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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

Chapter 6

Background on regression modeling

Note:
This chapter includes several code examples in R.
Feel free to skip those parts as they're not essential.
If you're feeling motivated, you can skim them and try reproducing them in Python

At a purely mathematical level, the methods described in this book have two purposes: prediction and comparison. We can use regression to predict an outcome variable, or more precisely the distribution of the outcome, given some set of inputs. And we can compare these predictions for different values of the inputs, to make simple comparisons between groups, or to estimate causal effects, a topic to which we shall return in Chapters 18–21. In this chapter we use our favored technique of fake-data simulation to understand a simple regression model, use a real-data example of height and earnings to warn against unwarranted causal interpretations, and discuss the historical origins of regression as it relates to comparisons and statistical adjustment.

6.1 Regression models

The simplest regression model is linear with a single predictor:

$$\text{Basic regression model: } y = a + bx + \text{error}.$$

The quantities a and b are called *coefficients* or, more generally, *parameters* of the model.

The simple linear model can be elaborated in various ways, including:

- Additional predictors: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \text{error}$, written in vector-matrix notation as $y = X\beta + \text{error}$.
- Nonlinear models such as $\log y = a + b \log x + \text{error}$.
- Nonadditive models such as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \text{error}$, which contains an *interaction* between the input variables x_1 and x_2.
- *Generalized linear models*, which extend the linear regression model to work with discrete outcomes and other data that cannot be fit well with normally distributed additive errors, for example predicting support for the Republican or Democratic presidential candidate based on the age, sex, income, etc. of the survey respondent.
- *Nonparametric models*, which include large numbers of parameters to allow essentially arbitrary curves for the predicted value of y given x.
- *Multilevel models*, in which coefficients in a regression can vary by group or by situation. For example, a model predicting college grades given admissions test scores can have coefficients that vary by college.
- *Measurement-error models*, in which predictors x as well as outcomes y are measured with error and there is a goal of estimating the relationship between the underlying quantities. An example is estimating the effect of a drug under partial compliance so that the dose taken by each individual patient is not exactly known.

In this book we shall focus on the first four of the generalizations above.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

6.2 Fitting a simple regression to fake data

Example: Regression fit to simulated data

We demonstrate linear regression with a simple example in R.[1] First we load in the `rstanarm` package, which allows us to fit regression models using the statistical inference engine Stan:

```
library("rstanarm")
```

We then simulate 20 fake data points y_i from the model, $y_i = a + bx_i + \epsilon_i$, where the predictor x_i takes on the values from 1 to 20, the intercept is $a = 0.2$, the slope is $b = 0.3$, and the errors ϵ_i are normally distributed with mean 0 and standard deviation $\sigma = 0.5$, so that we expect roughly two-thirds of the points to fall within ± 1 standard error of the line. Here is the code:[2]

```
x <- 1:20
n <- length(x)
a <- 0.2
b <- 0.3
sigma <- 0.5
y <- a + b*x + sigma*rnorm(n)
```

Fitting a regression and displaying the results

To fit the regression we set up a *data frame* containing predictor and outcome. The data frame can have any name; here we call it `fake` to remind ourselves that this is a fake-data simulation:

```
fake <- data.frame(x, y)
```

And then we can fit the model using the `stan_glm` (using Stan to fit a generalized linear model) function in R;[3] we can save the fit using any name:

```
fit_1 <- stan_glm(y ~ x, data=fake)
```

And then we can display the result:

```
print(fit_1, digits=2)
```

which yields:

```
            Median MAD_SD
(Intercept) 0.40   0.23
x           0.28   0.02

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.49   0.08
```

The first two rows of output tell us that the estimated intercept is 0.40 with uncertainty 0.23, and the estimated slope is 0.28 with uncertainty 0.02. The residual standard deviation σ is estimated at 0.49 with an uncertainty of 0.08.

Under the hood, fitting the model in Stan produced a set of simulations summarizing our inferences about the parameters a, b, and σ, and the output on the screen shows the median and mad sd (see Section 5.3) to produce a point estimate and uncertainty for each parameter.

It can be helpful to plot the data and fitted regression line:

```
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)
```

For convenience we can also put the formula on the graph:

[1] See Appendix A for instructions on how to set up and use R.
[2] Code for this example is in the folder `Simplest`.
[3] See Sections 1.6 and 8.4 for background on the `stan_glm` function.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

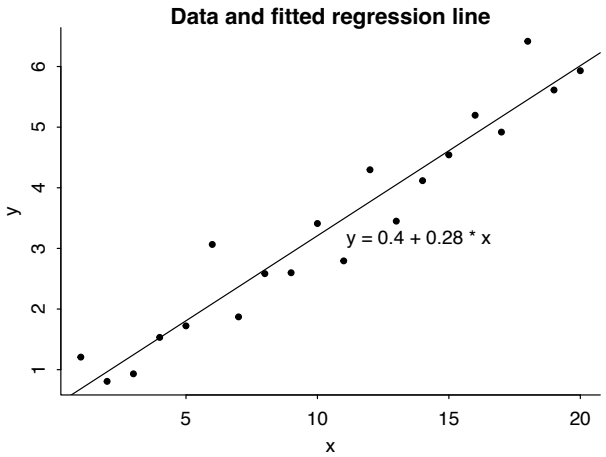

Figure 6.1 *Simple example of a regression line fit to fake data. The 20 data points were simulated from the model, $y = 0.2 + 0.3x + \text{error}$, with errors that were independent and normally distributed with mean 0 and standard deviation 0.5.*

Parameter	Assumed value	Estimate	Uncertainty
a	0.2	0.40	0.23
b	0.3	0.28	0.02
σ	0.5	0.49	0.08

Figure 6.2 *After simulating 20 fake data points from a simple linear regression, $y_i = a + bx_i + \epsilon_i$, with errors ϵ_i drawn from a normal distribution with mean 0 and standard deviation σ, we then fit a linear regression to these data and obtain estimates and uncertainties for the three parameters from the model. We can then see that the estimates are roughly consistent with the specified parameter values.*

```
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)
```

The result is shown in Figure 6.1.

Comparing estimates to assumed parameter values

Having fit the model to fake data, we can now compare the parameter estimates to their assumed values. For simplicity we summarize the results in Figure 6.2, which simply repeats the results from the fitted regression model on page 82.

To read these results, start with the intercept a, which we set to 0.2 in the simulations. After fitting the model to fake data, the estimate is 0.40, which is much different from the assumed 0.2—but the uncertainty, or standard error, in the estimate is 0.23. Roughly speaking, we expect the difference between the estimate and the true value to be within 1 standard error 68% of the time, and within 2 standard errors 95% of the time; see Figure 4.1. So if the true value is 0.2, and the standard error is 0.23, it's no surprise for the estimate to happen to be 0.40. Similarly, the estimates for b and σ are approximately one standard error away from their true values.

As just illustrated, any given fake-data simulation with continuous data would not exactly reproduce the assumed parameter values. Under repeated simulations, though, we should see appropriate coverage, as illustrated in Figure 4.2. We demonstrate fake-data simulation for linear regression more fully in Section 7.2.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

6.3 Interpret coefficients as comparisons, not effects

Example: Height and earnings

Regression coefficients are commonly called "effects," but this terminology can be misleading. We illustrate with an example of a regression model fit to survey data from 1816 respondents, predicting yearly earnings in thousands of dollars, given height in inches and sex, coded as `male` = 1 for men and 0 for women:[4]

```
earnings$earnk <- earnings$earn/1000
fit_2 <- stan_glm(earnk ~ height + male, data=earnings)
print(fit_2)
```

This yields,

```
            Median MAD_SD
(Intercept) -26.0   11.8
height        0.6    0.2
male         10.6    1.5

Auxiliary parameter(s):
      Median MAD_SD
sigma 21.4    0.3
```

The left column above shows the estimated parameters of the model, and the right column gives the uncertainties in these parameters. We focus here on the estimated model, putting off discussion of inferential uncertainty until the next chapter.

The table begins with the regression coefficients, which go into the fitted model:

$$\text{earnings} = -26.0 + 0.6 * \text{height} + 10.6 * \text{male} + \text{error}.$$

Then comes `sigma`, the residual standard deviation, estimated at 21.4, which indicates that earnings will be within $\pm$ 21 400 of the linear predictor for about 68% of the data points and will be within $\pm 2 * \$21\,400 = \$42\,800$ of the linear predictor approximately 95% of the time. The 68% and 95% come from the properties of the normal distribution reviewed in Figure 3.7; even though the errors in this model are not even close to normally distributed, we can use these probabilities as a rough baseline when interpreting the residual standard deviation.

We can get a sense of this residual standard deviation by comparing it to the standard deviation of the data and then estimating the proportion of variance explained, which we compute as 1 minus the proportion of variance unexplained:

```
R2 <- 1 - sigma(fit_2)^2 / sd(earnings$earnk)^2
```

which returns the value $R^2 = 0.10$, meaning that the linear model accounts for only 10% of the variance in earnings in these data. This makes sense, given that people's earnings vary a lot, and most of this variation has nothing to do with height or sex. We discuss R^2 further in Section 11.6; for now, you can just think of it as a way of putting a scale on σ, the residual standard deviation.

We have to be careful not to overinterpret the fitted model. For example, it might seem natural to report that the estimated effect of height is \$600 and the estimated effect of sex is \$10 600.

Strictly speaking, though, it is inappropriate to label these as "effects"—at least, not without a lot of assumptions. We say this because we define an *effect* as the change associated with some *treatment*, or intervention. To say that "the effect of height on earnings" is \$600 is to suggest that, if we were to increase someone's height by one inch, his or her earnings would increase by an expected amount of \$600. But this is not really what's being estimated from the model. Rather, what is observed is an observational pattern, that taller people in the sample have higher earnings on average. These data allow between-person comparisons, but to speak of effect of height is to reference a hypothetical within-person comparison.

[4] The survey was conducted in 1990, and for the analyses in this book we exclude respondents with missing values of height or earnings. Data and code for this example are in the folder `Earnings`.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

Figure 6.3 *(a) Scatterplot adapted from data from Pearson and Lee (1903) of the heights of mothers and their adult daughters, along with the* regression line *predicting daughters' from mothers' heights. (b) The regression line by itself, just to make the pattern easier to see. The line automatically goes through the mean of the data, and it has a slope of 0.54, implying that, on average, the difference of a daughter's height from the average (mean) of women's heights is only about half the difference of her mother's height from the average.*

How, then, can we think of the coefficient for height in the fitted model? We can say that, under the fitted model, the average difference in earnings, comparing two people of the same sex but one inch different in height, is $600. *The safest interpretation of a regression is as a comparison*.

Similarly, it would be inappropriate to say that the estimated "effect of sex" is $10 600. Better to say that, when comparing two people with the same height but different sex, the man's earnings will be, on average, $10 600 more than the woman's in the fitted model.

Under some conditions, the between-person inferences from a regression analysis can be interpreted as causal effects—see Chapters 18–21—but as a starting point we recommend describing regression coefficients in predictive or descriptive, rather than causal, terms.

To summarize: regression is a mathematical tool for making predictions. Regression coefficients can sometimes be interpreted as effects, but they can always be interpreted as average comparisons.

6.4 Historical origins of regression

Example: Mothers' and daughters' heights

"Regression" is defined in the dictionary as "the process or an instance of regressing, as to a less perfect or less developed state." How did this term come to be used for statistical prediction? This connection comes from Francis Galton, one of the original quantitative social scientists, who fit linear models to understand the heredity of human height. Predicting children's heights from parent's heights, he noticed that children of tall parents tended to be taller than average but less tall than their parents. From the other direction, children of shorter parents tended to be shorter than average but less short than their parents. Thus, from one generation to the next, people's heights have "regressed" to the average or *mean*, in statistics jargon.

Daughters' heights "regressing" to the mean

We illustrate with a classic study of the heredity of height, published in 1903 by Karl Pearson and Alice Lee.[5] Figure 6.3a shows the data of mothers' and daughters' heights along with the *regression*

[5]Data and code for this example are in the folder `PearsonLee`.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

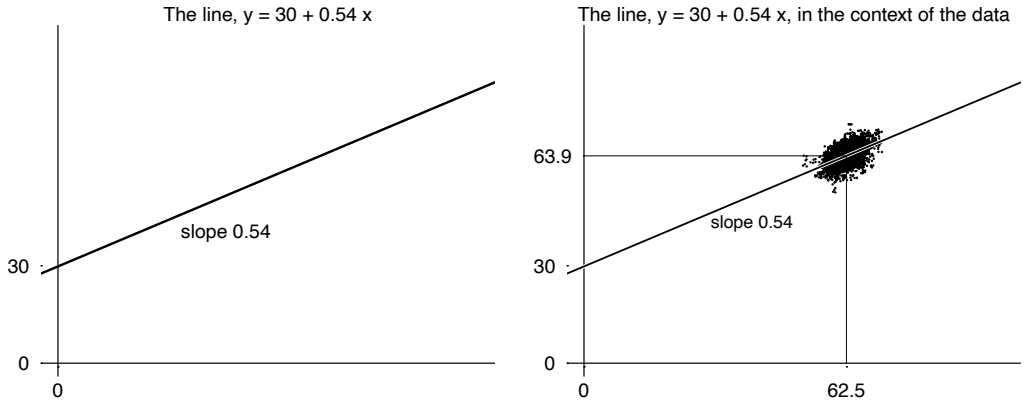

Figure 6.4 *(a) Fitted regression line, $y = 30 + 0.54\,x$, graphed using intercept and slope. (b) Difficulty of the intercept-slope formulation in the context of the data in the height example. The intercept of 30 inches corresponds to the predicted height of a daughter whose mother is a meaningless 0 inches tall.*

line—the best-fit line for predicting daughters' from mothers' heights. The line goes through the mean (average) of x and y, shown with a large dot in the center of the graph.

Figure 6.3b shows the line by itself, the formula, $y = 30 + 0.54x$, which we can also write as,

$$y = 30 + 0.54x + \text{error}, \tag{6.1}$$

to emphasize that the model does not fit individual data points perfectly. We shall give R code for displaying the data and fitting the line, but first we briefly discuss the line itself.

The equation $y = 30 + 0.54x$ describes a line with intercept 30 and slope 0.54, as shown in Figure 6.4a. The intercept-slope formula is an easy way to visualize a line, but it can have problems in various real-world settings, as we demonstrate in Figure 6.4b. The line's slope of 0.54 is clearly interpretable in any case—adding one inch to mother's height corresponds to an increase of 0.54 inches in daughter's predicted height—but the intercept of 30 is hard to understand on its own: it corresponds to the predicted height of a daughter whose mother is a meaningless 0 inches tall.

Instead we can use a different expression of the regression line, centering it not at 0 but at the mean of the data. The equation $y = 30 + 0.54x$ can equivalently be written as,

$$y = 63.9 + 0.54(x - 62.5), \tag{6.2}$$

as shown in Figure 6.3b. This formula shows that when $x = 62.5$, y is predicted to be 63.9.

To put this in the context of the example, if a mother has average height, her adult daughter is predicted to have average height. And then for each inch that a mother is taller (or shorter) than the average height, her daughter is expected to be about half an inch taller (or shorter) than the average for her generation.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

6.5 The paradox of regression to the mean

Now that we have gone through the steps of fitting and graphing the line that predicts daughters' from mothers' heights, we can return to the question of heights "regressing to the mean."

When looked at a certain way, the regression slope of 0.54 in Figure 6.3—indeed, any slope other than 1—seems paradoxical. If tall mothers are likely to have daughters who are only tallish, and short mothers are likely to have shortish daughters, does this not imply that daughters will be more average than their mothers, and that if this continues, each generation will be more average than the last, until, after a few generations, everyone will be just about of average height? For example, a mother who is 8 inches taller than average is predicted to have a daughter 4 inches taller than average, whose

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

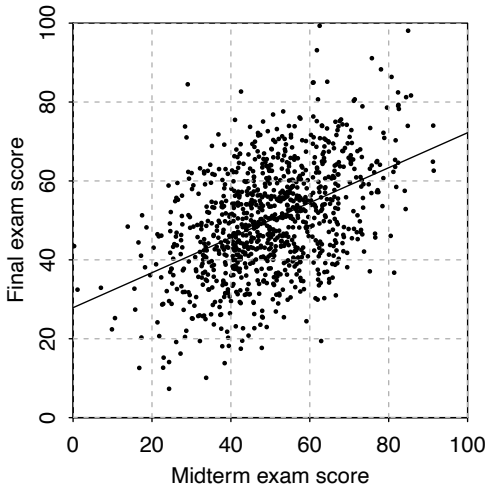

Figure 6.5 *Scatterplot of simulated midterm and final exam scores with fitted regression line, which has a slope of 0.45, implying that if a student performs well on the midterm, he or she is expected to do not so well on the final, and if a student performs poorly on the midterm, he or she is expected to improve on the final; thus, regression to the mean.*

daughter would be predicted to be only 2 inches taller than average, with *her* daughter predicted to be only an inch taller than average, and so forth.

But clearly this is not happening. We are already several generations after Pearson and Lee, and women's heights are as variable as ever.

The resolution of the apparent paradox is that yes, the *predicted* height of a woman is closer to the average, compared to her mother's height, but the actual height is not the same thing as the prediction, which has error; recall equation (6.1). The point predictions regress toward the mean—that's the coefficient less than 1—and this reduces variation. At the same time, though, the error in the model—the imperfection of the prediction—*adds* variation, just enough to keep the total variation in height roughly constant from one generation to the next.

Regression to the mean thus will always arise in some form whenever predictions are imperfect in a stable environment. The imperfection of the prediction induces variation, and regression in the point prediction is required in order to keep the total variation constant.

How regression to the mean can confuse people about causal inference; demonstration using fake data

Example: Simulated midterm and final exams

Regression to the mean can be confusing and it has led people to mistakenly attribute causality. To see how this can happen, we move from heights of parents and children to the mathematically equivalent scenario of students who take two tests.

Figure 6.5 shows a hypothetical dataset of 1000 students' scores on a midterm and final exam. Rather than using real data, we have simulated exam scores using the following simple process representing signal and noise:[6]

1. Each student is assumed to have a true ability drawn from a distribution with mean 50 and standard deviation 10.
2. Each student's score on the midterm exam is the sum of two components: the student's true ability, and a random component with mean 0 and standard deviation 10, reflecting that performance on any given test will be unpredictable: a midterm exam is far from a perfect measuring instrument.

[6]Code for this example is in the folder `FakeMidtermFinal`.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

3. Likewise, each student's score on the final exam is his or her true ability, plus another, independent, random component.

the fitted regression line:

```
fit_1 <- stan_glm(final ~ midterm, data=exams)
```

And here's the regression output:

```
(Intercept) 24.8
midterm      0.5
```

The estimated slope is 0.5 (see also Figure 6.5), which by being less than 1 is an example of regression to the mean: students who score high on the midterm tend to score only about half as high, compared to the average, on the final; students who score low on the midterm score low, but typically not as low, compared to the average, on the final. For example, on the far left of Figure 6.5 are two students who scored zero on the midterm and 33 and 42 on the final; on the far right of the graph are three students who scored 91 on the midterm and between 61 and 75 on the final.

It might seem natural to interpret this causally, to say that students who score well on the midterm have high ability but then they tend to get overconfident and goof off; hence, they typically don't do so well on the final. From the other direction, the appealing causal story is that poor-scoring students on the midterm are motivated to try extra hard, so they improve when the final exam comes along.

Actually, though, the data were simulated from a theoretical model that contained *no* motivational effects at all; both the midterm and the final were a function of true ability plus random noise. We know this because we created the simulation!

The pattern of regression to the mean—that is, the slope of the line in Figure 6.5 being less than 1—is a consequence of variation between the first and second observations: a student who scores very well on the midterm is likely to have been somewhat lucky and also to have a high level of skill, and so in the final exam it makes sense for the student to do better than the average but worse than on the midterm.

The point is that a naive interpretation of the data in Figure 6.5 could lead you to infer an effect (better-scoring students being lazy on the final; worse-scoring students studying harder) that is entirely spurious. This error is called the "regression fallacy."

Example: Flight school

A famous real-world example was reported by the psychologists Amos Tversky and Daniel Kahneman in 1973:

> The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with this training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try.

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

Actually, though, they explain:

> Regression is inevitable in flight maneuvers because performance is not perfectly reliable and progress between successive maneuvers is slow. Hence, pilots who did exceptionally well on one trial are likely to deteriorate on the next, regardless of the instructors' reaction to the initial success. The experienced flight instructors actually discovered the regression but attributed it to the detrimental effect of positive reinforcement. This true story illustrates a saddening aspect of the human condition. We normally reinforce others when their behavior is good and punish them when their behavior is bad. By regression alone, therefore, they are most likely to improve after being punished and most likely to deteriorate after being rewarded. Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding.

The point of this story is that a *quantitative* understanding of prediction clarifies a fundamental *qualitative* confusion about variation and causality. From purely mathematical considerations, it is expected that the best pilots will decline, relative to the others, while the worst will improve in their rankings, in the same way that we expect daughters of tall mothers to be, on average, tall but not quite as tall as their mothers, and so on.

Relation of "regression to the mean" to the larger themes of the book

The regression fallacy described above is a particular example of a misinterpretation of a comparison. The key idea is that, for causal inference, you should compare like with like.

We can apply this idea to the examples of regression to the mean. In the test scores problem, the causal claim is that doing poorly on the midterm exam is a motivation for students to study hard for the final, while students who do well on the midterm are more likely to relax. In this comparison, the outcome y is the final exam score, and the predictor x is the midterm score. The striking result is that, comparing students who differ by 1 unit on x, their expected difference is only $\frac{1}{2}$ unit on y.

And why is this striking? Because it is being compared to the slope of 1. The observed pattern as shown in the regression table and in Figure 6.5 is being compared to an implicit default model in which midterm and final exam scores are the same. But the comparison between these two models is inappropriate because the default model is not correct—there is not, in fact, any reason to suspect that midterm and final exam scores would be identical in the absence of any motivational intervention.

Our point here is not that there is a simple analysis which would allow us to perform causal inference in this setting. Rather, we are demonstrating regression to the mean, along with a comparison to an implicit (but, upon reflection, inappropriate) model can lead to incorrect causal inferences.

Again, in the flight school example, a comparison is being made to an implicit model in which, absent any positive or negative reinforcement, individual performance would stay still. But such a model is inappropriate in the context of real variation from trial to trial.

6.6 Bibliographic note

For background on the height and earnings example, see Ross (1990) and the bibliographic note at the end of Chapter 12.

The data on mothers' and daughters' heights in Figure 6.3 come from Pearson and Lee (1903); see also Wachsmuth, Wilkinson, and Dallal (2003), and Pagano and Anoke (2013) for more on this example. The idea of regression coefficients as comparisons relates to the four basic statistical operations of Efron (1982).

The historical background of regression to the mean is covered by Stigler (1986), and some of its connections to other statistical ideas are discussed by Stigler (1983). Lord (1967, 1969) considers how regression to the mean can lead to confusion about causal inference. The story of the pilots' training comes from Kahneman and Tversky (1973).

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

6.7 Exercises

6.1 *Data and fitted regression line*: 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.

6.2 *Programming fake-data simulation*: Write an R function to: (i) simulate n data points from the model, $y = a + bx +$ 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 σ; (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, σ, 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, σ.

6.3 *Variation, uncertainty, and sample size*: Repeat the example in Section 6.2, varying the number of data points, n. What happens to the parameter estimates and uncertainties when you increase the number of observations?

6.4 *Simulation study*: Perform the previous exercise more systematically, trying out a sequence of values of n, for each simulating fake data and fitting the regression to obtain estimate and uncertainty (median and mad sd) for each parameter. Then plot each of these as a function of n and report on what you find.

6.5 *Regression prediction and averages*: The heights and earnings data in Section 6.3 are in the folder `Earnings`. Download the data and compute the average height for men and women in the sample.

(a) Use these averages and fitted regression model displayed on page 84 to get a model-based estimate of the average earnings of men and of women in the population.

(b) Assuming 52% of adults are women, estimate the average earnings of adults in the population.

(c) Directly from the sample data compute the average earnings of men, women, and everyone. Compare these to the values calculated in parts (a) and (b).

6.6 *Selection on x or y*:

(a) Repeat the analysis in Section 6.4 using the same data, but just analyzing the observations for *mothers'* heights less than the mean. Confirm that the estimated regression parameters are roughly the same as were obtained by fitting the model to all the data.

(b) Repeat the analysis in Section 6.4 using the same data, but just analyzing the observations for *daughters'* heights less than the mean. Compare the estimated regression parameters and discuss how they differ from what was obtained by fitting the model to all the data.

(c) Explain why selecting on daughters' heights had so much more of an effect on the fit than selecting on mothers' heights.

6.7 *Regression to the mean*: Gather before-after data with a structure similar to the mothers' and daughters' heights in Sections 6.4 and 6.5. These data could be performance of athletes or sports teams from one year to the next, or economic outcomes in states or countries in two successive years, or any other pair of measurements taken on a set of items. Standardize each of the two variables so it has a mean of 0 and standard deviation of 1.

(a) Following the steps of Section 6.4, read in the data, fit a linear regression, and plot the data and fitted regression line.

(b) Repeat the above steps with fake data that look similar to the data you have gathered.

6.8 *Regression to the mean with fake data*: Perform a fake-data simulation as in Section 6.5, but using the flight school example on page 89. Simulate data from 500 pilots, each of whom performs two maneuvers, with each maneuver scored continuously on a 0–10 scale, that each pilot has a

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, Jennifer Hill, and Aki Vehtari 2020–2023. The book web page https://avehtari.github.io/ROS-Examples/

true ability that is unchanged during the two tasks, and that the score for each test is equal to this true ability plus independent errors. Further suppose that when pilots score higher than 7 on the scale during the first maneuver, that they get praised, and that scores lower than 3 on the first maneuver result in negative reinforcement. Also suppose, though, that this feedback has no effect on performance on the second task.

(a) Make a scatterplot with one dot for each pilot, showing score on the second maneuver vs. score on the first maneuver. Color the dots blue for the pilots who got praised, red for those who got negative reinforcement, and black for the other cases.

(b) Compute the average change in scores for each group of pilots. If you did your simulation correctly, the pilots who were praised did worse, on average, and the pilots who got negative reinforcement improved, on average, for the second maneuver. Explain how this happened, given that your data were simulated under a model in which the positive and negative messages had no effects.

6.9 *Working through your own example*: Continuing the example from the final exercises of the earlier chapters, find two variables that represent before-and-after measurements of some sort. Make a scatterplot and discuss challenges of "regression to the mean" when interpeting before-after changes here.