pop), sd(effect_pop))
```
  - iv.  

```
c(coef(fit)["z"], se(fit)["z"])
```
- (b) A campaign is studied that is intended to increase blood donation. The outcome measurement is the number of times that a person gives blood in the forthcoming year. The campaign is estimated to have a positive effect of 0.2. Which of the following stories is *not* consistent with this information?
- i. The campaign increases each person's response by 0.2 units.
  - ii. For each person, the campaign has a 20% chance of increasing the response by 1 unit.
  - iii. The campaign increases the response by 1 unit for 20% of the people and leaves it unchanged for the other 80%.
  - iv. The campaign increases the response by 1 unit for 40% of the people, decreases it by 1 unit for 20% of the people, and leaves it unchanged for the remaining 40%.
- (c) 1000 people in an experiment are randomly assigned to read two different vignettes and are then asked their opinion about a particular political issue. Assume a binary response, that is, a yes/no answer. Of the 500 who receive vignette A, 300 respond Yes to the issue question and 200 respond No. Of the 500 who receive vignette B, 250 respond Yes to the issue question and 250 respond No. A logistic regression is fit, predicting the survey response given a treatment indicator. Approximately what will be the estimated coefficient for the treatment indicator?
- i. -0.025
  - ii. -0.1
  - iii. -0.4
  - iv. -0.5

### B.3 Take-home exam

An alternative form of final exam is structured by giving students a series of tasks to be applied to a single example, going through all the materials in the course. Here are instructions for such an exam.

We will send you a topic and a dataset. You will have 48 hours to do this exam. The exam will be handed out at 8am. Submission deadline is two days later at 8am. You should be able to do it in one day; we're giving you two days so that you can do the exam on day 1, get a good sleep, and then go over your answers on day 2 before submitting.

Select questions that are jointly worth 130 points.

At the top of your submission you should indicate which questions you intend to address with the data. Answer in full sentences. Your submission should be a Rmarkdown file rendered as a pdf. Run `stan_glm` using the `refresh=0` setting to suppress intermediate output. Please also provide your code in a separate R file. The R file should include all commands from the Rmarkdown file and should run independently. It should include library calls and the data should be loaded in the environment from the same folder as the R file is in.

Answer all questions which you have chosen below in the context of this applied problem and the data you have downloaded.

1. Chapter 1 of *Regression and Other Stories* (10 points):

- Discuss the challenges in this example of generalizing from sample to population, from treatment to control group, and from observed measurements to the underlying construct of interest. Give two sentences for each.

2. Chapter 2 of *Regression and Other Stories* (10 points):

- Discuss issues of reliability and validity for this example, giving two sentences for each.
- Make a grid of graphs exploring your data, along with a paragraph describing what you have learned. Add another paragraph explaining how you have used the principles of statistical graphics in making your plots, and another paragraph discussing what important aspects of the data you were not able to include in these graphs. Your graphs should clearly communicate key features of the data. Use base graphics via `par(mfrow())` or `ggplot` via `facet_wrap/facet_grid`.

3. Chapter 3 of *Regression and Other Stories* (10 points):

- Consider a deterministic model on the linear or logarithmic scale that would arise here. Graph the model and discuss its relevance to this example.

4. Chapter 4 of *Regression and Other Stories* (10 points):

- Perform a simple comparison of treated vs. control or exposed vs. unexposed in your data. Compute the standard error and 95% confidence interval for this comparison.
- Discuss a possible source of bias or unmodeled uncertainty and estimate its magnitude, in comparison to the standard error you just calculated. Give a sentence discussing the relative importance of the bias or unmodeled uncertainty, compared to the uncertainty in the comparison from the data.

5. Chapter 5 of *Regression and Other Stories* (20 points):

- Construct a probability model—a function in R that first generates unobserved data and then generates potentially observed data, possibly with measurement error and selection bias. The process should be relevant to your applied problem but does not need to capture all its complexity. Graph your simulated data, compare to a graph of real data, and discuss the connections between your model and your larger substantive questions.

6. Chapter 6 of *Regression and Other Stories* (10 points):

### B.3. TAKE-HOME EXAM

355

- Take two variables that represent before-and-after measurements of some sort. Make a scatterplot and discuss challenges of “regression to the mean” when interpreting before-after changes here.
7. Chapter 7 of *Regression and Other Stories* (10 points):
- Fit a linear regression with a single predictor, graph the data along with the fitted line, and interpret the estimated parameters and their uncertainties. Write three sentences, one for each parameter in the model.
8. Chapter 8 of *Regression and Other Stories* (10 points):
- Use `lm` to fit a regression model with one predictor using least squares. Perform a calculation in R to demonstrate that the estimated slope equals a weighted average of slopes from all pairs of points.
9. Chapter 9 of *Regression and Other Stories* (10 points):
- Fit a linear regression to your data and use `predict`, `posterior_epred`, and `posterior_predict` to make predictions for a new data point with the predictor set to a reasonable value. In three sentences, summarize the results from these predictions and explain how they differ.
10. Chapter 10 of *Regression and Other Stories* (10 points):
- Fit a linear regression with multiple predictors including at least one interaction. The model should make sense; that is, there should be a good applied reason for fitting it. Explain each of the estimated parameters and their uncertainties, using one sentence for each parameter.
11. Chapter 11 of *Regression and Other Stories* (20 points):
- Fit a linear regression with multiple predictors. List all six of the assumptions of the model and explain, in one sentence for each, how these are relevant to this example.
  - Fit a linear regression with multiple predictors. Use residual plots (at least two) to assess the fit, and describe in two sentences what you found.
  - Fit two different linear regressions, each with multiple predictors. Both models should make sense; that is, there should be good applied reasons for fitting them. Compare the fits using leave-one-out cross validation, and in one sentence discuss what you found.
12. Chapter 12 of *Regression and Other Stories* (15 points):
- Fit a linear regression including a log transformation of predictors, outcome, or both. The model should make sense in the applied context. Graph the data and fitted model, and in a few sentences explain the result including the transformation.
  - Fit a linear regression with multiple predictors. Take one of the continuous predictors, bin it into discrete categories, and use these discrete indicators as predictors. Plot the data and both fitted models on the same graph.
13. Chapter 13 of *Regression and Other Stories* (10 points):
- Fit a logistic regression that makes sense in the applied context. Plot the data and fitted regression line. Interpret the coefficients and their uncertainties on the probability scale using the divide-by-4 rule.
14. Chapter 14 of *Regression and Other Stories* (20 points):
- Evaluate logistic regression using fake-data simulation. Check that the coefficients parameter estimates are approximately unbiased and that 95% intervals have approximate 95% coverage. Your simulation should be realistic in that the simulated data should be similar to the real data you are working with.

15. Chapter 15 of *Regression and Other Stories* (10 points):

- Fit a negative binomial regression that makes sense in the applied context. Plot the data and fitted regression line. Interpret the coefficients and their uncertainties.

16. Chapter 16 of *Regression and Other Stories* (15 points):

- Design a new study, including decisions about measurements and sample size. Use existing data and knowledge to come up with reasonable assumptions about effect size and variation. Then simulate fake data from this design, analyze the fake data, and discuss the results.

17. Chapter 17 of *Regression and Other Stories* (15 points):

- Consider how to generalize inferences from your fitted model to new scenarios. What information would be required for poststratification? Make some reasonable assumptions and use these to generalize your inferences to a relevant population.
- If your data have missingness, come up with a procedure to randomly impute the missing data and apply it to your example.

18. Chapter 18 of *Regression and Other Stories* (10 points):

- Frame a question of interest in terms of the effect of a binary treatment. For this example, explain what are the outcome variable  $y$ , the treatment indicator  $z$ , the pre-treatment variables  $x$ , and the potential outcomes  $y^0$  and  $y^1$ . Be precise.

19. Chapter 19 of *Regression and Other Stories* (10 points):

- Estimate a causal effect by linear or logistic regression of an outcome on a binary treatment variable, some pre-treatment predictors, and at least one treatment interaction. Interpret the coefficient estimates and standard errors from your fitted model. What are the assumptions required for you to interpret the coefficients on the treatment indicator as causal effects? Are these assumptions reasonable here?

20. Chapter 20 of *Regression and Other Stories* (20 points):

- Set up a causal inference problem with your data where you have an outcome variable, a binary treatment indicator, and some pre-treatment predictors, where there is noticeable lack of overlap, comparing treatment and control groups. Assess balance and overlap with graphs. Perform matching to get a subset of data with adequate overlap, and then fit a regression model on the subset. Discuss your results and their interpretation. In particular, how is your interpretation affected by the fact that you've analyzed only this subset?

21. Chapter 21 of *Regression and Other Stories* (20 points, either/or):

- Clearly define a causal effect of interest and perform an instrumental variables analysis using your data using the two regressions. Interpret each regression result along with the instrumental variables estimate. Compare to the result of a direct regression. In the context of this example, discuss the assumptions under which the direct and instrumental-variables regressions give good estimates of the causal effect.
- Perform a regression discontinuity analysis using your data. Discuss the choices involved in setting up this regression. Include other pre-treatment predictors in your model, not just the assignment variable. Interpret your result and discuss the assumptions required for it to represent a good causal inference.

This book has been published by Cambridge University Press as Active Statistics by Andrew Gelman and Aki Vehtari. This PDF is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.  
© Copyright by Andrew Gelman and Aki Vehtari 2024. The book web page <https://avehtari.github.io/ActiveStatistics/>

---

## **Appendix C**

# **Outlines of classroom activities**

---

## C.1 First semester

Week	Stories	Activities	Computer demonstrations	Discussion problems
1. Introduction to quantitative social science	Wikipedia experiment	Design a study to measure some quantity of interest	Collect and analyze simulated data	Find the hidden assumption
	Literary Digest poll of 1936	Design experiment to distinguish two hypotheses	Predict elections from economy	Find the hidden assumption
2. Overview of applied regression (Chapter 1 of <i>Regression and Other Stories</i> )	United Nations peacekeeping	Bag of candies and sampling bias	Graph of data and fitted line	Height and earnings
	Girls and sports	Gather and plot data from students	Tinker with an example	Graph hypothetical data
3. Data collection and visualization (Chapter 2 of <i>ROS</i> )	Political leanings of sports fans	Measure handedness	Download and work with data	Tell stories with graphs
	Use comparisons to redraw a graph	Scatterplot charades	Make plots clearer	Plots of baby names
4. Basics of math and probability (Chapter 3 of <i>ROS</i> )	Death rate in the pandemic	Amoebas and exponential growth and decline	Matrix manipulations	College admissions
	Galton's giants	Squares, cubes, and power-law growth	Compute weighted averages	Probability of a rare event
5. Statistical inference (Chapter 4 of <i>ROS</i> )	They got the wrong standard error	Design a bogus study	Simulate fake data and conf interval	Confidence intervals and true values
	Claims of implausibly large effects	Think about effect sizes	Proportions, means, and differences	Standard error for feeling thermometers
6. Simulation (Chapter 5 of <i>ROS</i> )	Proportion of identical twins	Real vs. fake coin flips	Break R functions	Discrete / continuous distribution
	Simulate a process of innovation	Simulate a probability process	Simulate 100 coin flips	Simulate clustering of buses

Week	Stories	Activities	Computer demonstrations	Discussion problems
7. Background on regression modeling (Chapter 6 of <i>ROS</i> )	Slope when predicting elections from the economy	Simulate fake data and fit a regression	Play with a simulated regression	Examples of regression to the mean
	Clinton/Trump vote vs. polls, and predictions	Memory quiz and regression to the mean	Challenges in setting up a simulation	Uniform partisan swing
8. Linear regression with a single predictor (Chapter 7 of <i>ROS</i> )	$5^2 + 12^2 = 13^2$	African countries in the U.N.	Regression, transformations, and sample size	Predict elections from incumbency
	Interpret the regression of earnings on height	Socioeconomic status and political views	Take average or regress on a constant term	How large was the sample size?
9. Fitting regression models (Chapter 8 of <i>ROS</i> )	Ronald Reagan and the evangelical vote	Simulate and recover regression lines	Play with the regression estimate	From inference to decision
	Does having a girl make you more conservative/liberal?	Move a point and shift the regression line	Compare <code>lm</code> and <code>stan_glm</code>	Sample size and statistical significance
10. Prediction and Bayesian inference (Chapter 9 of <i>ROS</i> )	Fairness of random exams	Coverage of prediction intervals	Different forms of predictive uncertainty	Coverage of prediction intervals
	Uncertainties in election forecasts	Prior for a real-world parameter	Bayes estimate of childhood intervention	Prior for a real-world parameter
11. Linear regression with multiple predictors (Chapter 10 of <i>ROS</i> )	Incumbency advantage in elections	Memory quiz with pre-test and treatment	Regression with interactions	Regression adjustment
	Beauty and teaching evaluations	Design a study with regression in mind	Adding interactions to a model	Why look at a pre-test?
12. Assumptions, diagnostics, evaluation (Chapter 11 of <i>ROS</i> )	Actual vs. guessed exam scores	Sample size and statistical significance	Take difference or regress on an indicator	Assumptions of regression
	Model checking for baseball analytics	Assumptions of regression	Simulate and debug	Patterns of residuals
13. Transformations and regression (Chapter 12 of <i>ROS</i> )	Logarithm of world population	Predictive uncertainties	Centered and standardized predictors	When to use the log scale
	Price elasticity of demand	Combining predictors to create a score	Regressions with logged variables	Straight line fit to non-linear data

## C.2 Second semester

Week	Stories	Activities	Computer demonstrations	Discussion problems
14. Review of statistics and regression (Chapters 1–12 of <i>Regression and Other Stories</i> )	Biased samples and coverage of intervals	Self-selected treatment assignment	Causal inference adjusting for pre-treatment	Sampling and adjustment
	The problem of too much talent?	Design a study to explore nonlinearity	Simulating patterns of bias	Causal inference, adjustment
15. Logistic regression (Chapter 13 of <i>ROS</i> )	Item-response analysis of final exams	“Two truths and a lie” game	Displaying a logistic curve	Real-world logistic regression
	Survey nonresponse	Predict the views of others	Logistic regression probabilities	Where logistic regression makes no sense
16. Working with logistic regression (Chapter 14 of <i>ROS</i> )	“Keys to the White House”	Job training and predictive comparisons	Predictions from logistic regression	Experimental design
	Opiate of the masses	Logistic regression with interactions	Linear or logistic regression	Design with pre-test
17. Other generalized linear models (Chapter 15 of <i>ROS</i> )	Patterns of gun ownership	How similar are you to your friends?	Simulating overdispersed data	Identification in linear models
	Structure in social networks	Alternative models for discrete data	Generalized linear model with offset	Functional forms for non-linear models
18. Design and sample size decisions (Chapter 16 of <i>ROS</i> )	The multiverse and the feedback loop	Design an experiment from scratch	Design analysis by simulation	Designing a survey
	Lucky golf balls and implausible effect sizes	Hypothetical study of left-handedness	Design for estimating interactions	Designing future studies
19. Poststratification and missing-data imputation (Chapter 17 of <i>ROS</i> )	Estimating state-level opinion	Generalizing from class to population	Regression and post-stratification	Network sampling
	Environmental Sustainability Index	Experimental design and effect sizes	Random imputation	Problems with missing data

Week	Stories	Activities	Computer demonstrations	Discussion problems
20. Causal inference and randomized experiments (Chapter 18 of <i>ROS</i> )	Varying treatment effects	Potential outcomes for basketball	Data analysis for basketball activity	Randomization and ethics
	Ballot-order effects	Potential outcomes for ballot order	Sample and population averages	Assumptions in randomized experiments
21. Causal inference using regression on treatment (Chapter 19 of <i>ROS</i> )	Pest control experiment	Adjustments in causal inference	Benefits of pre-treatment data	Causal logistic regression
	Social penumbras	Average treatment effects	Combining pre-treatment predictors	Holding all else equal?
22. Causal inference (Chapters 18–19 of <i>ROS</i> )	No effect of heart stents?	Components of an observational study	Playing with least squares	Individual and average effects
	The freshman fallacy	Study makers vs. study breakers	Don't adjust for intermediate outcomes	Nudge meta-analysis
23. Observational studies with measured confounders (Chapter 20 of <i>ROS</i> )	Retrospective evaluation of a policy	Imbalance and lack of overlap	Poststratification for causal inference	Effects of campaign contributions
	Postal service modeling	Victimization and views on crime policy	Measurement error models	Effects and variation
24. Additional topics in causal inference (Chapter 21 of <i>ROS</i> )	Deterrent effect of death penalty	Two measures of the same quantity	Instrumental variables	Effects of masks
	Regression discontinuity mishaps	"Why" questions and causal inference	Adjustment in regression discontinuity	Admissions test coaching
25. Advanced regression and multilevel models (Chapter 22 of <i>ROS</i> )	Nonlinearity in leafout dates	Nonlinear treatment effect	Modeling golf putting in Stan	Noisy time series
	Governors' elections and lifespans	When do students stop coming to class?	Opinions on same-sex marriage	20 data points and 16 predictors
26. Review of the course (Chapters 1–22 of <i>ROS</i> )	Randomized trials in international development	Designing a paper helicopter	Quadratic regression	Design using simulation
	Is North Carolina less democratic than North Korea?	Review in groups	Bias and unmodeled uncertainty	Electoral integrity index