

JAMES G. SCOTT

# DATA SCIENCE: A GENTLE INTRODUCTION

Copyright ©2017 James G. Scott

[HTTP://JGSCOTT.GITHUB.COM/](http://jgscott.github.com/)

These lecture notes are copyrighted materials, and are made available solely for educational and not-for-profit use. Any unauthorized use or distribution without written consent is prohibited.

*First published January 2010*

# *Contents*

1	<i>Data exploration</i>	9
2	<i>Fitting equations to data</i>	31
3	<i>Predictable and unpredictable variation</i>	57
4	<i>Grouping variables in regression</i>	77
5	<i>Quantifying uncertainty using the bootstrap</i>	107
6	<i>Multiple regression: the basics</i>	127
7	<i>Testing hypotheses</i>	147
8	<i>Building predictive models</i>	163
9	<i>Understanding cause and effect</i>	175



# *Introduction*

THIS BOOK is about data science. This term has no precise definition. Data science involves some statistics, some probability, some computing—and above all, some knowledge of your data set (the “science” part).

The goal of data science is to help us understand patterns of variation in data: economic growth rates, dinosaur skull volumes, student SAT scores, genes in a population, Congressional party affiliations, drug dosage levels, your choice of toothpaste versus mine . . . really any variable that can be measured.

To do that, we often use *models*. A model is a metaphor, a description of a system that helps us to reason more clearly. Like all metaphors, models are approximations, and will never account for every last detail. A useful mantra here is: all models are wrong, but some models are useful.<sup>1</sup> Aerospace engineers work with physical models—blueprints, simulations, mock-ups, wind-tunnel prototypes—to help them understand a proposed airplane design. Geneticists work with animal models—fruit flies, mice, zebrafish—to help them understand heredity. In data science, we work with statistical models to help us understand *variation*.

Like the weather, most variation in the world exhibits some features that are predictable, and some that are unpredictable. Will it snow on Christmas day? It’s more likely in Boston than Austin, and more likely still at the North Pole; that’s predictable variation. But even as late as Christmas eve, and even at the North Pole, nobody knows for sure; that’s unpredictable variation.

Statistical models describe both the predictable and the unpredictable variation in some system. More than that, they allow us to partition observed variation into its predictable and unpredictable components—and not just in some loose allegorical way, but in a precise mathematical way that can, with perfect accuracy, be described as Pythagorean. (More on that later.)

This focus on the structured quantification of uncertainty is what distinguishes data science from ordinary evidence-based reasoning. It’s important to know what the evidence says, goes this

<sup>1</sup> Attributed to George Box.

line of thinking. But it's also important to know what it doesn't say. Sometimes that's the tricky part.

We will learn data science for three purposes:

- (1) *to help us explore* a large body of data, so that we might identify predictable features or trends amid random variation.
- (2) *to test* our beliefs about relationships among things we can measure.
- (3) *to predict* the future behavior of some system, and to say something useful about what remains unpredictable.

These are the goals not merely of data science, but of the scientific method more generally.

*What data science isn't.* Many people assume that the job of a data scientist is to objectively summarize the facts, slap down a few error bars, and get out of the way.

This view is mistaken. To be sure, data science demands a deep respect for facts, and for not allowing one's wishes or biases to change the story one tells *with* the facts. But the process of analyzing data is inescapably subjective, in a way that should be embraced rather than ignored. Data science requires much more than just technical knowledge of ideas from statistics and computing. It also requires care and judgment, and cannot be reduced to a flowchart, a table of formulas, or a tidy set of numerical summaries that wring every last drop of truth from a data set. There is almost never a single "right" data-science approach for some problem. But there are definitely such things as good models and bad approaches, and learning to tell the difference is important. Just remember: calling a model good or bad requires knowing both the tool and the task. A shop-window mannequin is good for displaying clothes, but bad for training medical students about vascular anatomy. A big part of your statistical education is to hone this capacity for deciding when a statistical model is fit for its intended purpose.

Second, many people assume that data science must involve complicated models and calculations in order to do justice to the real world. Not always: complexity sometimes comes at the expense of explanatory power. We must avoid building models calibrated so perfectly to past experience that they do not generalize to future cases. This idea—that theories should be made as

complicated as they need to be, and no more so—is often called “Occam’s Razor.” A good model will be simple enough to understand and interpret, but not so simple that it does any major intellectual violence to the system being modeled. All models of the world must balance these goals, and statistical models are no exception.

Finally, many people also assume that data science involves difficult, tedious mathematics. Happily, this isn’t true at all. In fact, virtually all common techniques in data science are accessible to anyone with a high-school math education, and these days all the tedious calculations are taken care of by computers.

### *Data science then and now*

On the time scale of important post-Enlightenment ideas, the key tools of data science are middle-aged. A German astronomer named Tobias Mayer was using something vaguely like linear regression modeling (a data-science workhorse) as early as 1750.<sup>2</sup> But most scholars credit two later mathematicians—Legendre, a Frenchmen; and Gauss, a German—with independently inventing the *method of least squares* some time between 1794 and 1805. As you will soon discover (or may already know), the method of least squares is our primary mathematical workhorse for fitting models to data. That makes regression modeling newer than the invention of calculus (credited jointly to Leibniz and Newton in the late 1600’s), but older than the idea of evolution by natural selection (credited jointly to Darwin and Wallace over a period spanning the 1830’s to the 1850’s).

For most of the nineteenth century, data science largely remained the concern of a highly specialized group of astronomers and geophysicists. But in our own age—one of fast, cheap computers and abundant data—it has become ubiquitous. The very same principle of least squares proposed by Legendre and Gauss remains, over two hundred years later, an important part of the day-to-day toolkit for solving problems in fields from aeronautics to zoology and everywhere in between. If you’ve ever wondered why your social media accounts are eerily prescient—about your friends, about headlines that might appeal to you, about products you might want to buy—you can thank a data scientist.

Of course, our political and cultural climate still exhibits a streak of distrust toward data. Why else would Winston Churchill’s brazen instructions to a young protégé sound so depressingly fa-

<sup>2</sup> Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900*, pp. 16–25. Harvard University Press, 1986

miliar?

I gather, young man, that you wish to be a Member of Parliament. The first lesson that you must learn is that, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister.<sup>3</sup>

And why else would the famous remark, popularized by Twain and attributed to Disraeli, remain so apt, even a century later?

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: ‘There are three kinds of lies: lies, damned lies, and statistics.’<sup>4</sup>

How do you tell the difference between “robust, unbiased evidence,” misleading irrelevance, and cynical fraud? In considering this question, you will already have appreciated at least two good reasons to learn data science:

- (1) To use data honestly and credibly in the service of an argument you believe in.
- (2) To know how and when to be skeptical of someone else’s damned lies.

For as John Adams put it,

Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passion, they cannot alter the state of facts and evidence.<sup>5</sup>

<sup>3</sup> Quoted in *The Life of Politics* (1968), Henry Fairlie, Methuen, pp. 203–204

<sup>4</sup> *Chapters from My Autobiography*, North American Review (1907)

<sup>5</sup> ‘Argument in Defense of the Soldiers in the Boston Massacre Trials’ (1770)

# 1

## *Data exploration*

SUPPLY AND DEMAND, chocolate and peanut butter, education and income . . . some things just go hand in hand. In each case, a particular idea about how things work turns upon the interpretation of an observed relationship between things we can measure. To do this correctly requires care, judgment—and the right toolkit. The goal of this chapter is to equip you with some basic visual and numerical tools for exploring multivariate data sets, with an eye towards finding interesting relationships among variables.

*Cases and variables.* In statistics, we typically refer to the *cases* and *variables* of a data set. The cases are the basic observational units that we’re interested in: people, houses, cars, guinea pigs, etc. The variables are the different kinds of information we have about each case—for example, the horsepower, fuel economy, and vehicle class for a car. We typically organize a data set into a *data frame*. A data frame is like a simple spreadsheet where each case is a row and each variable is a column, like in Table 1.1.

Variables come in two basic kinds. Numerical variables are represented by a number, like horsepower. Categorical variables are described by the answer to a multiple-choice question, like vehicle class. This chapter will describe some strategies for summarizing relationships among both kinds of variables, as well as some further refinements to this basic “numerical versus categorical” distinction.

Table 1.1: A simple example of a data frame. Each case is a car, and there are five variables: horsepower, city gas mileage, highway gas mileage, weight (in pounds), and vehicle class.

	Horsepwer	CityMPG	HighwayMPG	Weight	Class
BMW 325xi	184	19	27	3461	Sedan
Chevrolet Corvette	350	18	25	3248	Sports
Mercedes-Benz CL500	302	16	24	4085	Sedan
Dodge Neon	132	29	36	2626	Sedan
Acura MDX	265	17	23	4451	SUV

## Variation across categories

MANY OF the data sets you'll meet will involve categories: chocolate or vanilla; rap or country; Toyota, Honda, or Hyundai; butcher or baker or candlestick maker. A simple, effective way to summarize these categorical variables<sup>1</sup> is to use a *contingency table*. On the Titanic, for example, a simple two-way table reveals that women and children survived in far greater numbers than adult men:

	Girl	Woman	Boy	Man
Survived	50	242	31	104
Died	22	74	51	472

We call this a two-way or bivariate table because there are two variables being compared: survival status versus type of person. The categories go along the rows and columns of the table; the cell counts show how many cases fall into each class. The process of sorting cases into the cells of such a table is often called *cross-tabulation*.

We can also make multi-way tables that show more than two variables at once. Given the constraints of a two-dimensional page, multiway tables are usually displayed as a series of two-way tables. As the following three-way table reveals, richer passengers, of either sex, fared better than others.

		Cabin Class	1st	2nd	3rd
		Survived	139	94	106
		Died	5	12	110
Male	Survived	61	25	75	
	Died	118	146	418	

Tables are almost always the best way to display categorical data sets with few classifying variables, for the simple reason that they convey a lot of information in a small space.<sup>2</sup>

*Ordinal and binary variables.* If a categorical variable has only two options (heads or tails, survived or died), we often call it an indicator, binary, or dummy variable. (These names can be used interchangeably.)

<sup>1</sup> Categorical variables are sometimes referred to as *factors*, and the categories themselves as the *levels* of the factor. The R statistical software package uses this terminology.

Table 1.2: A two-way table, because there are two categorical variables by which cases are classified. The data are available in the R package *effects*. Originally compiled by Thomas Cason from the *Encyclopedia Titanica*.

Table 1.3: An example of a *multi-way table*, where counts are classified by cabin class, sex, and survival. NB: passengers of unknown age are included in this table, but not the previous one.

<sup>2</sup> This animation provides some good guidelines for formatting tables.

Some categories have a natural ordering, like measures of severity for a hurricane, or responses to a survey about consumer satisfaction. (Has your experience with our call center been Atrocious, Merely Bad, Acceptable, Good, or Excellent?) These are called *ordinal variables*. Ordinal variables differ from numerical variables in that, although they can be placed in a definite order, they cannot be compared using the laws of arithmetic. For example, we can't subtract "Good" from "Excellent" and get a meaningful answer, in the way we can subtract \$1000 from \$5000 and get a number.

### *Relative risk*

The **relative risk**, sometimes also called the risk ratio, is a widely used measure of association between two categorical variables. To introduce this concept, let's examine a tidbit of data from the PREDIMED trial, a famous study on heart health conducted by Spanish researchers that followed the lifestyle and diet habits of thousands of people over many years, beginning in 2003.<sup>3</sup>

The main purpose of the PREDIMED trial was to assess the effect of a Mediterranean-style diet on the likelihood of someone experiencing a major cardiovascular event (defined by the researchers as a heart attack, stroke, or death from cardiovascular causes). But as part of the study, the researchers also collected data on whether the trial participants were, or had ever been, regular smokers. The table below shows the relationship between smoking and whether someone experienced a cardiovascular event during the study period.

Current or former smoker?		
	No ( <i>n</i> = 3892)	Yes ( <i>n</i> = 2432)
No event	3778	2294
Event	114	138

Let's compare the absolute risk of cardiovascular events for smokers, versus that of non-smokers.<sup>4</sup> Among the smokers, 138 of 2432 people (5.7%) experienced an event; while among the non-smokers, 114 of 3892 people (2.9%) experienced an event. To compute the relative risk of cardiovascular events among smokers, we take the ratio of these two absolute risks:

$$\text{Relative risk} = \frac{138/2432}{114/3892} = 1.94.$$

<sup>3</sup> Estruch R, Ros E, Salas-Salvado J, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. *N Engl J Med* 2013;368:1279-1290. The full text of the article is available at <http://www.nejm.org/doi/full/10.1056/NEJMoa1200303>

<sup>4</sup> By "absolute risk," we simply mean the chance of an event happening.

This ratio says that smokers were 1.94 times more likely than non-smokers to experience a cardiovascular event during the study.<sup>5</sup>

More generally, for any event (a disease, a car accident, a mortgage default) and any notion of “exposure” to some factor (smoking, driving while texting, poor credit rating), the relative risk is

$$\text{Relative risk} = \frac{\text{Risk of event in exposed group}}{\text{Risk of event in non-exposed group}}.$$

The relative risk tells us how much more (or less) likely the event is in one group versus another. It’s important to remember that the relative risk (in our example, 1.94 for smokers) is quite different from the *absolute risk* (in our example, 0.057 for smokers). This distinction is often missed or elided in media coverage of health issues. See, for example, [this blog post](#) from the UK’s cancer-research funding body about news reports of cancer studies.

<sup>5</sup> Of course, this doesn’t prove that the smoking caused the cardiovascular events. One could argue that the smokers may have had other systematically unhealthier habits that did them in instead, and the smoking was merely a marker of these other habits. We’ll soon talk about this issue of confounding much more.

## Variation of numerical variables

FIGURE 1.1 depicts a histogram of daily average temperatures in two American cities—San Diego, CA, and Rapid City, SD—for every day from January 1995 to November 2011. Temperature is an example of a *numerical variable*, or something for which numerical comparisons are meaningful (twice as far, six times as fast, \$17 cheaper, and so forth). Numerical variables can be *discrete* or *continuous*. Temperature is continuous; we measure it in arbitrarily small increments. Marbles, on the other hand, are discrete; we count them on our fingers and toes.

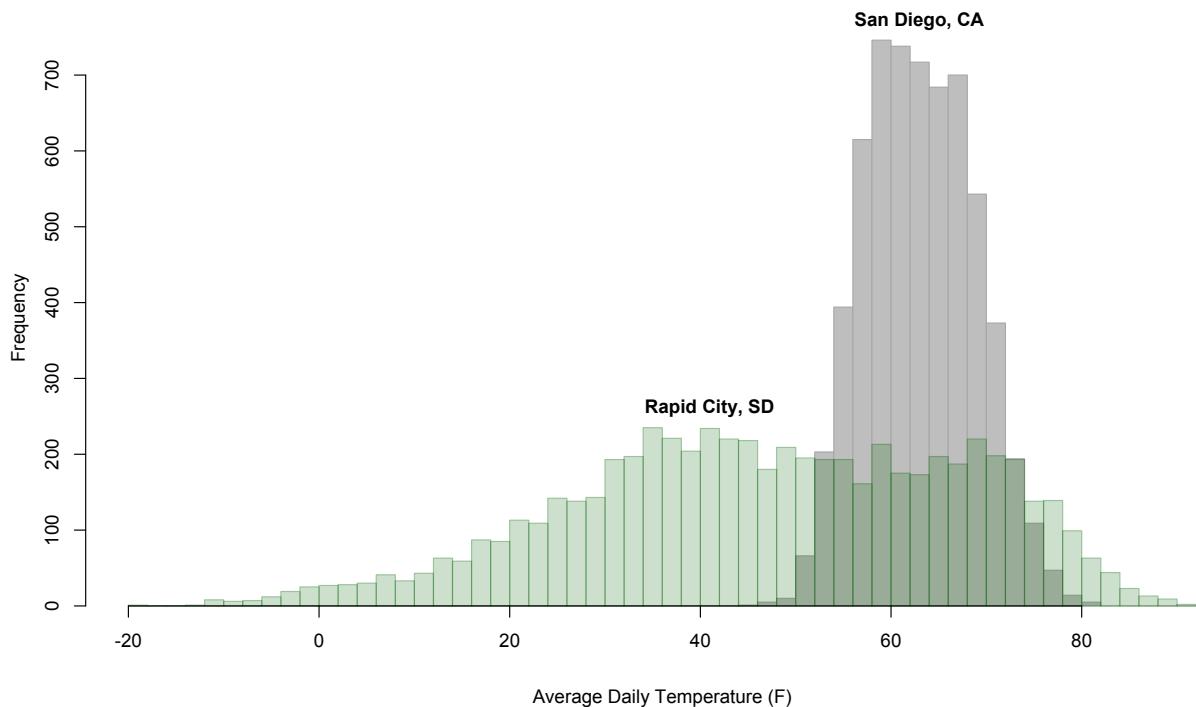
A histogram is a great way to depict the distribution of a numerical variable. To construct one, we first partition the range of possible outcomes (here, temperatures) into a set of disjoint intervals (“bins”). Next, we count the number of cases that fall into each bin. Finally, we draw a rectangle over each bin whose height is equal to the count within each bin.<sup>6</sup>

The histogram in Figure 1.1 suggest two obvious, meaningful questions we can ask about a numerical variable like temperature: where is the middle of the sample, and how much does a typical case vary from the middle?

You’re probably already aware of more than one way to answer the question, “Where is the middle?”

- There’s the sample mean, written as  $\bar{y}$ . If we have  $n$  data

<sup>6</sup> Technically this is called a frequency histogram; one could also make a *density histogram* in which the heights of the bars are scaled appropriately so that the total area of all the bars sums to 1.



points  $\{y_1, \dots, y_n\}$ , then

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

The subscript  $i$ 's run from case 1 to case  $n$ , where  $n$  is the number of data points in the sample. In many data sets the actual ordering of cases won't matter, and will just reflect the arbitrary ordering of the rows in your data frame.<sup>7</sup>

- There's the median, or the halfway point in a sample.
- There's also the mode, or the most common value.

These different ways of quantifying the middle value all have different properties. For example, the median is less sensitive than the mean to extreme values in your sample; there can be more than one mode in a sample, but only one mean or median.<sup>8</sup>

### *Sample standard deviation and sample variance*

Another important question is, "How spread out are the data points from the middle?" Figure 1.1 drives home the importance

Figure 1.1: Daily average temperatures for San Diego and Rapid City, 1995–2011. These data are visualized in a histogram, which is a simple and effective way to depict the variation of a single numerical variable across many cases.

<sup>7</sup> An obvious exception is in the analysis of time-series data, where the ordering of observations in time may be highly meaningful.

<sup>8</sup> For example, consider the data set  $\{1, 2, 3, 3, 4, 4, 5\}$ .

of dispersion in making useful comparisons. Not only are average temperatures lower overall in Rapid City than in San Diego, but they are also a lot more variable: the coldest days are much colder in Rapid City, but the hottest days are hotter, too.

As with the notion of “middle” itself, there is more than one way of quantifying variability, and each way is appropriate for different purposes. Let’s follow the line of thinking that leads us to the *standard deviation*, which is probably the most common way of measuring dispersion. Suppose we choose to measure the middle of a sample  $y_1, \dots, y_n$  using the mean,  $\bar{y}$ . Each case varies from this middle value by its *deviation*,  $y_i - \bar{y}$ . Why not, therefore, just compute the average deviation from the mean? Well, because

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}) &= \frac{1}{n} \sum_{i=1}^n y_i - \frac{n}{n} \bar{y} \\ &= \bar{y} - \bar{y} \\ &= 0.\end{aligned}$$

The positives and negatives cancel each other out. We could certainly fix this by taking the absolute value of each deviation, and then averaging those:

$$M = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}|.$$

This quantity is a perfectly sensible measure of the “typical deviation” from the middle. Fittingly enough, it is called the *mean absolute deviation* of the sample.

But it turns out that, for the purposes of statistical modeling, a quantity called the *sample variance* makes more sense:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

That is, we *square* each deviation from  $\bar{y}$ , rather than take the absolute value. Remember that when we square a negative number, it becomes positive, so that we don’t have the problem of the positives and negatives cancelling each other out.

The definition of sample variance raises two questions:

- (1) Why do we divide by  $n - 1$ , when dividing by  $n$  would seem to make more sense for computing an average?
- (2) Why do we square the deviations, instead of taking absolute values as above?

To answer the first question: we divide by  $n - 1$  rather than  $n$  for obscure technical reasons that, despite what you may read in other statistics textbooks, just aren't that important. (It has to do with "unbiased estimators," which, despite the appealing name, are overrated.) Mainly we use  $n - 1$  to follow convention.

As for the second question: because sums of squares are special! In all seriousness, there are deep mathematical reasons why we choose to measure dispersion using sums of squared deviations, rather than the seemingly more natural sums of absolute deviations. You'll learn why in a future chapter, but if you want a preview, think about Pythagoras and right triangles. . . .

Of course, computing the sample variance leaves us in the awkward position of measuring variation in the *squared* units of whatever our variable is measured in. This is not intuitive; imagining telling someone that the mean temperature in Rapid City over the last 17 years was 47.3 degrees Fahrenheit, with a sample variance of 402 degrees squared. This is a true statement, but nearly uninterpretable.

Luckily, this is easily fixed by taking the square root of the sample variance, giving us the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}. \quad (1.1)$$

Now we're back to the original units, and an interpretable measure of "typical deviation from the middle"—for Rapid City, 20.1 degrees. This looks about right from the histogram below; the blue dot is the sample mean, and the blue line stretches 1 sample standard deviation to either side of the mean.

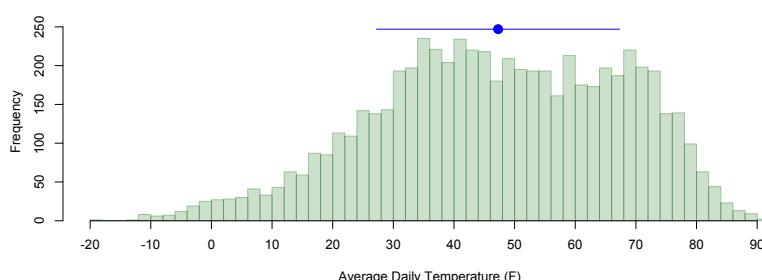


Figure 1.2: The histogram shows average daily temperatures in Rapid City. The blue dot is the sample mean, and the blue line shows an interval encompassing one sample standard deviation to either side of the sample mean.

Two other simple measures of spread are worth mentioning briefly. First, there's the *range*, or the difference between the largest

and smallest values in the sample. There's also the *interquartile range*, or the difference between the 75<sup>th</sup> and 25<sup>th</sup> percentiles. This is robust to extreme values, since it involves only the middle 50% of the sample.

### *Percentiles, quantiles, and coverage intervals*

Another useful way to summarize the variation of a numerical variable across cases is to compute a set of percentiles, also called quantiles. A familiar example is the median: it happens that exactly 50% of the daily average temperatures in Rapid City fall below fall below 47.6 degrees, and we call this point the median (or the 50th percentile). Similarly, 10% of days in Rapid City are colder than 20.7 degrees, and 90% of days are colder than 73.2 degrees; these are the 10th and 90th percentiles, respectively. A quantile is just a percentile expressed in terms of a decimal fraction; the 80th percentile and 0.8 quantile are the same number.

A common way to summarize a distribution of a numerical variable is to quote a *coverage interval* defined by two percentiles, like the 10th and 90th percentiles (which covers 80% of the cases) or the 2.5th and 97.5th percentiles (which covers 95% of the cases). So, for example, we might quote an 80% coverage interval for daily average temperatures in Rapid City as (20.7, 73.2), whose endpoints are formed from the 10th and 90th percentiles.

### *Standardization by z-scoring*

Which temperature is more extreme: 50 degrees in San Diego, or 10 degrees in Rapid City? In an absolute sense, of course 10 degrees is a more extreme temperature. But what about in a relative sense? In other words, is a 10-degree day more extreme *for Rapid City* than a 50-degree day is *for San Diego*? This question could certainly be answered using quantiles, which you've already learned how to handle. But let's discuss a second way: by calculating a z-score for each temperature.

The z-score of some quantity  $x$  is the number of standard deviations by which  $x$  is above its mean. If a z-score is negative, then the corresponding observation is below the mean.

To calculate a z-score for a number  $x$ , we subtract the corresponding mean  $\mu$  and divide by the standard deviation  $\sigma$ :

$$z = \frac{x - \mu}{\sigma}.$$

For a 50-degree day in San Diego, this is:

$$z = \frac{50 - 63.1}{5.7} \approx -2.3.$$

Or about 2.3 standard deviations below the mean. On the other hand, for a 10-degree day in Rapid City, the z-score is

$$z = \frac{10 - 47.3}{20.1} \approx -1.9.$$

Or about 1.9 standard deviations below the mean. Thus a 50-degree day in San Diego is actually more extreme than a 10-degree day in Rapid City! The reason is that temperatures in Rapid City are both colder on average (lower mean) and more variable (higher standard deviation) than temperatures in San Diego.

As this example suggests, z-scores are useful for comparing numbers that come from different distributions, with different statistical properties. It tells you how extreme a number is, relative to other numbers from that same distribution. We often think of the normal distribution as a useful reference here for interpreting z-scores. The normal distribution has the property that about 68% of observations fall within  $z = 1$  standard deviation of the mean, and about 95% fall within  $z = 2$  standard deviations.

## Variation between, and within, groups

A COMMON situation is that we have both categorical and numerical data about each case in a data set. For example, Table 1.4 below shows the average SAT math and verbal scores, stratified by college, for undergraduates in the incoming fall of 2000 freshmen class at the University of Texas at Austin. All 5,191 students who went on to receive a bachelor's degree within 6 years are included; those who dropped out, for whatever reason, are not.

The table tells you something about how the numerical variables (test scores) change depending upon the categorical variable (college), and they are superficially similar to the contingency tables we just encountered. They highlight interesting and useful facts about variation between the groups. Math skills, for example, are probably more important for engineering majors than English majors, and this is reflected in the differences between the group-level means.

College	Average SAT	
	Math	Verbal
Architecture	685	662
Business	633	597
Communications	592	609
Education	555	546
Engineering	675	606
Fine Arts	597	594
Liberal Arts	598	590
Natural Sciences	633	597
Nursing	561	555
Social Work	602	589

Table 1.4: Average SAT math and verbal scores, stratified by college, for entering freshmen at UT–Austin in the fall of 2000. Collected under the Freedom of Information Act from the state of Texas.

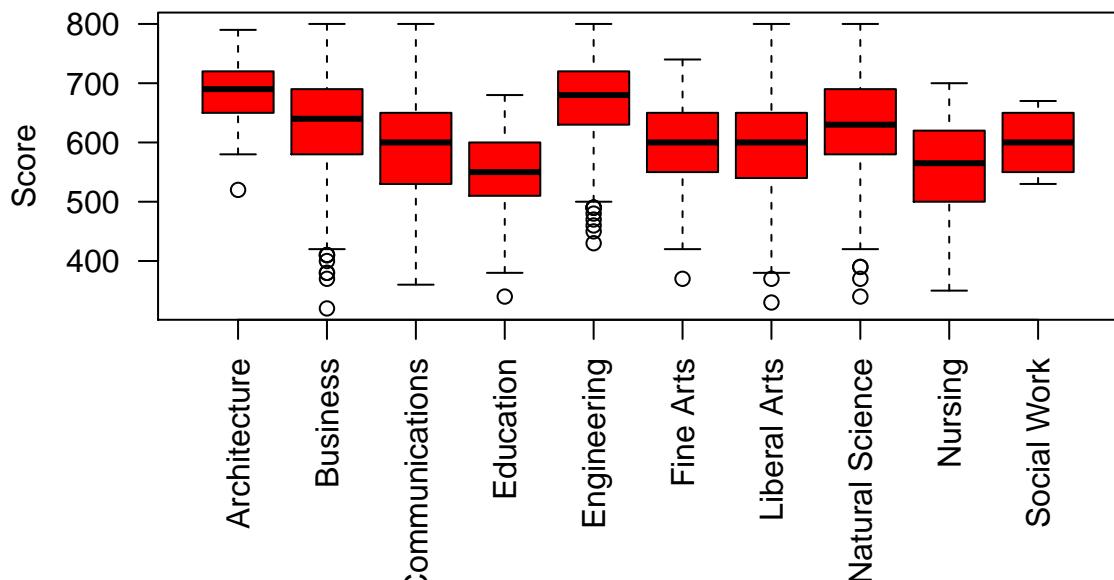
Table 1.4 does differ from a contingency table, however, in one crucial respect: the entries in the table are not counts, but group-level averages. Notice that, to depict between-group variation, the table has reduced each college to a typical case, represented by some hypothetical student who earned the college-wide average SAT scores on both the math and verbal sections. In doing so, it has obscured the underlying variability of students *within* the colleges. But as our example of city temperatures demonstrated, sometimes this variability is an important part of the story as well.

### Boxplots

This is where boxplots are useful: they allow you to assess variability both between and within the groups. In a boxplot, like the ones shown in Figure 1.3, there is one box per category. (The top panel shows a boxplot for SAT Math scores; the bottom, for SAT Verbal scoers.) Each box shows the *within-group variability*, as measured by the interquartile range of the numerical variable (SAT score) for all cases in that category. The middle line within each box is the median of that category, and the differences between these medians give you a sense of the *between-group variability*. In this boxplot, the whiskers extend outside the box no further than 1.5 times the interquartile range. Points outside this interval are shown as individual dots.

A table like 1.4 focuses exclusively on the between-group variability; it reduces each category to a single number, and shows how those numbers vary from one category to the next. But in

### SAT Math Scores by UT College



### SAT Verbal Scores by UT College

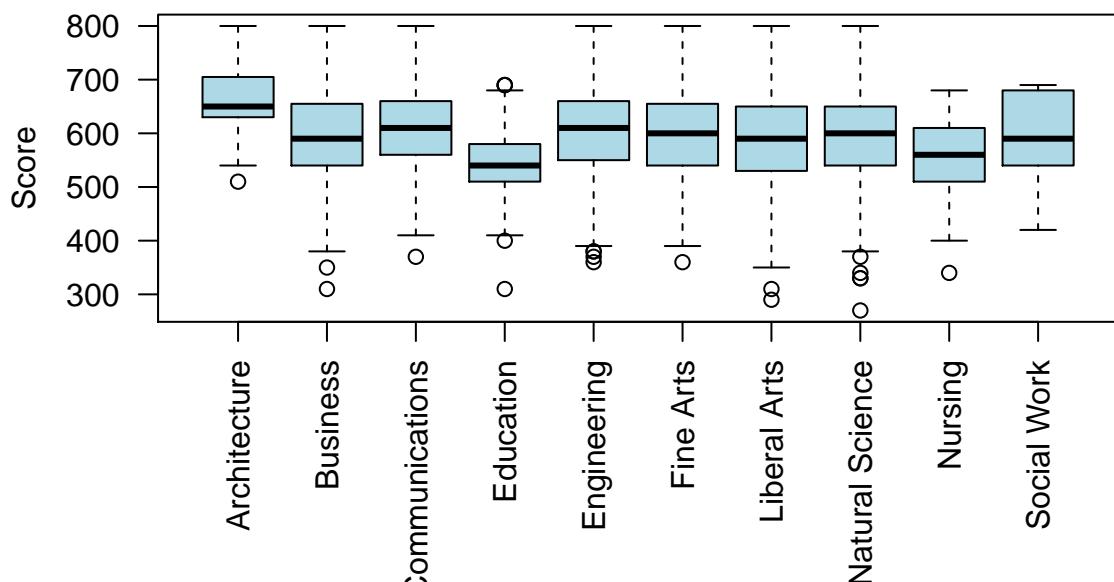


Figure 1.3: Boxplots of the full data set used to form the means in Table 1.4.

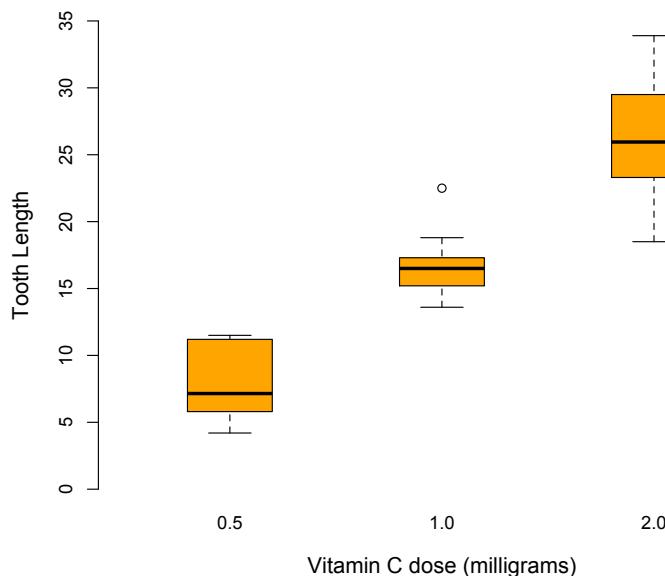


Figure 1.4: For comparison, the table of within-group means is below. Notice how the within-group variability evident in the boxplots at left simply disappears when presented in the form of a summary table, below:

Dose (mg)	Tooth len.
0.5	7.98
1.0	16.77
2.0	26.14

many data sets, it is actually the within-group variability that matters most. For example, as Figure 1.3 shows, SAT scores vary much more within a college as they do between colleges. For example, there is 52-point difference in average SAT math scores between Architecture students and Natural Science students. But within Natural Sciences, the interquartile range is nearly twice as large: 100 points.

The situation is quite different Figure 1.4. These boxplots show the growth of guinea pigs' teeth versus their daily dosage of Vitamin C. Like humans, but unlike most other mammals, guinea pigs need Vitamin C to keep rollin', yet they cannot synthesize their own. Their vitamin C intake is strongly predictive of their overall health, measured in this case by the length of their teeth. In this boxplot, we see comparatively more variability between the groups, whose boxplots almost don't overlap.

The same comparison will come up again and again: between-group variability (the differences between typical or average group members) versus within-group variability (the variation of cases within a single group). We'll soon make this comparison mathematically rigorous, but these examples convey the essence of the idea:

- A UT student's college tells you something, though not ev-



erything, about his or her likely SAT scores.

- A guinea pig's Vitamin C regimen tells you something, though not everything, about its tooth growth. But in a relative sense, it tells you more than a UT student's college tells you about his or her SAT scores.

Always remember that a table of group-wise means does not depict "data" as such, but an abstraction of some typical group member. This abstraction may be useful for some purposes. But within-group variability is also important, and may even be the dominant feature of interest. In this case, presenting the group-wise means alone, without the corresponding plots or measures of variability, may obscure more than it reveals.

### *Dot plots*

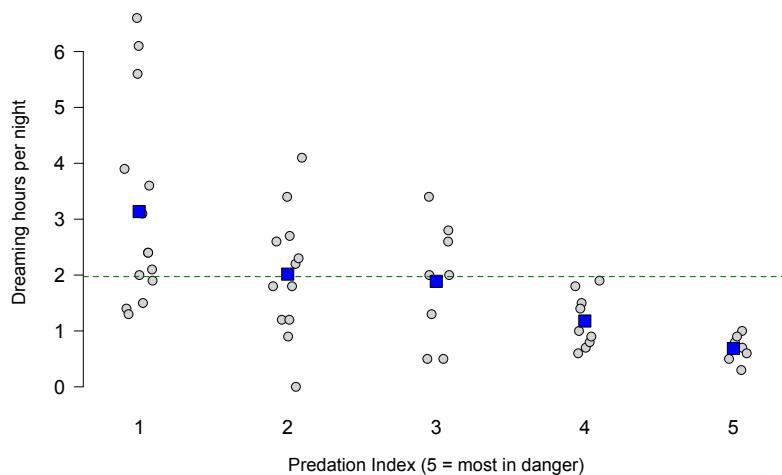
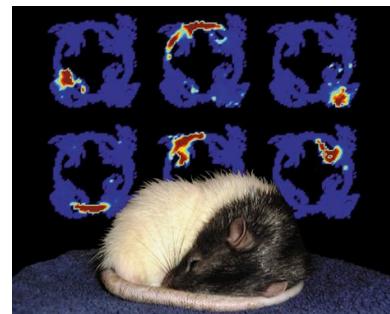
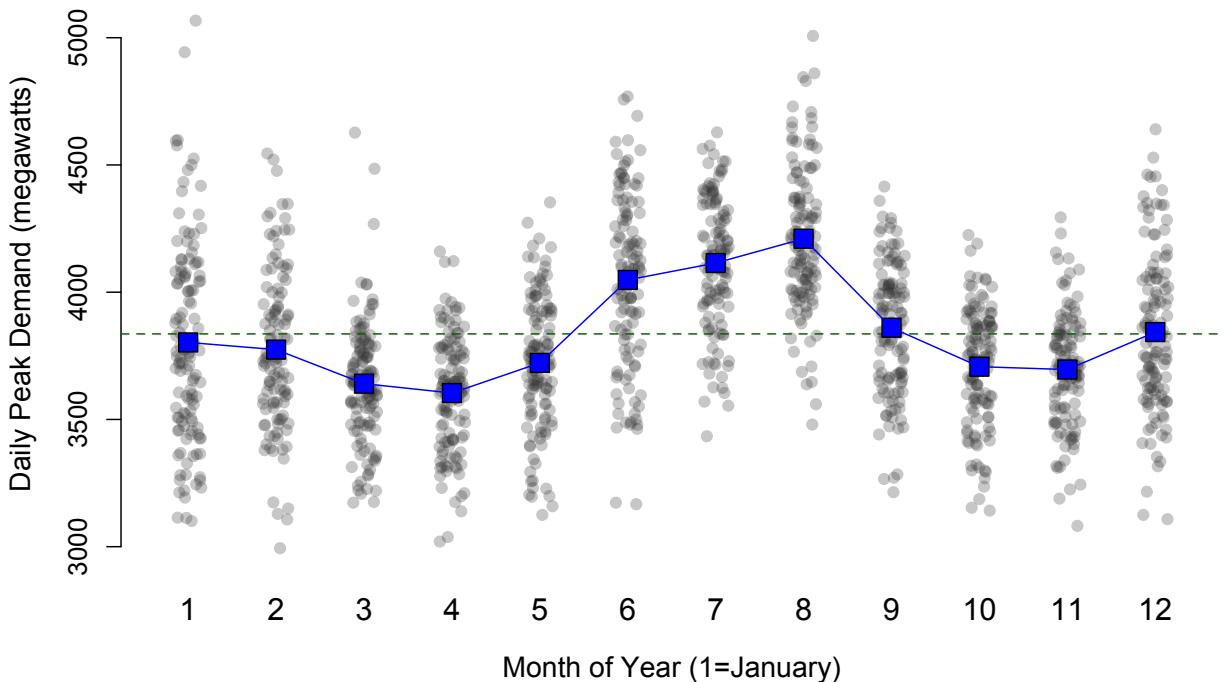


Figure 1.5: Dreaming hours per night versus danger of predation for 50 mammalian species. Data from: "Sleep in Mammals: Ecological and Constitutional Correlates," Allison and Cicchetti (1976). *Science*, November 12, vol. 194, pp. 732-734. Photo of the dreaming critter from the MIT News office ([web.mit.edu/newsoffice/2001/dreaming.html](http://web.mit.edu/newsoffice/2001/dreaming.html)).

The *dot plot* is a close cousin of the boxplot. For example, the plot in Figure 1.5 depicts a relationship between the length of a mammal's dreams (as measured in a lab by an MRI machine) and the severity of the danger it faces from predators. Each dot is a single species of mammal—like, for example, the dreaming critter at right. The predation index is an ordinal variable running from 1 (least danger) to 5 (most danger). It accounts both for how likely an animal is to be preyed upon, and how exposed it is when sleeping. Notice the direction of the trend—you'd sleep poorly too if you were worried about being eaten.





As you can see, the dot plot is useful for small data sets, when a boxplot is no simpler than just plotting the cases group by group. Strictly speaking, the points should all line up vertically with their corresponding values of predation index, on the  $x$ -axis. But a small amount of artificial horizontal jitter has been added to the dots, which allows the eye to distinguish the individual cases more easily.

Dot plots can also be effective for larger data sets. In Figure 1.6 we see four years of data on daily peak electricity demand for the city of Raleigh, NC, stratified by month of the year. Both the between-group and within-group variation show up clearly.

#### *Group means and grand means*

If you looked carefully, you may have noticed two extra features of the dot plots in Figures 1.5 and 1.6. The square blue dots show the *group means* for each category. The dotted green line shows the *grand mean* for the entire data set, irrespective of group identity.

Figure 1.6: Daily peak electricity demand (stratified by month) in Raleigh, NC from 2006–09. The dashed line is the average peak demand for the whole data set, and the blue dots are the month-by-month means.

Notice that, in plotting these means along with the data, we have implicitly partitioned the variability:

$$\text{Individual case} = \text{Group mean} + \text{Deviation of that case}$$

$$\text{Individual case} = \text{Grand mean} + \text{Deviation of group} + \text{Deviation of that case}$$

This is just about the simplest statistical model we can fit, but it's still very powerful. We'll revisit it soon.

## More than one numerical variable

Our basic tool for visualizing a bivariate relationship between two numerical variables is the *scatter plot*. Figure 1.7 shows a plot of the daily returns for Microsoft stock versus Apple stock for every trading day in 2015. Every dot corresponds to a day. The location of the dot along the horizontal axis shows the Apple return, and the location on the vertical axis shows the Microsoft return, for that day. In this case, we can see that Microsoft and Apple stocks tend to move up and down together. (Most stocks do.) We can also see the speckling of outliers: those points that are visibly separate

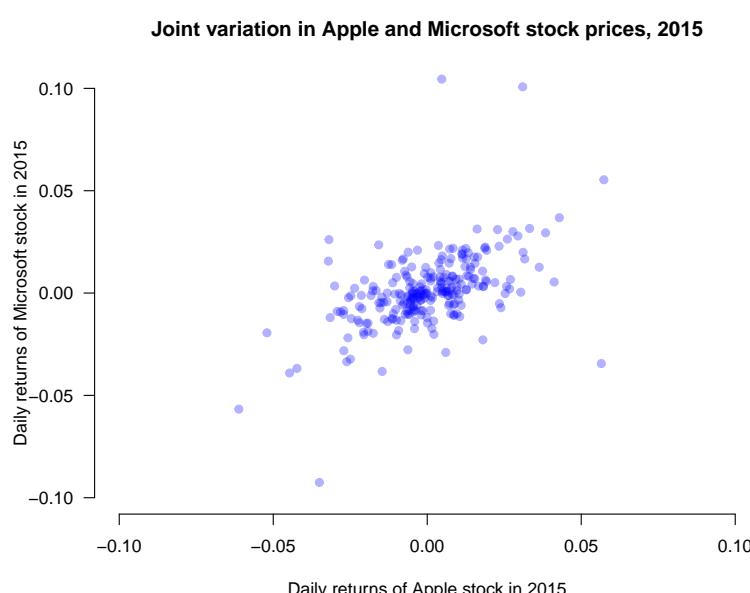


Figure 1.7: A scatter plot of the daily returns for Microsoft stock, versus those of Apple stock, for every trading day in 2015. The daily return is the implied interest rate from holding a stock from the end of one trading day to the end of the next. For example, Apple stock closed at \$105.95 per share on January 7th and at \$110.02 on January 8th. Thus the return for January 8th was

$$\frac{110.02 - 105.95}{105.95} \approx 0.038,$$

or about a 3.8% daily return. On the same day, holders of Microsoft stock enjoyed a 2.9% return.

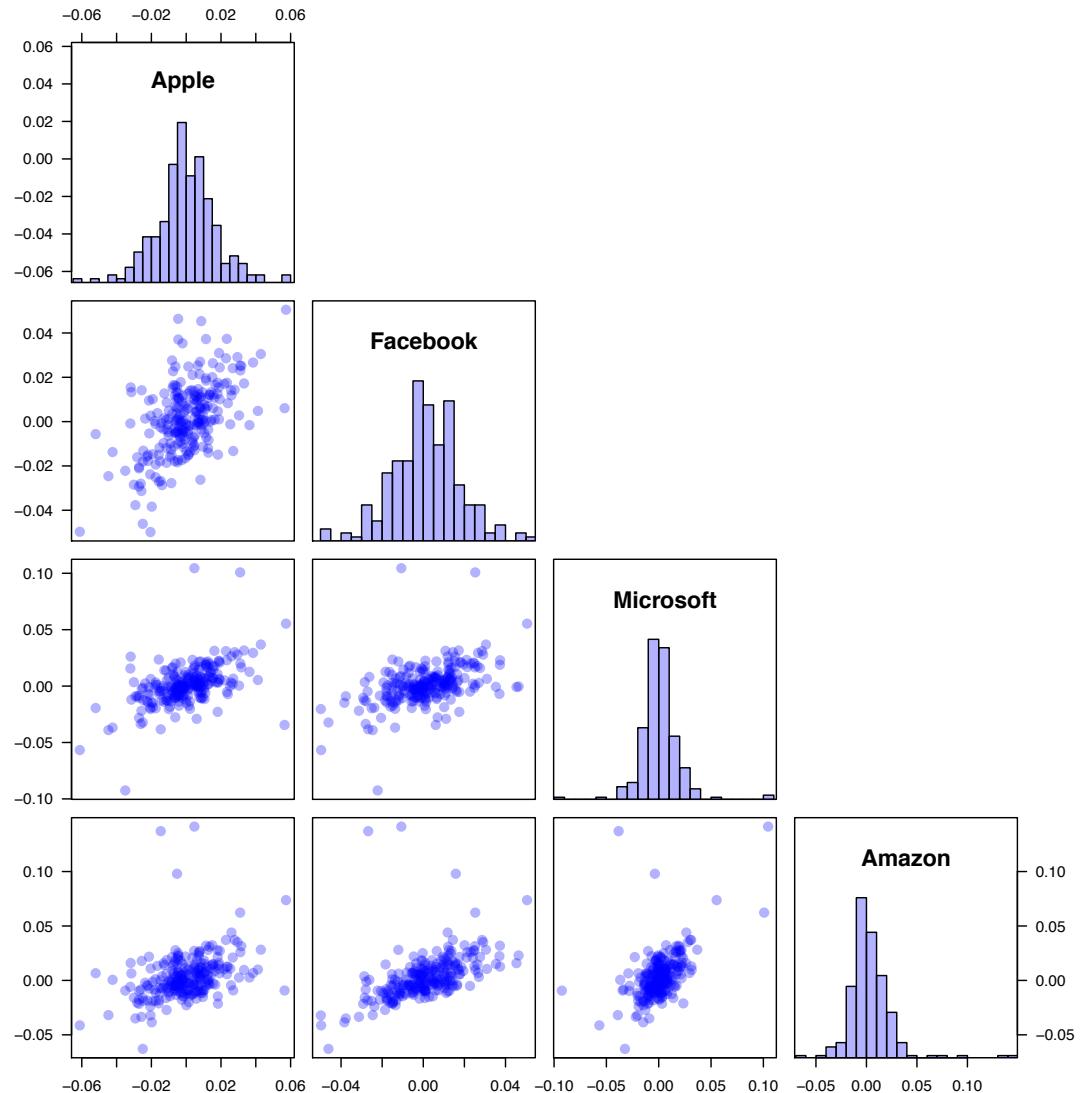


Figure 1.8: A pairs plot: a matrix of four pairwise scatter plots for the daily returns of Apple, Facebook, Microsoft, and Amazon stocks in 2015. The histograms along the diagonal also label the rows and columns of the matrix: e.g. the plot in the second row has Facebook returns along the vertical axis, while the plots in the second column both have Facebook returns along the horizontal axis.

from the main cloud and that represent very good (or bad) days for holders of these two stocks.

A simple way to visualize three or more numerical variables is via a *pairs plot*, as in Figure 1.8. A pairs plot is a matrix of simpler plots, each depicting a bivariate relationship. In Figure 1.8, we see scatterplots for each pair of the daily returns for Microsoft, Facebook, Apple, and Amazon stocks. The histograms on the diagonal serve a dual purpose: (1) they show the variability of each stock in isolation; and (2) they label the rows and columns, so that you know which plots compare which variables.

*Sample correlation.* The *sample correlation coefficient* is a standard measure of the strength of linear dependence between two variables in a sample. If we label the first variable as  $x_1, \dots, x_n$  and the second as  $y_1, \dots, y_n$ , then the correlation coefficient is defined as

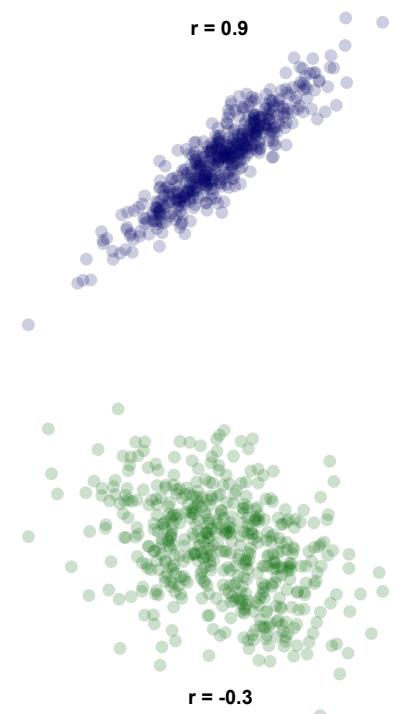
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad (1.2)$$

where  $s_x$  and  $s_y$  are the sample standard deviations of the  $X$  and  $Y$  variables. At right you see scatter plots that depict examples of strong positive (top) and weak negative (bottom) correlation. Sample correlation is between 1 and  $-1$ , which are the extremes of perfect positive and perfect negative correlation.

To summarize the correlation among a set of more than two variables, we typically calculate a *correlation matrix* whose entry in row  $i$ , column  $j$  is the correlation between variable  $i$  and variable  $j$ . For the four stocks depicted in Figure 1.8, the correlation matrix is below. Notice that the matrix is symmetric and has ones along the diagonal (because a variable is perfectly correlated with itself):

	Apple	Microsoft	Facebook	Amazon
Apple	1.00	0.52	0.55	0.36
Microsoft	0.52	1.00	0.47	0.52
Facebook	0.55	0.47	1.00	0.50
Amazon	0.36	0.52	0.50	1.00

*Caveats.* A key fact to remember is that correlation measures the strength of *linear* dependence. If two variables don't fall roughly along a straight line in a scatter plot, then correlation can be misleading. For example, consider Figure 1.9: four different data sets, four different stories about what's going on. Yet all have the same correlation:  $r = 0.816$ .



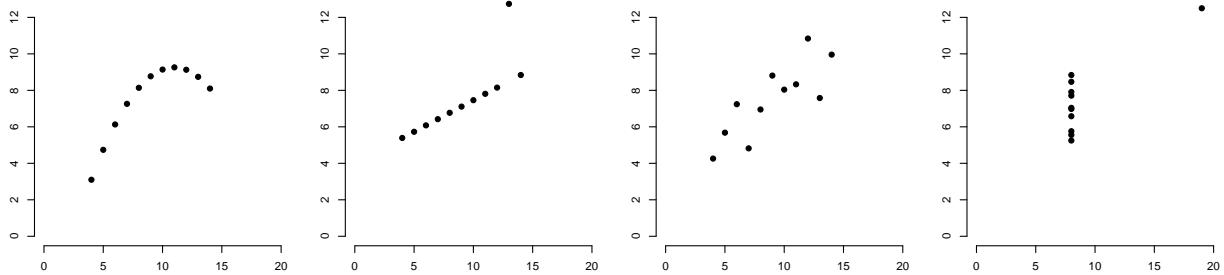


Figure 1.9: above. Data taken from F.J. Anscombe, "Graphs in Statistical Analysis." *American Statistician*, 27 (1973), pp. 17–21

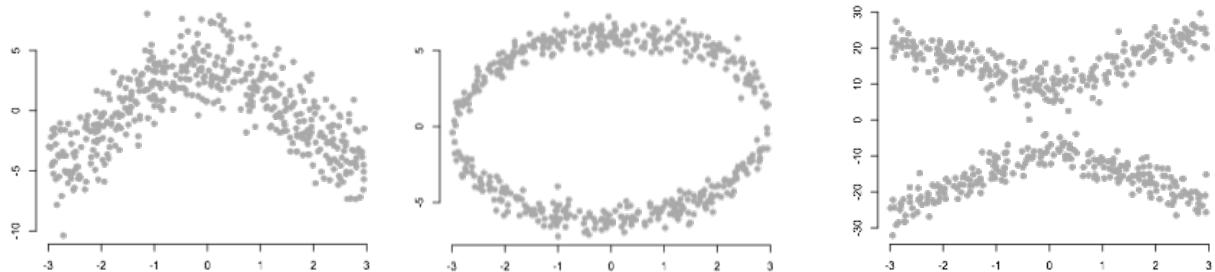


Figure 1.10: left. Each panel shows obvious dependence, but has a sample correlation of  $r = 0$ .

Another important fact is that a sample correlation of 0 ("uncorrelated") does not necessarily mean that two variables are unrelated. In fact, the correlation coefficient is so intimately tied up with the assumption of a linear relationship that it breaks down entirely when used to quantify the strength of nonlinear relationships. In each of the three plots in Figure 1.10, for example, there is an obvious (nonlinear) relationship between the two variables. Yet the sample correlation coefficient for each of them turns out to be exactly zero.

The lesson of these two plots is that you should always plot your data. After all, a sample correlation coefficient is just one number. It can only tell you so much about the relationship between two variables, and a scatterplot (or boxplot, or dot plot) is a much, much richer summary of that relationship.

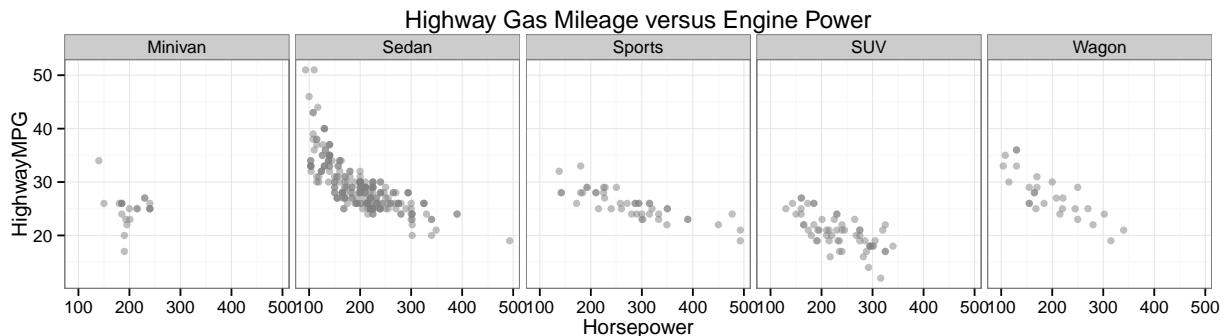


Figure 1.11: Highway gas mileage versus engine power for 387 vehicles in five different classes.

## Further multivariate plots

In visualizing data, we are usually constrained by the limitations of the two-dimensional page or screen. Nevertheless, there are many cool techniques for showing more than two variables at once, despite these limitations.

### Lattice plots

Figure 1.11 shows three variables from a data set on 387 vehicles: the highway gas mileage, the engine power (in horsepower), and the class of the vehicle (minivan, sedan, sports car, SUV, or wagon). This is done via a *lattice plot*, which displays the relationship between two variables, stratified by the value of some third variable. In this case the main relationship of interest is between mileage and engine power, and the stratifying variable is vehicle class. Notice how figure 1.11 repeats a scatterplot of MPG versus horsepower five times: one plot for the vehicles in each class. To facilitate comparisons across the strata, both the horizontal and vertical axes are identical in each plot.

The figure suggests several facts:

- Nobody makes a powerful minivan.
- The overall MPG–horsepower trend is negative for all classes.
- The SUVs have the worst gas mileage overall, and in particular have worse mileage than the sports cars and wagons despite having similar or lower power. (Compare the average vertical location in the SUV panel versus the others).
- The MPG–horsepower relationship becomes nonlinear for

Another term for a lattice plot is a trellis plot.

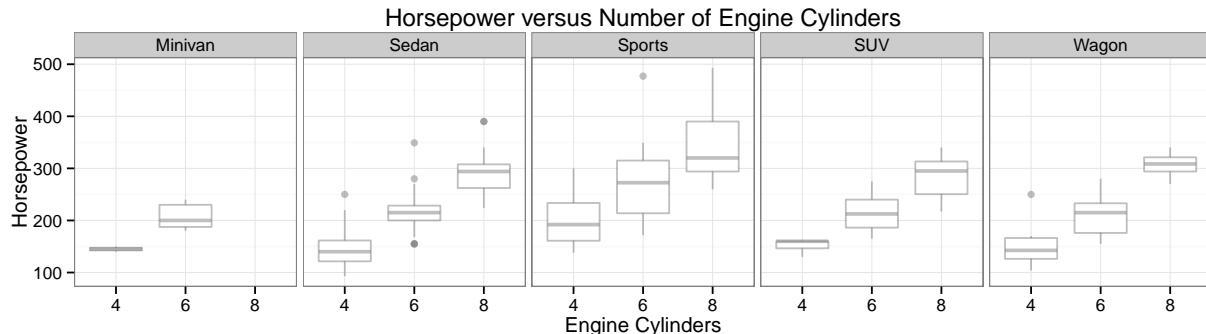


Figure 1.12: Highway gas mileage versus engine power for 387 vehicles in five different classes.

sedans at low horsepower, but perhaps not for wagons.

- As engine power increases, the dropoff in gas mileage looks steeper for SUVs than for sports cars.
- For a fixed level of engine power, there is considerable variability in fuel economy. (Pick a fixed point on the horizontal axis and focus on the cars near there. Now look at the corresponding variability along the vertical axis for those cars.)

We can make a lattice of boxplots as well. For example, Figure 1.12 shows boxplots of engine power versus number of engine cylinders, stratified by vehicle class. This suggests an explanation for the fact that engine power is not a perfect predictor of fuel economy: some cars get more power out of a smaller engine, and are presumably more efficient as a result.

*With a numerical variable.* In Figure 1.11, the stratifying variable is categorical. But we can also stratify a data set according to a numerical variable, by *discretizing* that variable into bins—much in the same way we do when we make a histogram. Figure 1.13 shows the latitude, longitude, and depth (in kilometers) beneath the earth’s surface for the epicenter of every earthquake recorded since 1963 near Fiji, an island in the South Pacific Ocean. The “depth” variable has been discretized into nine equal-length bins. The nine panels show the latitude and longitude of the quakes whose depths fell in each interval, labeled at the top of each panel.

As depth increases (going left to right, top to bottom), a spatial pattern emerges. The shallower earthquakes are at the intersection of two major tectonic plates. The deeper quakes emanate from the Tonga Trench—35,702 feet below the sea at its deepest point.<sup>9</sup>

<sup>9</sup> And the final resting place of 3.9 kilograms of radioactive plutonium-238 from the ill-fated Apollo 13 mission.

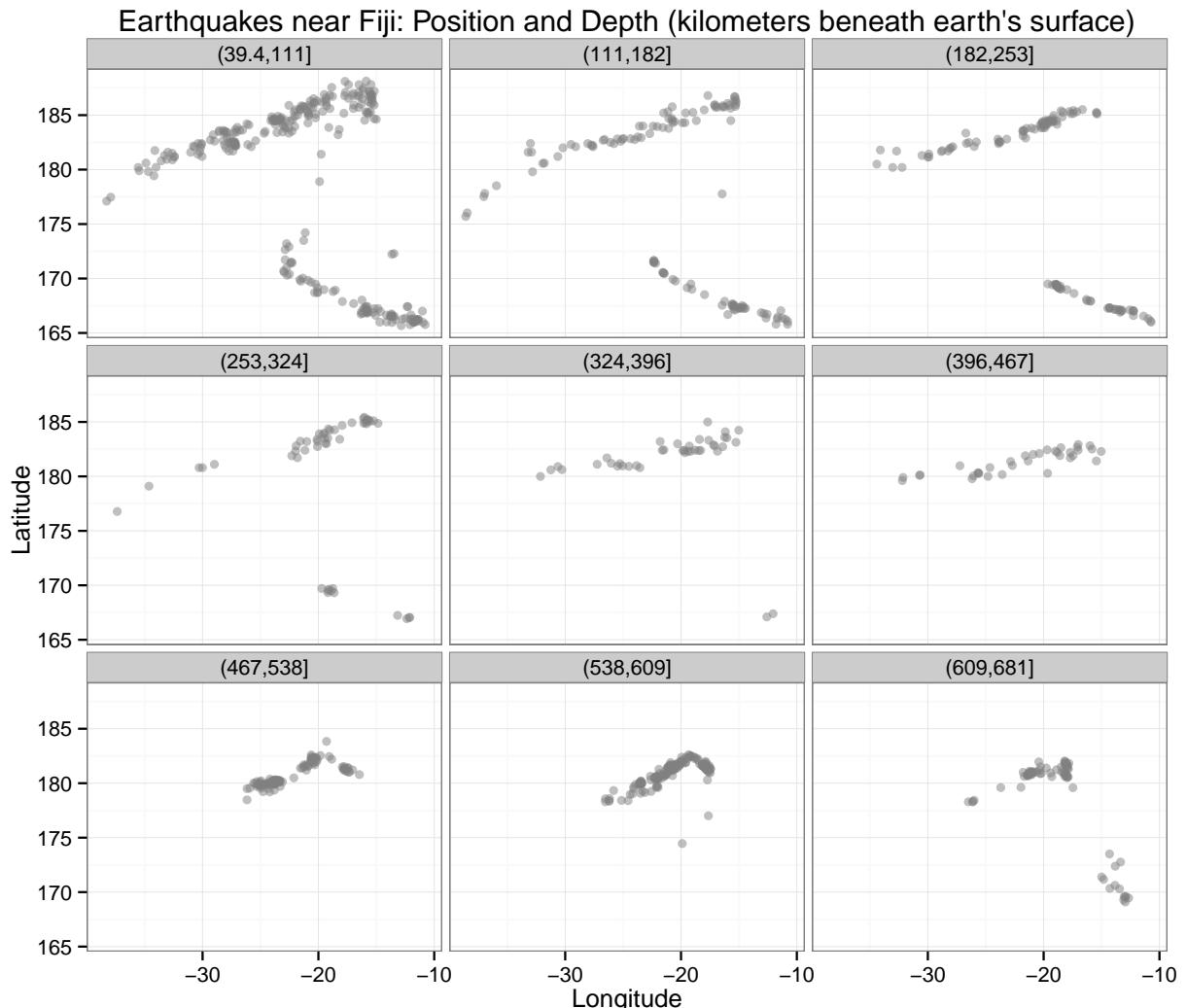


Figure 1.13: Earthquakes in Fiji: latitude versus longitude for quakes within each of nine different depth intervals. Here the range of depths beneath the earth's surface (in kilometers) is labeled at the top of each panel.



## 2

# Fitting equations to data

SO FAR we've concentrated on relatively simple visual and numerical summaries of data sets. In many cases we will want to go further, by fitting an explicit equation—usually called a *regression model*—that describes how one variable changes as a function of some other variables. There are many reasons we might want to do this. Here are four that we'll explore at length:

- to make a forecast;
- to summarize the trend in a data set;
- to make comparisons that adjust statistically for some systematic effect; and
- to quantify the amount of variability in some variable that cannot be predicted, in the context of what *can* be predicted.

This chapter introduces the idea of a regression model and builds upon these themes.

## Fitting straight lines

As a running example we'll use the data from Figure 2.1, which depicts a sample of 104 restaurants in the vicinity of downtown Austin, Texas. The horizontal axis shows the restaurant's "food deliciousness" rating on a scale of 0 to 10, as judged by the writers of a popular guide book entitled *Fearless Critic: Austin*. The vertical axis shows the typical price of a meal for one at that restaurant, including tax, tip, and drinks. The line superimposed on the scatter plot captures the overall "bottom-left to upper-right" trend in the data, in the form of an equation: in this case,  $y = -6.2 + 7.9x$ . On average, it appears that people pay more for tastier food.

This is our first of many data sets where the response (price,  $Y$ ) and predictor (food score,  $X$ ) can be described by a linear regression model. We write the model in two parts as " $Y = \beta_0 + \beta_1 X + \text{noise}$ ." The first part, the function  $\beta_0 + \beta_1 X$ , is called the *linear predictor*—linear because it is the equation of a straight

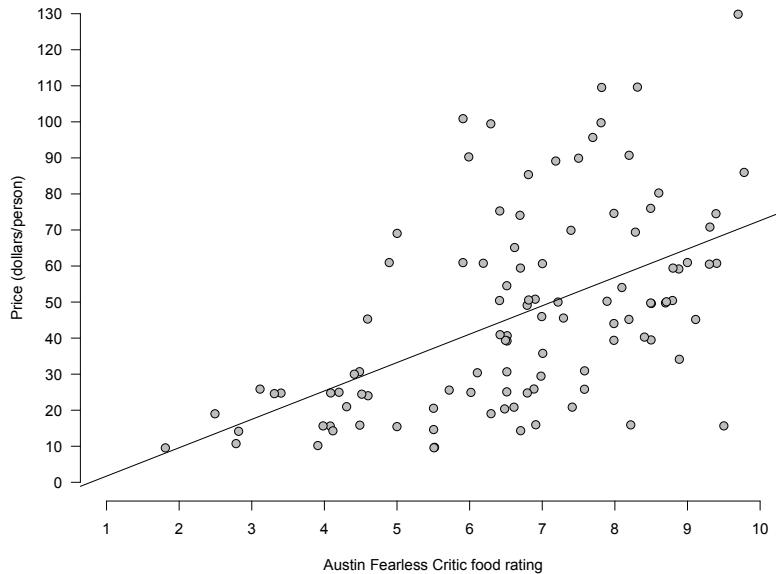


Figure 2.1: Price versus reviewer food rating for a sample of 104 restaurants near downtown Austin, Texas. The data are from a larger sample of 317 restaurants from across greater Austin, but downtown-area restaurants were chosen to hold location relatively constant. Data from Austin Fearless Critic, [www.fearlesscritic.com/austin](http://www.fearlesscritic.com/austin). Because of ties in the data, a small vertical jitter was added for plotting purposes only. The equation of the line drawn here is  $y = -6.2 + 7.9x$ .

line, predictor because it predicts  $Y$ . The second part, the noise, is a crucial part of the model, too, since no line will fit the data perfectly. In fact, we usually denote each individual noise term explicitly:

$$y_i = \beta_0 + \beta_1 x_i + e_i. \quad (2.1)$$

An equation like (2.1) is our first example of a regression model. The *intercept*  $\beta_0$  and the *slope*  $\beta_1$  are called the *parameters* of the regression model. They provide a mathematical description of how price changes as a function of food score. The little  $e_i$  is called the *residual* for the  $i$ th case—residual, because it's how much the line misses the  $i$ th case by (in the vertical direction). The residual is also a fundamental part of the regression model: it's what's “left over” in  $y$  after accounting for the contribution of  $x$ .

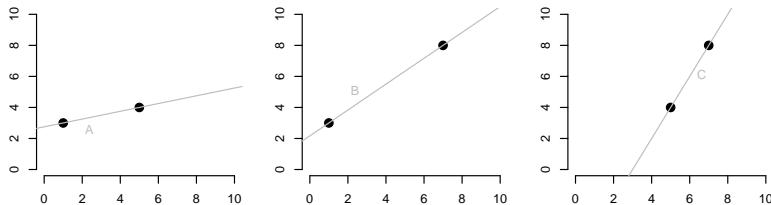
*For every two points. . . .*

A natural question is: how do we fit the parameters  $\beta_0$  and  $\beta_1$  to the observed data? Historically, the standard approach, still in widespread use today, is to use the method of least squares. This involves choosing  $\beta_0$  and  $\beta_1$  so that the sum of squared residuals (the  $e_i$ 's) will be as small as possible. This is what we did to get the equation  $y_i = -6.2 + 7.9x_i$  in Figure 2.1.

The method of least squares is one of those ideas that, once

you've encountered it, seems beautifully simple, almost to the point of being obvious. But it's worth pausing to consider its historical origins, for it was far from obvious to a large number of very bright 18th-century scientists.

To see the issue, consider the following three simple data sets. Each has only two observations, and therefore little controversy about the best-fitting linear trend.



For every two points, a line. If life were always this simple, there would be no need for statistics.

But things are more complicated if we observe three points.

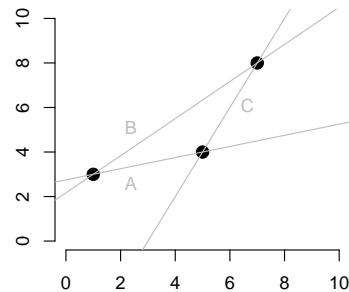
$$\begin{aligned} 3 &= \beta_0 + 1\beta_1 \\ 4 &= \beta_0 + 5\beta_1 \\ 8 &= \beta_0 + 7\beta_1 \end{aligned}$$

Two unknowns, three equations. There is no solution for the parameters  $\beta_0$  and  $\beta_1$  that satisfies all three equations—and therefore no perfectly fitting linear trend exists. Seen graphically, at right, it is clear that no line can pass through all three points.

Abstracting a bit, the key issue here is the following: how are we to combine inconsistent observations? Any two points are consistent with a unique line. But three points usually won't be, and most interesting data sets have far more than three data points.

Therefore, if we want to fit a line to the data anyway, we must allow the line to miss by a little bit for each  $(x_i, y_i)$  pair. We express these small misses mathematically, as follows:

$$\begin{aligned} 3 &= \beta_0 + 1\beta_1 + e_1 \\ 4 &= \beta_0 + 5\beta_1 + e_2 \\ 8 &= \beta_0 + 7\beta_1 + e_3. \end{aligned}$$



The three little  $e$ 's are the residuals, or misses.

But now we've created a different predicament. Before we added the  $e_i$ 's to give us some wiggle room, there was no solution

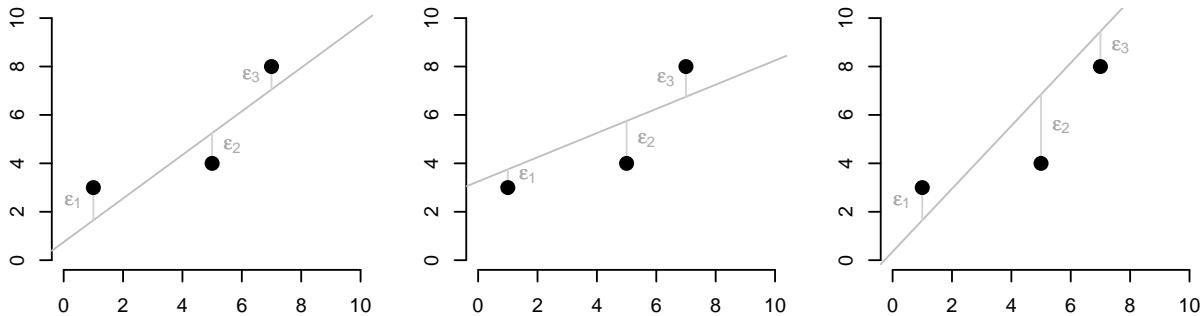


Figure 2.2: Three possible straight-line fits, each involving an attempt to distribute the “errors” among the observations.

to our system of linear equations. Now we have three equations and five unknowns: an intercept, a slope, and three residuals. This system has infinitely many solutions. How are we to choose, for example, among the three lines in Figure 2.2? When we change the parameters of the line, we change the residuals, thereby redistributing the errors among the different points. How can this be done sensibly?

Believe it or not, scientists of the 1700’s struggled mightily with this question. Many of the central scientific problems of this era concerned the combination of astronomical or geophysical observations. Astronomy in particular was a hugely important subject for the major naval powers of the day, since their ships all navigated using maps, the stars, the sun, and the moon. Indeed, until the invention of a clock that would work on the deck of a ship rolling to and fro with the ocean’s waves, the most practical way for a ship’s navigator to establish his longitude was to use a lunar table. This table charted the position of the moon against the “fixed” heavens above, and could be used in a roundabout fashion to compute longitude. These lunar tables were compiled by fitting an equation to observations of the moon’s orbit.

The same problem of fitting astronomical orbits arose in a wide variety of situations. Many proposals for actually fitting the equation to the data were floated, some by very eminent mathematicians. Leonhard Euler, for example, proposed a method for fitting lines to observations of Saturn and Jupiter that history largely judges to be a failure.

In fact, some thinkers of this period disputed that it was even a good idea to combine observations at all. Their reasoning was, roughly, that the “bad” observations in your sample would corrupt

the “good” ones, resulting in an inferior final answer. To borrow the phrase of Stephen Stigler, an historian of statistics, the “deceptively simple concept” that combining observations would improve accuracy, not compromise it, was very slow to catch on during the eighteenth century.<sup>1</sup>

<sup>1</sup> *The History of Statistics*, p. 15.

### *The method of least squares*

No standard method for fitting straight lines to data emerged until the early 1800’s, half a century after scientists first entertained the idea of combining observations. What changed things was the *method of least squares*, independently invented by two people. Legendre was the first person to publish the method, in 1805, although Gauss claimed to have been using it as early as 1794.

The term “method of least squares” is a direct translation of Legendre’s phrase “méthode des moindres carrés.” The idea is simple: choose the parameters of the regression line that minimize  $\sum_{i=1}^n e_i^2$ , the sum of the squared residuals. As Legendre put it:

In most investigations where the object is to deduce the most accurate possible results from observational measurements, we are led to a system of equations of the form

$$E = a + bx + cy + fz + \&c.,$$

in which  $a, b, c, f, \&c.$  are known coefficients, varying from one equation to the other, and  $x, y, z, \&c.$  are unknown quantities, to be determined by the condition that each value of  $E$  is reduced either to zero, or to a very small quantity. . . .

Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of the squares of the errors a minimum. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.<sup>2</sup>

The utility of Legendre’s suggestion was immediately obvious to his fellow scientists and mathematicians. Within two decades, least squares became the dominant method throughout the European scientific community.

Why was the principle adopted so quickly and comprehensively? For one thing, it offered the attractiveness of a single best answer, evaluated according to a specific, measurable criterion. This gave the procedure the appearance of objectivity—especially

<sup>2</sup> Adrien-Marie Legendre (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*. Translation p. 13, Stigler’s *A History of Statistics*.

compared with previous proposals, many of which essentially amounted to: “muddle around with the residuals until you get an acceptable balance of errors among the points in your sample.”

Moreover, unlike many previous proposals for combining observations, the least-squares criterion could actually be applied to non-trivially large problems. One of the many advantages of the least-squares idea is that it leads immediately from grand principle to specific instructions on how to compute the estimate  $(\hat{\beta}_0, \hat{\beta}_1)$ :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.2)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (2.3)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means of the  $X$  and  $Y$  variables, respectively. The line  $y = \hat{\beta}_0 + \hat{\beta}_1 x$  is the best possible linear fit to the data, in a squared-error sense. That is to say: among the family of all possible straight-line fits to the data, this particular line has the smallest sum of squared residuals. Deriving this solution involves solving a simple mathematical problem involving some calculus and matrix algebra—something that scientists of the nineteenth century could do easily, via pen and paper.

In statistics, a little hat on top of something usually denotes a guess or an estimate of the thing wearing the hat.

## Goals of regression analysis

WITH modern computers, the estimation of linear regression models by least squares is now entirely automatic for all but the very largest of data sets.<sup>3</sup> It’s so ordinary, in fact, that the method is often abbreviated as OLS: ordinary least squares.

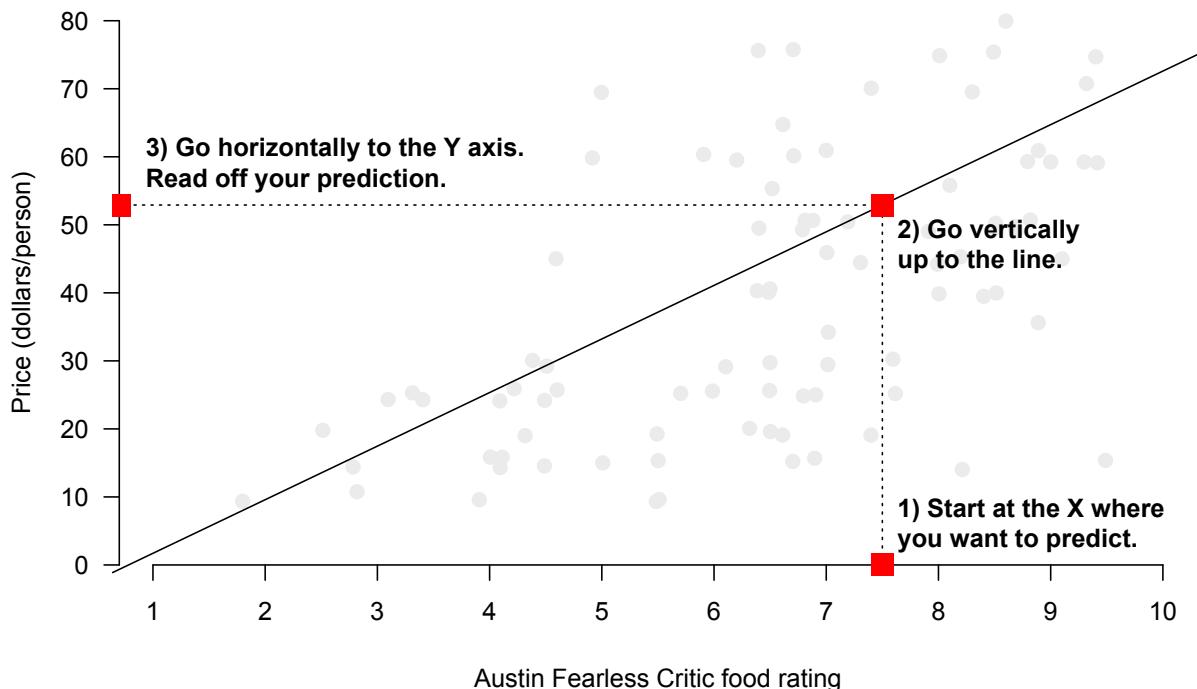
But don’t let the simplicity of the model-fitting step fool you: regression modeling is a wonderfully rich and complex subject. We’ll start by focusing on four kinds of stories one can tell with a regression model. Each is useful for a different purpose.

*Story 1: A regression model is a plug-in prediction machine.*

One way to interpret a regression model is as is a function  $\hat{y} = f(x)$  that maps inputs ( $x$ ) to expected outputs ( $\hat{y}$ ). When we plug in the original  $x$  values in to the least-squares equation, we get back the so-called *fitted values*, or *model values*, denoted  $\hat{y}_i$ :

$$\hat{y}_i = \hat{\beta}_0 + x_i \hat{\beta}_1. \quad (2.4)$$

<sup>3</sup> By “very largest,” think: every search that Google has every recorded, every post in the history of Facebook, and so forth. It’s still possible to fit regression models to those data sets, but doing so is far from automatic—and possessing the expertise necessary to do so is a large part of what makes the major Silicon Valley companies so extraordinary (and so valuable).



In this way, the regression model partitions each observed  $y$  value into two pieces:  $y_i = \hat{y}_i + e_i$ , a fitted value plus a residual.

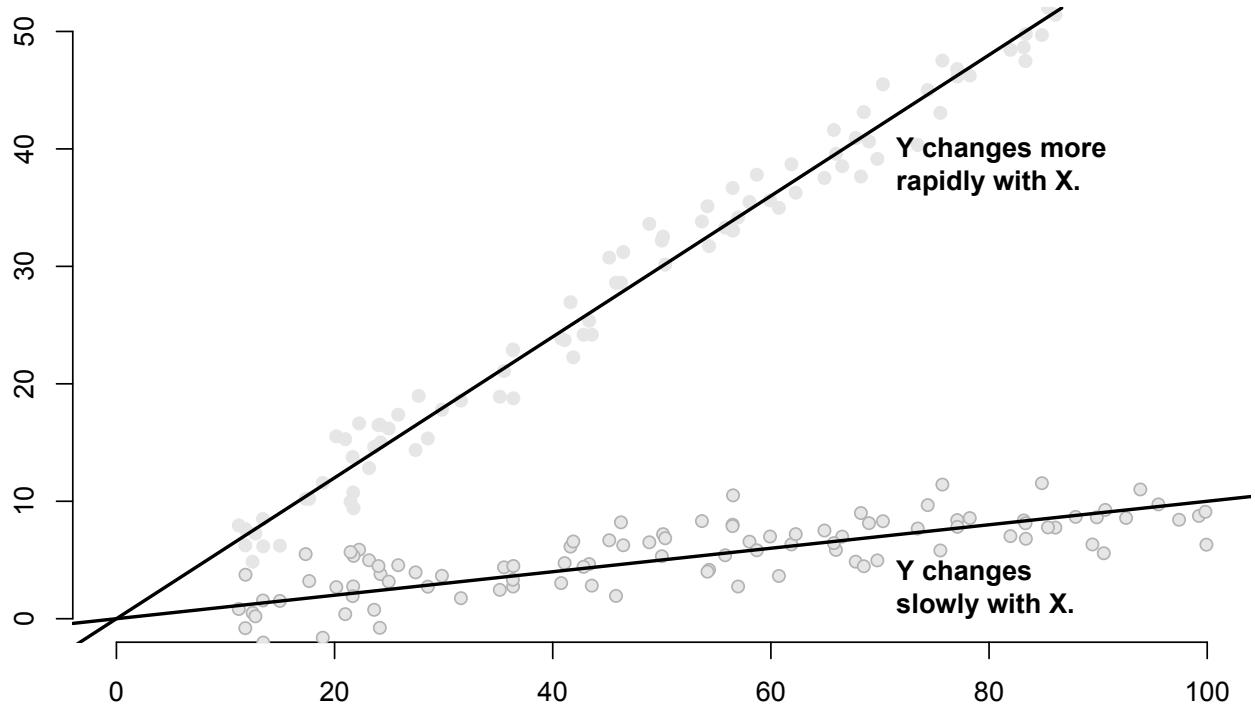
This is especially useful forecasting the response of a new case, where we know the value of the predictor but not the response. Specifically, if we see a new observation  $x^*$  and want to predict where the corresponding  $y^*$  will be, we can simply plug in  $x^*$  and read off our guess for  $y^*$  from the line:  $\hat{y}^* = \hat{\beta}_0 + x^* \hat{\beta}_1$ .

For example, if we know that a new restaurant earned a food rating of 7.5, our best guess for the cost of the meal—knowing nothing else about the restaurant—would be to use the linear predictor:  $\hat{y}^* = -6.2 + 7.9 \cdot 7.5$ , or \$53.05 per person. (See Figure 2.3). This, incidentally, is where the name *regression* comes from: we expect that future  $y$ 's will “regress to the mean” specified by the linear predictor.

Figure 2.3: Using a regression model for plug-in prediction of the price of a meal, assuming a food rating of 7.5.

*Story 2: A regression model summarizes the trend in the data.*

The linear predictor tells you how  $Y$  changes, on average, as a function of  $X$ . In particular, the slope  $\beta_1$  tells you how the re-



sponse tends to change as a function of the predictor:

$$\beta_1 = \frac{\Delta Y}{\Delta X},$$

read “delta-Y over delta-X,” or “change in Y over change in X.”

For the line drawn in Figure 2.1, the slope is  $\beta_1 = 7.9$ . On average, then, one extra Fearless Critic food rating point ( $\Delta X$ ) is associated with an average increase of \$7.90 ( $\Delta Y$ ) in the price of a meal. The slope is always measured in units of  $Y$  per units of  $X$ —in this case, dollars per rating point. It is often called the *coefficient* of  $X$ .

To interpret the intercept, try plugging in  $x_i = 0$  into the regression model and notice what you get for the linear predictor:  $\beta_0 + \beta_1 \cdot 0 = \beta_0$ . This tells you that the intercept  $\beta_0$  is what we’d expect from the response if the predictor were exactly 0.

Sometimes the intercept is easily interpretable, and sometimes it isn’t. Take the trend line in Figure 2.1, where the intercept is  $\beta_0 = -6.2$ . This implies that a restaurant with a Fearless Critic food rating of  $x = 0$  would charge, on average,  $y = -\$6.20$  for the privilege of serving you a meal.

Perhaps the diners at such an appalling restaurant would feel

Figure 2.4: The slope of a regression model summarizes how fast the  $Y$  variable changes, as a function of  $X$ .

Generally we use a capital letter when referring generically to the predictor or response variable, and a lower-case letter when referring to a specific value taken on by either one.

this is fair value. But a negative price is obvious nonsense. Plugging in  $x = 0$  to the price/rating model and trying to interpret the result is a good example of why extrapolation—using a regression model to forecast far outside the bounds of past experience—can give silly results.

*Story 3: A regression model takes the X-ness out of Y.*

We've seen how a regression model splits up every observation in the sample into two pieces, a fitted value ( $\beta_0 + \beta_1 x_i$ ) and a residual ( $e_i$ ):

$$\text{Observed } y \text{ value} = (\text{Fitted value}) + (\text{Residual}), \quad (2.5)$$

or equivalently,

$$\text{Residual} = (\text{Observed } y \text{ value}) - (\text{Fitted value}).$$

The residuals from a regression model are sometimes called “errors.” This is especially true in experimental science, where measurements of some  $Y$  variable will be taken at different values of the  $X$  variable (called design points), and where noisy measurement instruments can introduce random errors into the observations.

But in many cases this interpretation of a residual as an error can be misleading. A regression model can still give a nonzero residual, even if there is no mistake in the measurement of the  $Y$  variable. It's often far more illuminating to think of the residual as the part of the  $Y$  variable that it is left unpredicted by  $X$ .

In Figure 2.1, for example, the positive slope of the line says: yes, people generally pay more for tastier food. The residuals say: not always. There are many other factors affecting the price of a restaurant meal in Austin: location, service, decor, drinks, the likelihood that Matthew McConaughey will be eating overpriced tacos in the next booth, and so forth. Our simple model of price versus food rating collapses all of these other factors into the residuals.

A good way of summarizing this is that the regression model “takes the  $X$ -ness out of  $Y$ ,” leaving what remains in the residual  $e_i$ :

$$\underbrace{y_i}_{\text{Observed } y \text{ value}} = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Predictable by } x} + \underbrace{e_i}_{\text{Unpredictable by } x}.$$

This is easily seen in our example by plotting the residual price ( $e_i$ ) against food rating ( $x_i$ ), side by side with the original data, as in Figure 2.5. In the right panel, there is no evident correlation

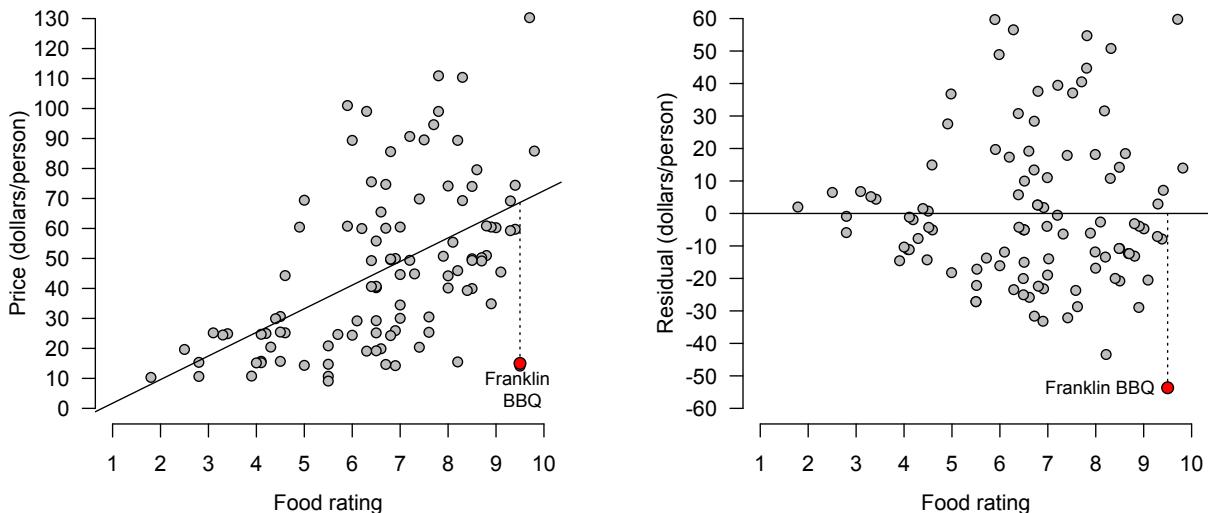


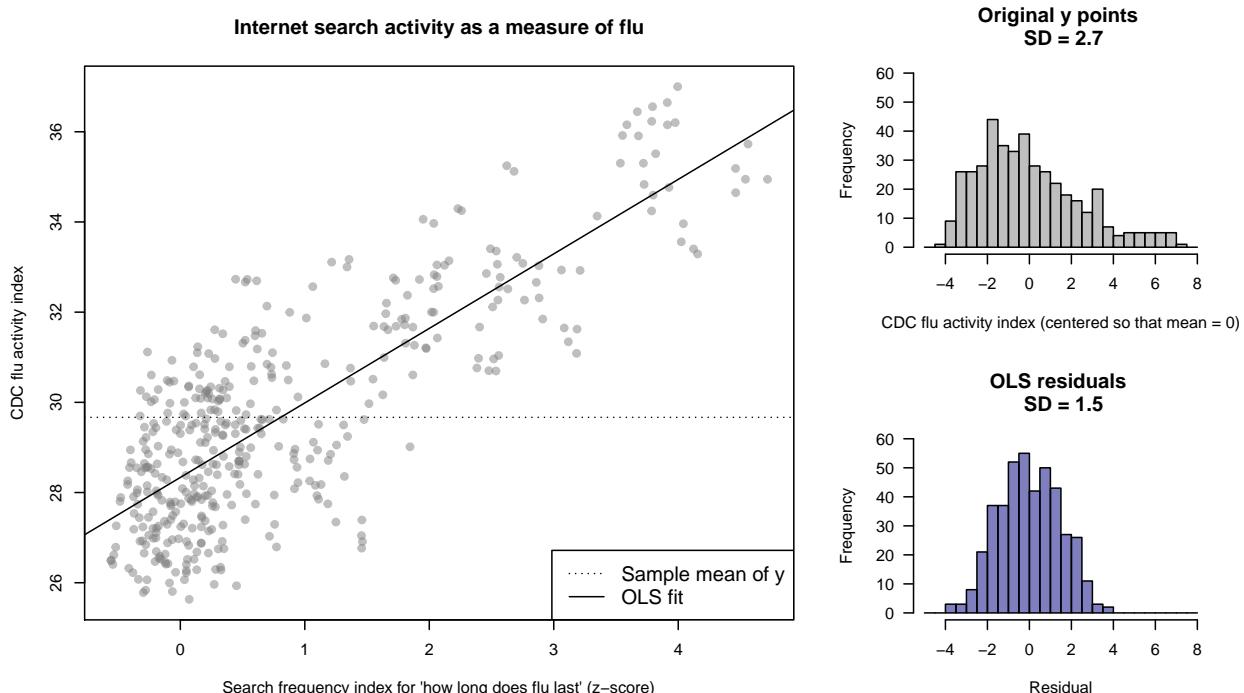
Figure 2.5: Left: the original data on price versus food rating. Right: the residuals from the least squares fit on the left. The residual for Franklin BBQ is the length of the dotted vertical line:  $e_i = -\$53.85$ .

between food rating and the residuals. This should always be true: a good regression model should take the  $X$ -ness out of  $Y$ , so that the residuals look independent of the predictor. If they don't, then the model hasn't done its job. (Always plot your residuals to check this.)

You've just seen your first example of *statistical adjustment*. Notice the red dot sitting in the lower right of Figure 2.5, with a low price and a high food rating? This isn't the least expensive restaurant near downtown Austin in an absolute sense. But it is the least expensive *after we adjust for food rating*. To do this, we simply subtract off the fitted value from the observed value of  $y$ , leaving the residual—which, you'll recall, captures what's over in the response (price) after the predictor (food score) has been taken into account. The restaurant in question has a food rating of 9.5, good for *Fearless Critic*'s third best score in the entire city. For such delicious food, you would expect to pay  $\hat{y}^* = -6.2 + 7.9 \cdot 9.5$ , or \$68.85 per person. In reality, the price of a meal at this restaurant is a mere \$15, or  $e_i = -\$53.85$  less than expected. That's the largest, in absolute value, of all the negative residuals.

This restaurant is Franklin Barbecue, declared "Best Barbecue in America" by *Bon Appétit* magazine, and undoubtedly the most delicious residual in the city:

Go to Austin and queue up at Franklin Barbecue by 10:30 a.m.  
When you get to the counter, Aaron Franklin will be waiting,



knife in hand, ready to slice up his brisket. (Order the fatty end.) Grab a table, a few beers, and lots of napkins and dig in. Take a bite, and don't tell me you're not convinced you've reached the BBQ promised land.

But visitors take note: this article ("A Day in the Life of a BBQ Genius," by food critic Andrew Knowlton) is from July of 2011, and its advice is dated. These days, queueing up at 10:30 would have you last in line!

*Story 4: A regression model quantifies the information in a predictor.*

The idea behind the Flu Prediction Project, run jointly by IBM Watson and the University of Osnabrück in Germany, is simple.<sup>4</sup> Researchers combine social-media and internet-search data, together with official data provided by government authorities, like the Centers for Disease Control (CDC) in the United States, to yield accurate real-time predictions about the spread of seasonal influenza. This kind of forecasting model allows public-health authorities to allocate resources (like antivirals and flu vaccines)

Figure 2.6: A scatter plot of the CDC's measure of flu activity versus Google search activity for the phrase "how long does flu last" (z score of search frequency). To the right of the scatter plot, we see two dot plots, both on the same scale: (1) the original deviations from the sample mean,  $y_i - \bar{y}$ ; and (2) the residuals from the regression equation,  $y_i - \hat{y}_i$ .

<sup>4</sup> <http://www.flu-prediction.com>

using the most up-to-date information possible. After all, the official government data can usually tell you what flu activity was like two weeks ago. Social-media and internet-search data, if used correctly, have the potential to tell what you it's like right now.

To give you a sense of how strong the predictive signal from internet-search data can be, examine Figure 2.6, focusing first on the scatter plot in the left panel. Here each dot corresponds to a day. On the  $x$ -axis is a measure of Google search activity for the term "how long does flu last," where higher numbers mean that more people are searching for that term on that day.<sup>5</sup> On the  $y$  axis, we see a measure of actual flu activity on that day, constructed from data provided by the CDC.

The search activity on a given day strongly predicts actual flu transmission, which makes sense: one of the first things that many people do when they fall ill is to commiserate with a search engine about the depth and duration of their suffering. But just how much information about flu does the search activity for this single term—"how long does flu last"—convey?

In principle, there are many ways of measuring this information content. In fact, you've already met one way to do so: by computing the correlation coefficient between the two variables. Our regression model provides another way, because it allows us to compare our predictions of flu activity both with and without the  $x$  variable.

- Without knowing the predictor variable, our best guess for the outcome is just the sample mean,  $\bar{y}$ , and the prediction error for each case is  $y_i - \bar{y}$ . You can think of the sample mean as our "baseline" prediction; it is obviously a pretty simple baseline.
- With the predictor variable, our best guess is given by the regression model,  $\hat{y}_i = \beta_0 + \beta_1 x_i$ , and the prediction error for each case is the residual,  $y_i - \hat{y}_i$ .

In each case, we would expect these errors to be distributed around zero. The question is: how much smaller do the errors of the regression model tend to be, compared with the errors we make by predicting the outcome using the sample mean alone? If our predictions errors get a lot smaller with the  $x$  variable than without it, then we'll know that this variable conveys a lot of information about response.

To answer this question, return to Figure 2.6. To the right of

<sup>5</sup> Specifically, it's a  $z$  score: how many standard deviations about the mean was the search frequency on that day for that particular term.

### LEAST SQUARES THEN AND NOW: AN HISTORICAL ASIDE

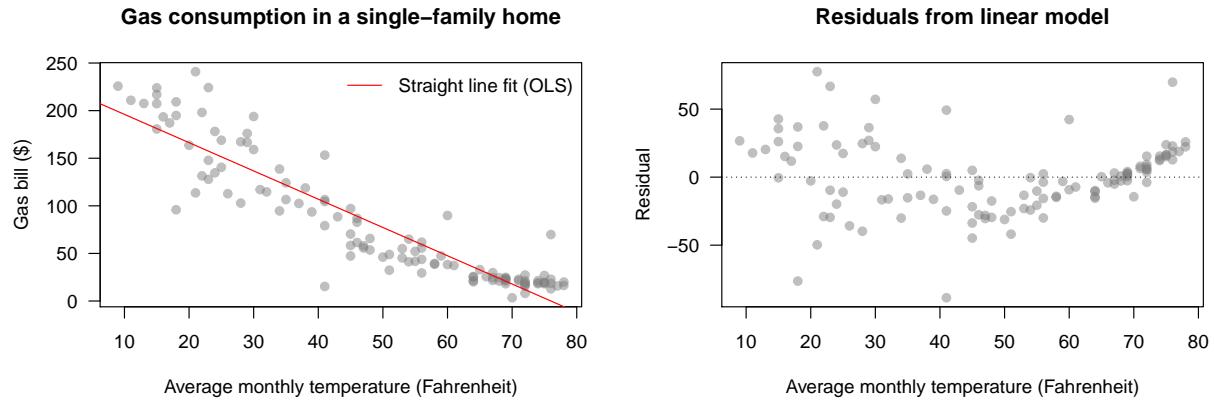
The Ordnance Survey is the governmental body in the United Kingdom charged with mapping and surveying the British Isles. “Ordnance” is a curious name for a map-making body, but it has its roots in the military campaigns of the 1700’s. The name just stuck, despite the fact that these days, most of the folks that use Ordnance Survey maps are probably hikers.



In the days before satellites and computers, map-making was a grueling job, both on the soles of your feet and on the pads of your fingers. Cartographers basically walked and took notes, and walked and took notes, ad infinitum. In the 1819 survey, for example, the lead cartographer, Major Thomas Colby, endured a 22-day stretch where he walked 586 miles—that’s 28 miles per day, all in the name of precision cartography. Of course, that was just the walking. Then the surveyors would have to go back home and crunch the numbers that allowed them to calculate a consistent set of elevations, so that they could correctly specify the contours on their maps.

They did the number-crunching, moreover, by hand. This is a task that would make most of us weep at the drudgery. In the 1858 survey, for example, the main effort involved reducing an enormous mass of elevation data to a system of 1554 linear equations involving 920 unknown variables, which the Ordnance Survey mathematicians solved using the principle of least squares. To crunch their numbers, they hired two teams of dozens of human computers each, and had them work in duplicate to check each other’s mistakes. It took them two and a half years to reach a solution.

A cheap laptop computer bought today takes a few seconds to solve the same problem.



the scatter plot you see two histograms: (1) the original deviations  $y_i - \bar{y}$ , and (2) the residuals from the regression model. You'll notice that some of the original variation has been absorbed by the regression model: the residuals are less variable (standard deviation 1.5) than the original  $y$  points (standard deviation 2.7).

This is how a regression model measures the information content of a predictor: information means reduction in prediction error for the response. The bigger this reduction in prediction uncertainty, the more informative the predictor.

Figure 2.7: Left: a scatterplot of monthly gas consumption (measured in dollars) versus average monthly temperature at a single-family home in Minnesota, together with a linear regression model fit by ordinary least squares. Right: a plot of the residuals from the linear model versus temperature, showing the deficits of the straight line fit. Data source: Daniel T. Kaplan, *Statistical Modeling: A Fresh Approach*, 2009.

## Beyond straight lines

UP TO this point, we've talked about fitting straight lines using the principle of least squares. For many data sets, however, a linear regression model doesn't provide an adequate description of what's going on. Consider, for example, the data on monthly gas consumption for a single-family home in Minnesota shown in the left panel Figure 2.7. As the temperature rises, the residents of the house use less gas for heating. But this trend is not well described using a straight line fit by least squares, in this case

$$\text{Gas Bill} = \$226 - 3 \cdot \text{Temperature} + \text{Residual}.$$

For example, consumption levels off when the temperature rises above 65 degrees F, but the straight line keeps going down.

The inadequacy of the linear model is revealed by the residual plot in the right panel. Here, the residuals  $e_i$  from the linear fit in

the left panel are plotted versus temperature. Remember, these residuals *should* be unrelated with the predictor if our regression model has done its job right. But here, this is clearly false:

- At very cold temperatures (10-20 degrees), the residuals are almost all positive, suggesting that the regression model made predictions that were systematically too low.
- At cool temperatures (40-60 degrees), the residuals are almost all negative, suggesting that the regression model made predictions that were systematically too high.
- At nice temperatures (65-80 degrees), the residuals are almost all positive, suggesting that the regression model made predictions that were systematically too low yet again.

Thus there is still information left in the temperature variable that can be exploited to do a better job at predicting the gas bill.

In such cases, we need to consider nonlinear regression models. In this section, we'll look at two restricted—but still very useful—families of nonlinear models that can still be fit easily using least squares:

- (1) polynomial models (like quadratic or cubic equations); and
- (2) models involving a simple mathematical transformation of the predictor, the response, or both.

### *Polynomial regression models*

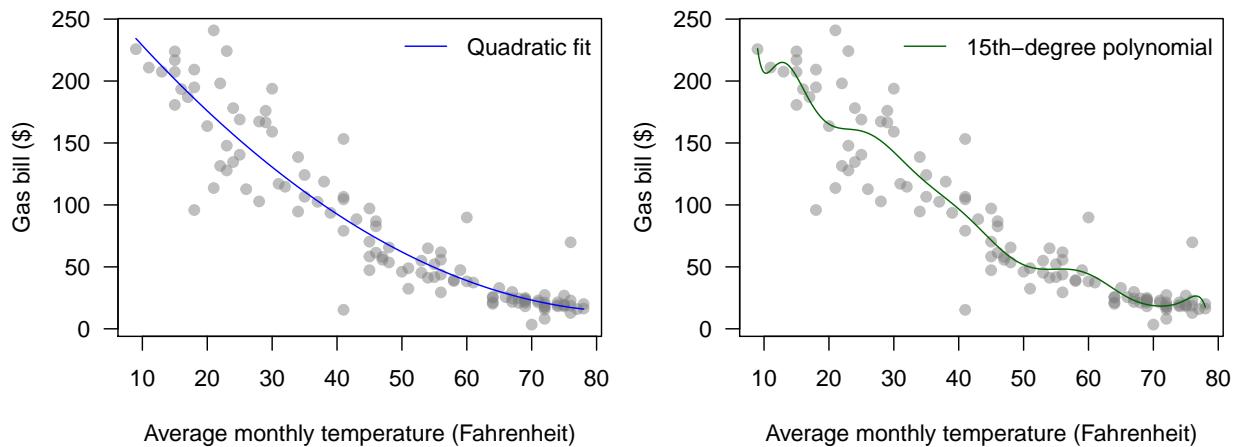
A polynomial is a mathematical function defined by sum of multiple terms, each containing a different power of the same variable (here, as elsewhere, denoted  $x$ ). A linear function is a special case of a polynomial that only has the first power of  $x$ :  $y = \beta_0 + \beta_1 x$ .

But we can fit other polynomials by least squares, too. For example, the left panel of Figure 2.8 shows the least-squares fit of a quadratic equation (another name for a second-degree polynomial) to the gas-consumption data set:

$$\text{Gas Bill} = \$289 - 6.4 \cdot \text{Temp} + 0.03 \cdot \text{Temp}^2 + \text{Residual}.$$

The quadratic model fits noticeably better than the straight line. In particular, it captures the leveling-off in gas consumption at high temperatures that was missed by the linear model.

Beyond these two families, there is a much wider class of nonlinear models that can still be fit by least squares, but not easily. (That is, Legendre's simple computational method won't work, and we need something fancier.) These are often called nonparametric regression models, and they are the subject of a more advanced course.



*Over-fitting.* If the quadratic model (a second-order polynomial) fits better than the straight line (a first-order polynomial), why not try a third-, fourth-, or higher-order polynomial to get an even better fit? After all, we can fit polynomial models of any degree by least squares, estimating equations of the form

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_K x^K,$$

for an arbitrary choice of  $K$  (the degree of the polynomial).

To want to fit the data as well as possible is an understandable impulse. But for most data sets, if we venture beyond  $K = 2$  (quadratic) or  $K = 3$  (cubic), we rapidly get into dangerous overfitting territory. *Over-fitting* occurs when a regression model starts to memorize the random noise in the data set, rather than describe the underlying relationship between predictor and response. We see a clear example of overfitting in the right panel of 2.8, which shows the result of using least-squares to estimate a 15th-degree polynomial for gas bill versus temperature. The fitted curve exaggerates minor dips and rises in the data, leading to an absurdly complex function. There's no reason for us to think that gas consumption responds to temperature in the way implied by the green curve on the right of Figure 2.8. For example, why would consumption rise systematically between 10 and 15 degrees, but then drop again between 15 and 20 degrees?

In regression modeling, we want to build models that are only as complex as they must be in order to describe the underlying relationship between predictor and response. But how do we

Figure 2.8: Left: the fit of a quadratic model (2nd-degree polynomial) estimated by least squares. Right: the fit of a 15th-degree polynomial. The model on the left provides an intuitively reasonable description of the underlying relationship, while the model on the right is a clear example of over-fitting.

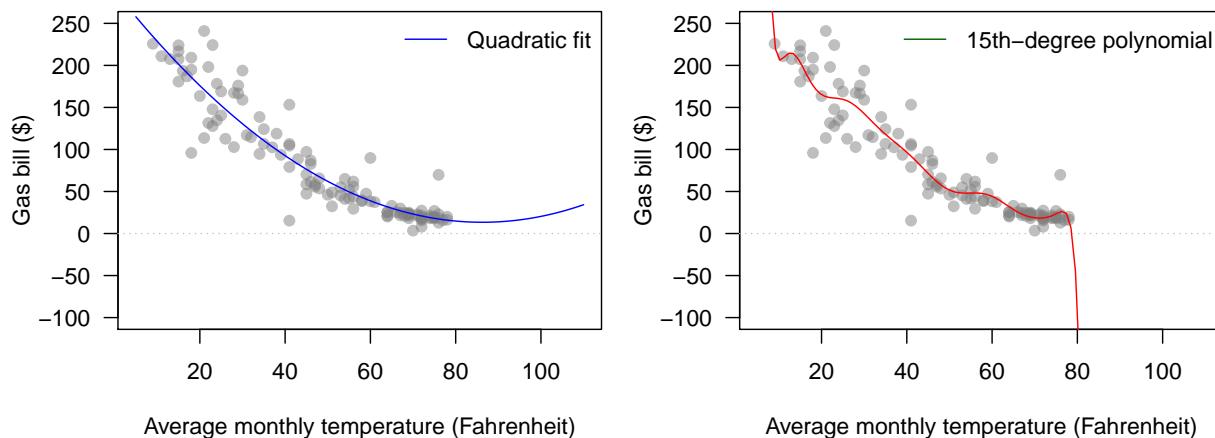


Figure 2.9: Extrapolation in polynomial regression models.

know reliably when we've crossed this line from "fitting well" into over-fitting?

We're still a few chapters away from being able to provide a solid answer to that question. For now, it's fine to let your intuition and your eyes be your guide:

- Does the fitted equation look implausibly wiggly?
- Is there a sound reason, grounded in knowledge of the phenomenon being measured, to believe in the complexity that your model postulates?

With apologies to [Potter Stewart](#): when it comes to overfitting, you'll often know it when you see it. This is one of many reasons why it is always a good idea to plot your data.

*Extrapolation.* Although the quadratic model fits the data well, its predictive abilities will deteriorate as we move above 80 degrees (i.e. as we use the model to extrapolate further and further beyond the range of past experience). As we can see in the left panel of Figure 2.9, that's because the fitted curve is a parabola: it turns upwards around 85 degrees, counterintuitively suggesting that gas bills would eventually rise with temperature.

This behavior is magnified dramatically with higher-order polynomials, which can behave in unpredictable ways beyond the endpoints of your data. The right panel of Figure 2.9 shows this clearly: notice that the predictions of the 15th-degree polynomial drop off a cliff almost immediately beyond the range of the available data, at 79 degrees. You'll sometimes hear this phenomenon—

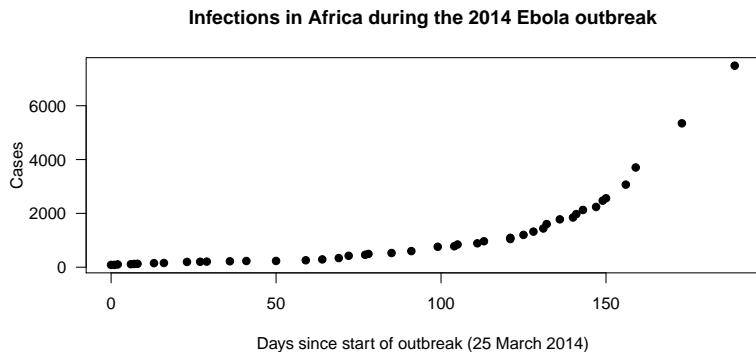


Figure 2.10: Cases of Ebola over time in West Africa, 2014. Compiled from [CDC reports](#) by Francis Smart, as described here.

ridiculous predictions beyond the endpoints of the data—referred to as an “endpoint artifact.”

This example offers a cautionary tale: never extrapolate with a polynomial regression model, unless you really know what you’re doing.

### *Exponential growth and decay*

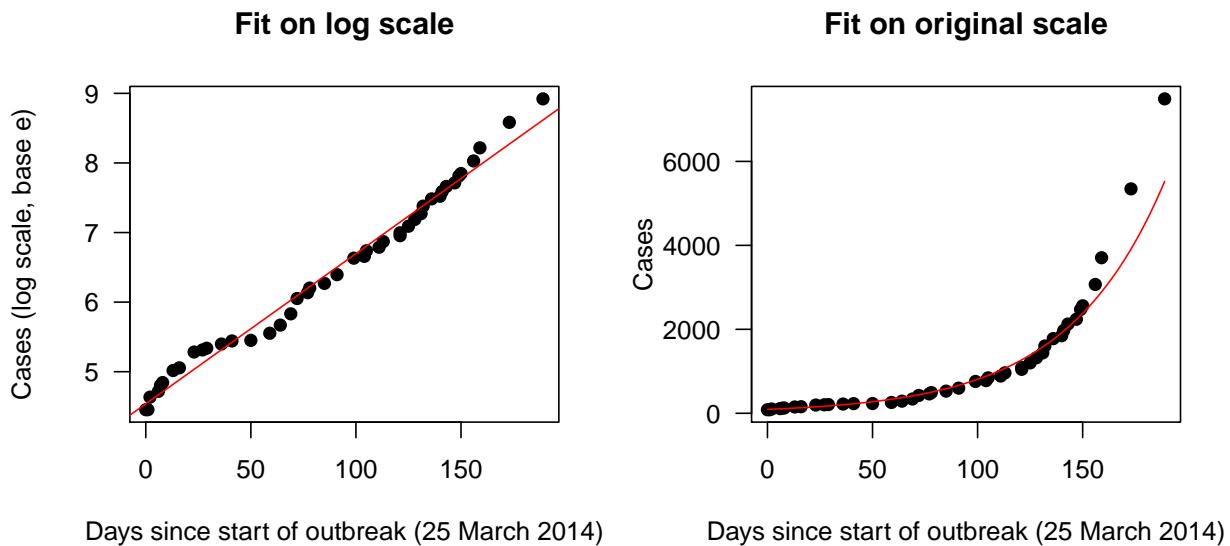
Beginning in March 2014, West Africa experienced the largest outbreak of the Ebola virus in history. Guinea, Liberia, Niger, Sierra Leone, and Senegal were all hit hard by the epidemic. Figure 2.10 shows the number of laboratory-confirmed cases of Ebola in these five countries over time, beginning on March 25.

If we wanted to fit a model to describe how the number of Ebola infections grew over time, we might be tempted to fit a polynomial function (since a linear model clearly won’t work well here). However, basic biology tells us that the transmission rate of a disease through a population is reasonably well described by an exponential growth model: 1 infection leads to 2, which lead to 4, which lead to 8, to 16, and so on. The equation for an exponential-growth model is

$$y = \alpha \cdot e^{\beta t}, \quad (2.6)$$

where  $y$  is the expected number of cases and  $t$  is the number of time intervals (e.g. weeks or days) since the start of the outbreak.

It turns out that we can use least squares to fit an exponential growth model of this form, using a new trick: *take the logarithm of the response variable* and fit a linear model to this new transformed variable. We can see why this works if we take the logarithm of



$y$  in the equation for exponential growth (labeled 2.6, above). To preserve equality, if we take the log of the left-hand side, we also have to take the log of the right-hand side:

$$\begin{aligned}\log y &= \log(\alpha \cdot e^{\beta_1 t}) \\ &= \log \alpha + \beta_1 t.\end{aligned}$$

The second equation says that the log of  $y$  is a linear function of the time variable,  $t$ , with intercept  $\beta_0 = \log \alpha$  and slope  $\beta_1$ .

Thus to fit the exponential growth model for any response variable  $y$ , we need to follow two steps:

- (1) Define a new variable  $z = \log y$  by taking the logarithm of the original response variable.
- (2) Fit a linear model for the transformed variable  $z$  versus the original predictor, using ordinary least squares.

Figure 2.11 shows the result of following these two steps for the Ebola data. The left panel shows the straight-line fit on the log scale:

$$\log \text{Cases} = 4.54 + 0.021 \cdot \text{Days}.$$

The right panel shows the corresponding exponential-growth curve on the original scale:

$$\text{Cases} = 93.5 \cdot e^{0.021 \cdot \text{Days}}.$$

Figure 2.11: An exponential-growth model fit to the Ebola data by ordinary least squares, where the  $y$  variable is shown on the log scale (left) and on the original scale (right).

The leading constant is calculated from the intercept on the log scale:  $93.5 \approx e^{4.54}$ . From Figure 2.11, we can see that the exponential-growth model fits adequately, although imperfectly: the rate of growth seems to be accelerating at the right of the picture, and the upward trajectory is visibly nonlinear on the log scale. (Remember: all models are wrong, but some models are useful.)

An exponential model with a negative slope  $\beta_1$  on the log scale is called an exponential decay model. Exponential decay is a good model for, among other things, the decay of a radioactive isotope.

*Interpreting the coefficient in an exponential model.* To interpret the coefficient in an exponential growth model, we will use it to calculate the doubling time—that is, how many time steps it takes for the response variable (here, Ebola cases) to double.

In terms of our estimated model, the number of cases doubles between days  $t_1$  and  $t_2$  whenever

$$\frac{\alpha e^{\beta_1 t_2}}{\alpha e^{\beta_1 t_1}} = 2,$$

so that the number of cases on day  $t_2$  (in the numerator) is precisely twice the number of cases on day  $t_1$ , in the denominator. If we simplify this equation using the basic [rules of algebra for exponentials](#), we find that the number of days that have elapsed between  $t_1$  and  $t_2$  is

$$t_2 - t_1 = \frac{\log 2}{\beta_1}.$$

This is our doubling time. For Ebola in West Africa, the number of cases doubled roughly every

$$\frac{\log 2}{0.021} \approx 32$$

days during the spring and early summer of 2014.

In an exponential decay model (where  $\beta_1 < 0$ ), a similar calculation would tell you the [half life](#), not the doubling time.<sup>6</sup>

### Double log transformations

In some cases, it may be best to take the log of both the predictor and the response, and to work on this doubly transformed scale. For example, in the upper left panel of Figure 2.12, we see a scatter plot of brain weight (in grams) versus body weight (in kilos) for 62 different mammalian species, ranging from the lesser short-tailed shrew (weight: 10 grams) to the African elephant (weight:

<sup>6</sup> Instead, solve the equation

$$\frac{\alpha e^{\beta_1 t_2}}{\alpha e^{\beta_1 t_1}} = 1/2$$

for the difference  $t_2 - t_1$ .

6000+ kilos). You can see that most species are scrunched up in a small box at the lower left of the plot. This happens because the observations span many orders of magnitude, and most are small in absolute terms.

But if we take the log of both body weight and brain weight, as in the top-right panel of Figure 2.12, the picture changes considerably. Notice that, in each of the top two panels, the red box encloses the same set of points. On the right, however, the double log transformation has stretched the box out in both dimensions, allowing us to see the large number of data points that, on the left, were all trying to occupy the same space. Meanwhile, the two points outside the box (the African and Asian elephants) have been forced to cede some real estate to the rest of Mammalia.

This emphasizes the taking the log is an “unsquishing” operator. To see this explicitly, look at the histograms in the second and third row of panels in Figure 2.12. Whenever the histogram of a variable looks highly skewed right, as on the left, a log transformation is worth considering. It will yield a much more nicely spread-out distribution of points, as on the right.

*Power laws.* It turns out that when we take the log of both variables, we are actually fitting a *power law* for the relationship between  $y$  and  $x$ . The equation of a power law is

$$y = \alpha \cdot x^{\beta_1}$$

for some choices of  $\alpha$  and  $\beta$ . This is a very common model for data sets that span many orders of magnitude (like the body/brain weight data). To see the connection with the double log transformation, simply take the logarithm of both sides of the power law:

$$\begin{aligned} \log y &= \log (\alpha \cdot x^{\beta_1}) \\ &= \log \alpha + \log x^{\beta_1} \\ &= \log \alpha + \beta_1 \log x. \end{aligned}$$

Therefore, if  $y$  and  $x$  follow a power law, then  $\log y$  and  $\log x$  follow a linear relationship with intercept  $\log \alpha$  and slope  $\beta_1$ . This implies that we can fit the parameters of the power law by applying the double log transformation and using ordinary least squares. For our mammalian brain weight data, applying this recipe yields the fitted equation

$$\log \text{brain} = 2.13 + 0.75 \cdot \log \text{body},$$

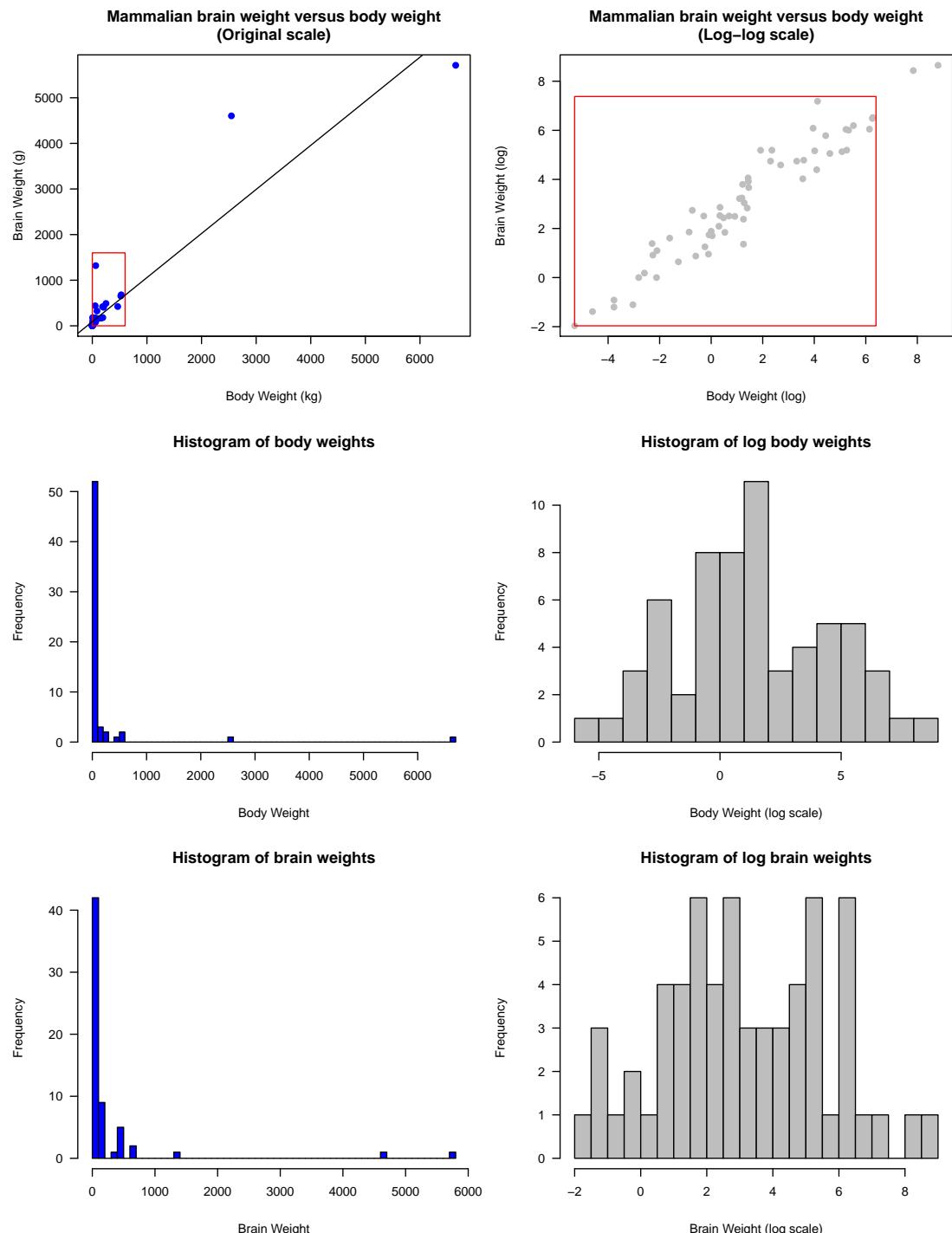
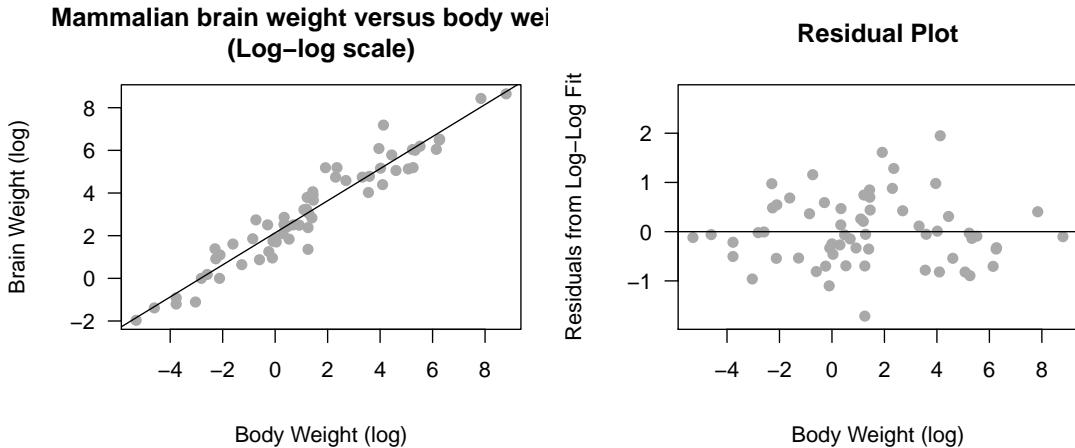


Figure 2.12: Brain weight versus body weight for 62 mammalian species, both on the original scale and the log scale. Notice how the log transformation “unsquishes” the points.



or expressed as a power law on the original scale,

$$\text{brain} = 8.4 \cdot \text{body}^{0.75}.$$

*The residuals in a power-law model.* As we've just seen, we can fit power laws using ordinary least squares after a log transformation of both the predictor and response. In introducing this idea, we ignored the residuals and focused only on the part of the model that describes the systematic relationship between  $y$  and  $x$ . If we keep track of these residuals a bit more carefully, we see that the model we're fitting for the  $i$ th response variable is this:

$$\log y_i = \log \alpha + \beta_1 \log x_i + e_i, \quad (2.7)$$

where  $e_i$  is the amount by which the fitted line misses  $\log y_i$ . We suppressed these residuals before the lighten the algebra, but now we'll pay them a bit more attention.

Equation 2.7 says that the residuals affect the model in an additive way on the log scale. But if we exponentiate both sides, we find that they affect the model in a multiplicative way on the original scale:

$$\begin{aligned} \exp(\log y_i) &= \exp(\log \alpha) \cdot \exp(\beta_1 \log x) \exp(e_i) \\ y_i &= \alpha x^{\beta_1} \exp(e_i). \end{aligned}$$

Therefore, in a power low, the exponentiated residuals describe the percentage error made by the model on the original scale. Let's work through the calculations for two examples:

Figure 2.13: A straight-line fit to the mammalian brain weight data after a double log transformation.

- If  $e_i = 0.2$  on the log–log scale, then the actual response is  $\exp(0.2) \approx 1.22$  times the value predicted by the model. That is, our model underestimates this particular  $y_i$  by 22%.
- If  $e_i = -0.1$  on the log–log scale, then the actual response is  $\exp(-0.1) \approx 0.9$  times the value predicted by the model. That is, our model overestimates this particular  $y_i$  by 10%.

The key thing to realize here is that the *absolute* magnitude of the error will therefore depend on whether the  $y$  variable itself is large or small. This kind of multiplicative error structure makes perfect sense for our body–brain weight data: a 10% error for a lesser short-tailed shrew will have us off by a gram or two, while a 10% error for an elephant will have us off by 60 kilos or more. Bigger critters mean bigger errors—but only in an absolute sense, and not if we measure error relative to body weight.

*Interpreting the slope under a double log transformation.* To correctly interpret the slope  $\beta_1$  under a double log transformation, we need a little bit of calculus. The power law that we want to fit is of the form  $y = \alpha x^{\beta_1}$ . If we take the derivative of this expression, we get

$$\frac{dy}{dx} = \beta_1 \alpha x^{\beta_1 - 1}.$$

We can rewrite this as

$$\begin{aligned}\frac{dy}{dx} &= \frac{\beta_1 \alpha x^{\beta_1}}{x} \\ &= \beta_1 \frac{y}{x}.\end{aligned}$$

If we solve this expression for  $\beta_1$ , we get

$$\beta_1 = \frac{dy/y}{dx/x}. \quad (2.8)$$

Since the  $dy$  in the derivative means “change in  $y$ ”, the numerator is the rate at which the  $y$  variable changes, as a fraction of its value. Similarly, since  $dx$  means “change in  $x$ ”, the denominator is the rate at which the  $x$  variable changes, as a fraction of its value.

Putting this all together, we find that  $\beta_1$  measures the ratio of percentage change in  $y$  to percentage change in  $x$ . In our the mammalian brain-weight data, the least-squares estimate of the slope on a log-log scale was  $\hat{\beta}_1 = 0.75$ . This means that, among mammals, a 100% change (i.e. a doubling) in body weight is associated with a 75% expected change in brain weight. The bigger you are, it

would seem, the smaller your brain gets—at least relatively speaking.

The coefficient  $\beta_1$  in a power law is often called an *elasticity* parameter, especially in economics, where it is used to quantify the responsiveness of consumer demand to changes in the price of a good or service. The underlying model for consumer behavior that's often postulated is that

$$Q = \alpha P^{\beta_1},$$

where  $Q$  is the quantity demanded by consumers,  $P$  is the price, and  $\beta_1 < 0$ . Economists would call  $\beta_1$  the **price elasticity of demand**,<sup>7</sup> which may be a familiar concept from a microeconomics course.

<sup>7</sup> They actually define elasticity as the ratio in Equation 2.8, but as we've seen, this is mathematically equivalent to the regression coefficient you get when you fit the  $x$ - $y$  relationship using a power law.



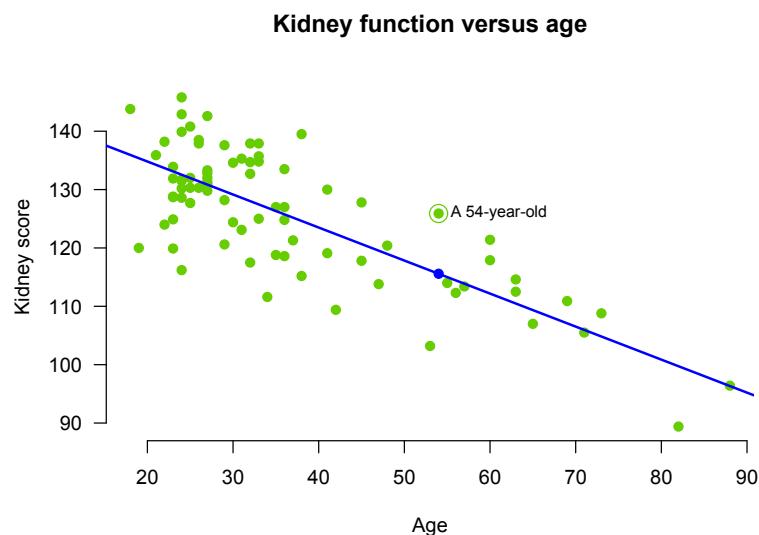
# 3

## *Predictable and unpredictable variation*

### Quantifying uncertainty in a prediction

THERE are many things we can look forward to as we age—for example, richer relationships, improved confidence, better self-knowledge, and the right to go to bed early without being judged.

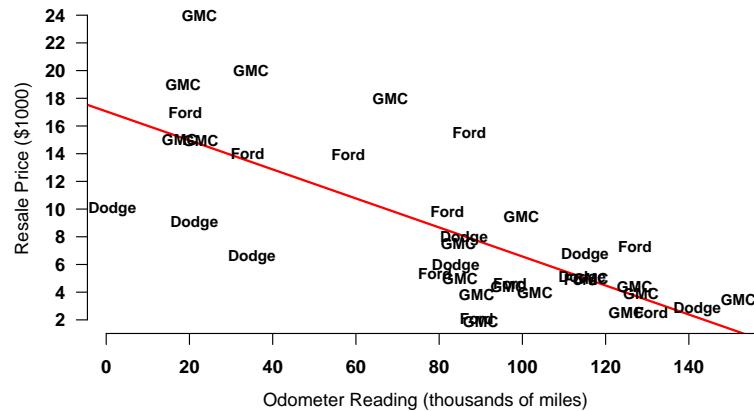
Unfortunately, improved kidney function isn't one of them. The following plot shows a sample of 78 patients from an ordinary doctor's office. The  $x$ -axis shows the patient's age, while the  $y$ -axis shows the patient's creatinine-clearance rate in mL/min, which is a common measure of kidney function (higher is better):<sup>1</sup>



<sup>1</sup> According to the National Institutes of Health, “The creatinine clearance test helps provide information about how well the kidneys are working. The test compares the creatinine level in urine with the creatinine level in blood. . . . Creatinine is removed, or cleared, from the body entirely by the kidneys. If kidney function is abnormal, creatinine level increases in the blood because less creatinine is released through the urine.”

Suppose you're the doctor running this clinic, and a 54-year old man walks through the door. He tests at 126 mL/min, which is 10 points above the prediction of the regression line (blue dot on the line). Is the man's score too high, or is it within the range of normal variation from the line?

This question is fundamentally about *prediction uncertainty*. Anytime we use a statistical model to make a prediction, some version of this question comes up. For example, among pickup trucks for sale on Craigslist, those with higher odometer readings tend to have lower asking prices:



Now imagine you have your eye on a pickup truck with 80,000 miles on it. The least squares fit says such that the expected price for such a truck is about \$8,700. If the owner is asking \$11,000, is this reasonable, or drastically out of line with the market?

Here's another example. Mammals more keenly in danger of predation tend to dream fewer hours.

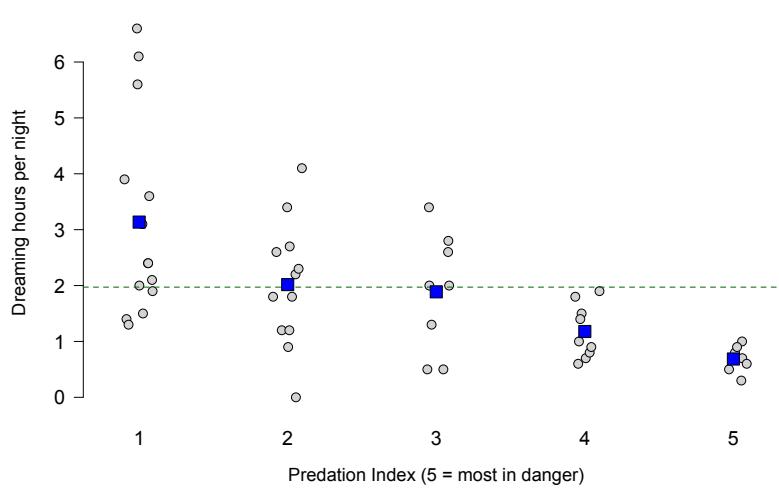


Figure 3.1: Dreaming hours per night versus danger of predation for 50 mammalian species. In this and in Figure 3.2, the blue squares show the group-wise means, while the dotted green line shows the grand mean for the entire data set.

But there is still residual variation that practically begs for a Zen

proverb. Why does the water rat dream at length? Why does the wolverine not?

Finally, the people of Raleigh, NC tend to use less electricity in the milder months of autumn and spring than in the height of winter or summer—but not uniformly. Many spring days see more power usage than average; many summer days see less. What is the normal range of electricity consumption for a day in August, the hottest month of the year?

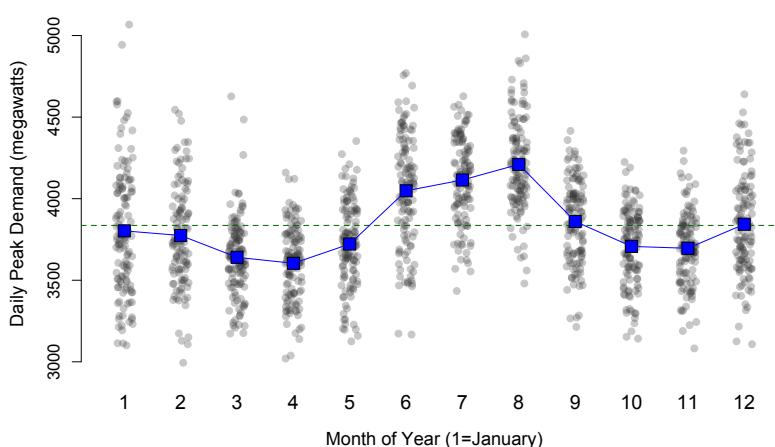


Figure 3.2: Daily peak demand for electricity versus month of the year in Raleigh, NC from 2006–2009.

In all of these cases, one must remember that the fitted values from a statistical model are generalizations about a typical case, given the information in the predictor. But no generalization holds for all cases. This is why we explicitly write models as

$$\text{Observed } y \text{ value} = \text{Fitted value} + \text{Residual}.$$

It is common to view a statistical model as nothing more than a recipe for calculating the fitted values, and to think that the residuals are just the errors made by this model. But we'll have a richer picture if instead we view the residuals as *part of* the model. If you've ignored the variation in the residuals, then you really haven't specified a complete forecast.

An important distinction here is that of a *point estimate*, or single best guess, versus an *interval estimate*, or a range of likely values. Fitted values are point estimates. Point estimates are useful, but interval estimates are much better. After all, variation from the average, far from being an “error,” is a normal part of life.

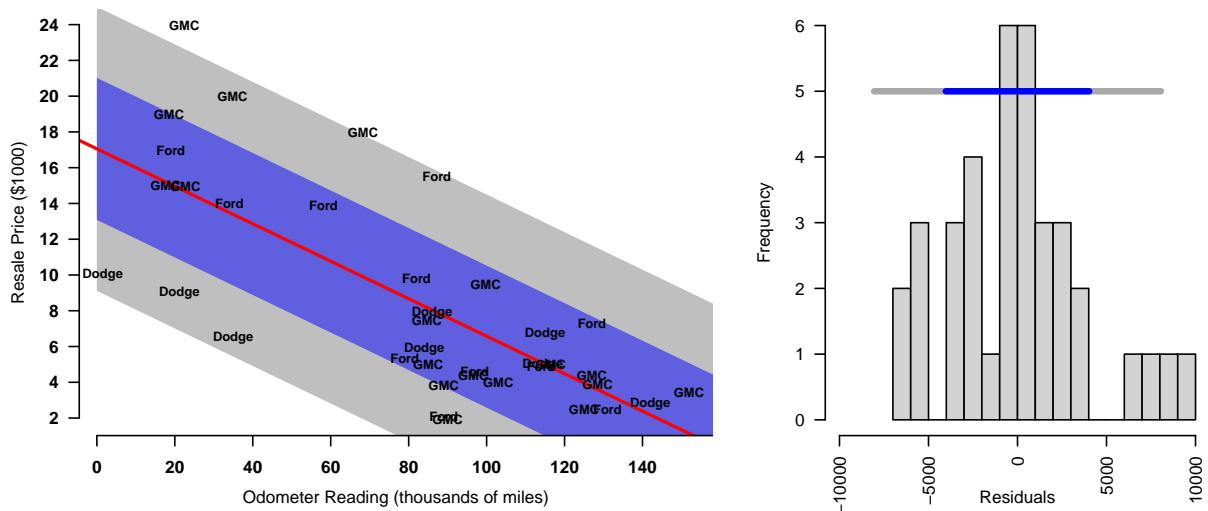


Figure 3.3: Prediction intervals for pickup-truck prices on Craigslist. Shaded grey area = 1 residual standard deviation to either side of the fitted line (red); shaded blue = 2 standard deviations. The right panel shows a histogram of the residuals, along with the width of the  $1s_e$  (blue) and  $2s_e$  (grey) envelopes.

## Prediction intervals

THE key question we must answer to quantify our prediction uncertainty is: “How much does a typical case vary from the prediction of the regression model?” We have a lot of ways to answer this question (box plots, histograms, dot plots, and so forth). The most common way is to calculate the *residual standard deviation*:

$$s_e = \sqrt{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

where  $p$  is the number of free parameters in the model (e.g. two for a straight-line fit: the slope and the intercept). This quantity describes how much a typical case deviates from the fitted line, just like the ordinary standard deviation tells us how much a typical case deviates from sample mean  $\bar{y}$  (page 15).

To see how the residual standard deviation  $s_e$  can be used to quantify prediction uncertainty, let’s take another look at the data set of pickup trucks advertised on Craigslist. In Figure 3.3, the red line is the least-squares fit:  $Y = 17054 - 0.105 x$ . The residual standard deviation is \$3,971, compared to the original standard deviation of \$5,584. That is, a typical truck deviates from the sample mean  $\bar{y}$  by about \$5,584, and from the least-squares line by about \$3,971. Knowledge of the truck’s mileage has improved our

The residual standard deviation is also called the *residual standard error*. The formula for the residual standard deviation is *almost* identical to the formula for the sample standard deviation of the residuals. The minor difference is the divisor:  $n - p$  instead of  $n - 1$ . The reason is that the sample standard deviation centers  $y_i$  by the sample mean, which involves computing 1 extra number ( $\bar{y}$ ) from the data. The residual standard deviation centers  $y_i$  by the OLS fitted values, which involves computing  $p$  extra numbers ( $\hat{\beta}_0$  and  $\hat{\beta}_1$  in the case of a straight-line fit).

predictive accuracy by about  $\$5584 - \$3971 = \$1613$ , but there is still a lot of uncertainty. The two shaded strips in Figure 3.3 depict this uncertainty visually. The blue extends to 1 residual standard deviation (line  $\pm \$3,971$ ) on either side of the line, while the grey strip extends to 2 residual standard deviations (line  $\pm \$7,942$ ).

The key idea of a prediction interval is that these grey strips can be used to provide an interval estimate for forecasting the price of a future truck—that is, one not in our original data set. For our hypothetical pickup truck with 80,000 miles, the point estimate for the expected price (from the least-squares line) is \$8,672. But if we go out one residual standard deviation, the *interval estimate* is  $\$8,672 \pm \$3,971$ , or (4701, 12643). You can see where the wide of these one- and two-standard-deviation envelopes comes from, in the histogram in the right panel of Figure 3.3.

How accurate is the interval estimate? A simple way to quantify this is just to count the number of cases that fall within the one-standard deviation band to either side of the line, as a fraction of the total number of cases. Since the medium grey strip,

$$y \in 17054 - 0.105 \cdot x \pm 3971,$$

captures 27 out of 37 total cases, it therefore constitutes a family of *prediction intervals* at a *coverage level* of 73% (27/37). We call it a family of intervals, because there is actually one such prediction interval for every possible value of  $x$ . At  $x = 80000$ , the interval is (4701, 12643); at  $x = 40,000$ , the interval is (8892, 16834).

To summarize, forming a prediction interval requires two steps: constructing the interval, and quantifying its accuracy. In a simple linear regression model, the interval itself takes the form

$$y \in \hat{\beta}_0 + \hat{\beta}_1 x \pm k \cdot s_e,$$

or more concisely,  $y \in \hat{y} \pm k \cdot s_e$ . Here  $s_e$  is the residual standard deviation, and  $k$  is a chosen multiple that characterizes the width of the intervals. There is a clear trade-off here: larger choices of  $k$  mean wider intervals, which mean more uncertainty, but greater coverage. Typical values for  $k$  are 1 or 2. To quantify the accuracy of the interval, we look at its coverage: that is, what fraction of examples in our original data set are contained within their corresponding interval.

Most good statistical software makes it easy to calculate prediction intervals. In R, for example, the `predict` function allows you to specify a given coverage level (e.g. 95%) and will output the

Here the notation  $y \in c \pm h$  means that  $y$  (the response) is in the interval centered at  $c$  that extends  $h$  units to either side. Thus  $h$  is the half-width of the interval. The sign  $\in$  is concise mathematical notation for “is in.”

lower and upper bounds of a prediction interval at that specified coverage.

*Standardized residuals.* Return to the question we posed on the beginning of the chapter. You're the doctor at a clinic, and a 54-year old man has score of 126 mL/min for his creatinine clearance test. Is the man's score too high, or is it within the range of normal variation from the line?

We can answer this question by calculating a *standardized residual*, which is just a  $z$ -score based on dividing the residual by the residual standard deviation:

$$z = \frac{y_i - \hat{y}_i}{s_e} = \frac{e_i}{s_e}.$$

In this example,  $y_i = 126$ ,  $\hat{y}_i = 116$ , and the residual standard deviation is  $s_e = 7.2$ . Therefore the man's  $z$ -score is  $(126 - 116)/7.2$ , or about 1.4 standard deviations above normal. This is on the higher side, but within the range of typical values seen in the clinic.

*A caveat.* The technique we've learned for forming prediction intervals is pretty useful, but it's not perfect. That's because it ignores uncertainty about the parameters of the model itself, and only accounts for uncertainty about residuals, assuming that the fitted model is true. (That is, we're ignoring the fact that we might have been a bit off in our estimates of the slope and intercept, due to sampling variability.) As a result, these prediction intervals actually underestimate the total amount of uncertainty that we'd like to incorporate into our interval estimate. We'll soon learn how to quantify these additional forms of uncertainty. But imperfections aside, even these slightly naïve prediction intervals that don't account for parameter uncertainty are much better than a point estimate.

## Partitioning sums of squares

WHEN we introduced the concept of the sample standard deviation, we asked the question: what's so great about sums of squares for measuring variation? The answer is: because linear statistical models *partition the total sum of squares* into predictable and unpredictable components. This isn't true of any other simple measure of variation. Sums of squares are special.

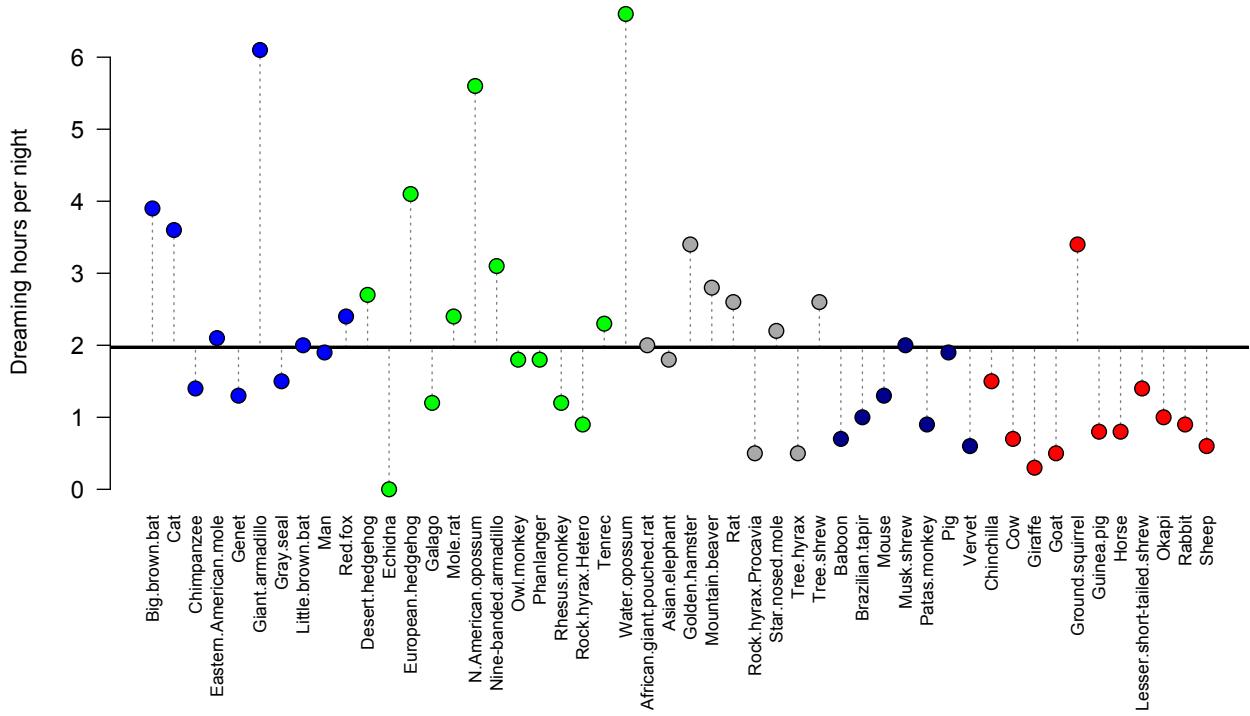


Figure 3.4: Dreaming hours by species, along with the grand mean. For reference, the colors denote the predation index, ordered from left to right in increasing order of danger (1–5). The vertical dotted lines show the deviations from the grand mean:  $y_i - \bar{y}$ .

Let's return to those grand and group means for the mammalian sleeping-pattern data. We will use sums of squares to measure three quantities: the total variation in dreaming hours; the variation that can be predicted using the predation index; and unpredictable variation that remains "in the wild."

In Figure 3.4, we see the observed  $y$  value (dreaming hours per night) plotted for every species in the data set. The horizontal black line shows the grand mean,  $\bar{y} = 1.97$  hours. The dotted vertical lines show the deviations between the grand mean and the actual  $y$  values,  $y_i - \bar{y}$ .

To account for the information in the predictor, we fit the model "dreaming hours  $\sim$  predation index," computing a different mean for each group:

$$\underbrace{y_i}_{\text{Observed value}} = \underbrace{\hat{y}_i}_{\text{Group mean}} + \underbrace{e_i}_{\text{Residual}}.$$

There are three quantities to keep track of here:

- The observed values,  $y_i$ .

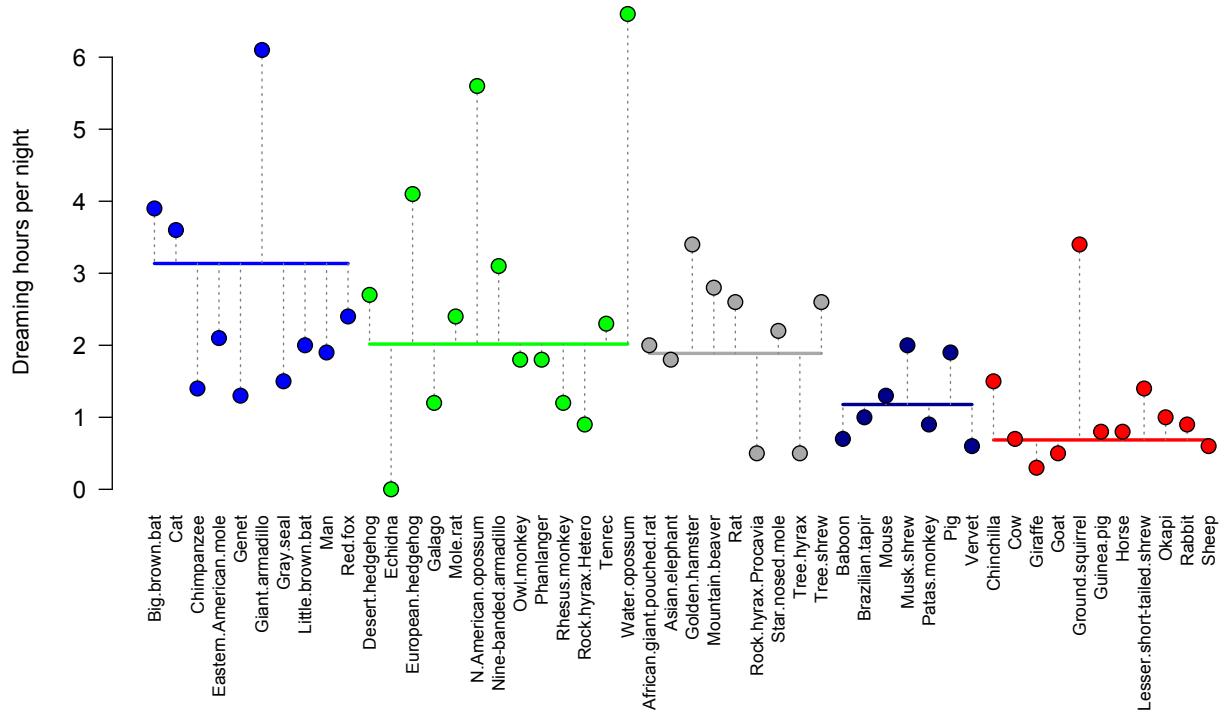


Figure 3.5: Dreaming hours by species, along with the group means stratified by predation index. The vertical dotted lines show the residuals from the group-wise model “Dreaming hours  $\sim$  predation index.”

- The grand mean,  $\bar{y}$ .
- The fitted values,  $\hat{y}_i$ , which are just the group means corresponding to each observation. These are shown by the colored horizontal lines in Figure 3.5 and again as diamonds in Figure 3.6. For example, cats and foxes in group 1 (least danger, at the left in dark blue) both have fitted values of 3.14; goats and ground squirrels in group 5 (most danger, at the right in bright red) both have fitted values of 0.68. Notice that the fitted values also have a sample mean of  $\bar{y}$ : the average fitted value is the average observation.

There are also three important relationships among  $y_i$ ,  $\hat{y}_i$ , and  $\bar{y}$  to keep track of. We said we’d measure variation using sums of squares, so let’s plunge ahead.

- The total variation, or the sum of squared deviations from the mean  $\bar{y}$ . This measures the variability in the original data:

$$TV = \sum_{i=1}^n (y_i - \bar{y})^2 = 102.1.$$

This equation says that the number 102.1 comes from summing all the squared deviations in the data set—that is,  $(3.9 - \bar{y})^2 + (3.6 - \bar{y})^2 + \dots + (0.6 - \bar{y})^2 = 102.1$ .

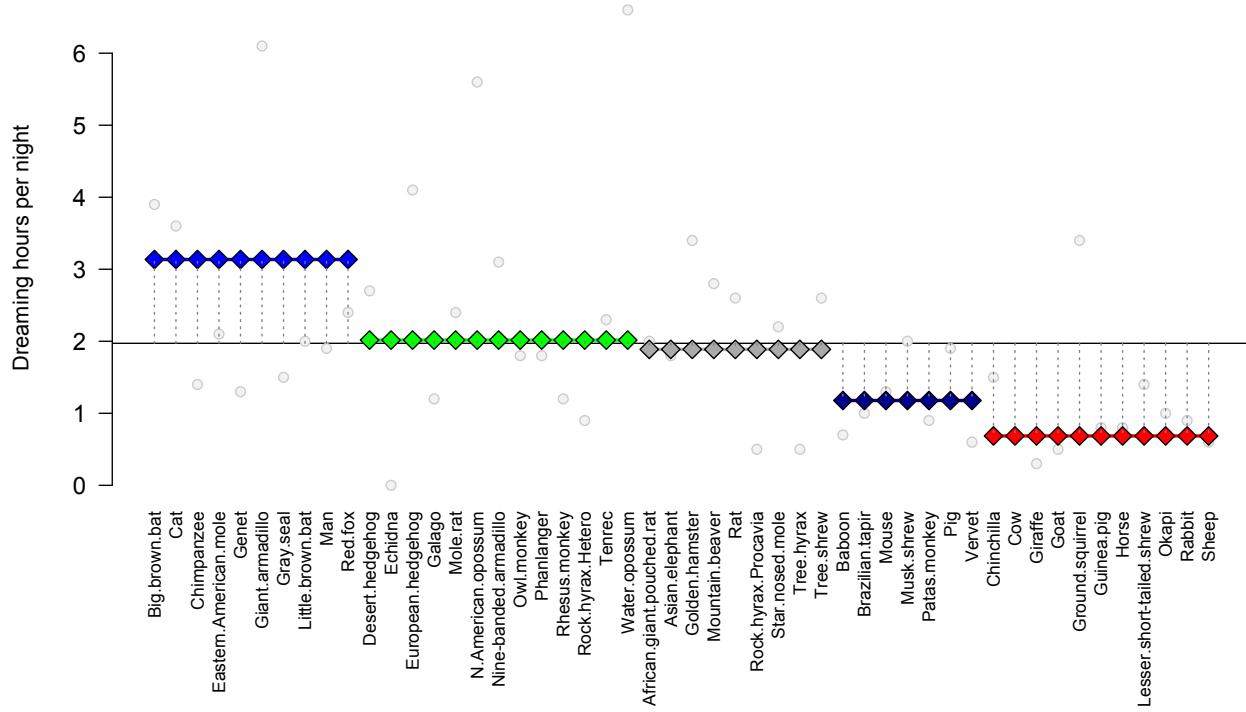


Figure 3.6: Dreaming hours by species (in grey), along with the fitted values (colored diamonds) from the group-wise model using predation index as a predictor. The vertical lines depict the differences  $\hat{y}_i - \bar{y}$ .

- The predictable variation, or the sum squared differences between the fitted values and the grand mean. This measures the variability described by the model:

$$PV = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 36.4.$$

- The unpredictable variation, or the sum of squared residuals from the group-wise model. This is the variation left over in the observed values after accounting for group membership:

$$UV = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = 65.7.$$

What's special about these numbers? Well, notice that

$$102.1 = 36.4 + 65.7,$$

so that  $TV = PV + UV$ . The model has cleanly partitioned the original sum of squares in two components: one predicted by the model, and one not.

What if we measured variation using sums of absolute values instead? Let's try it and see:

$$\sum_{i=1}^n |y_i - \bar{y}| = 53.0$$

$$\sum_{i=1}^n |\hat{y}_i - \bar{y}| = 33.7$$

$$\sum_{i=1}^n |y_i - \hat{y}_i| = 42.5.$$

Clearly  $53.0 \neq 33.7 + 42.5$ . If this had been how we'd defined TV, PV, and UV, we wouldn't have such a clean "partitioning effect" like the kind we found for sums of squares.

Is this partition effect a coincidence, or a meaningful generalization? To get further insight, let's try the same calculations on the peak-demand data set from Figure 3.2, seen again at right. First, we sum up the squared deviations  $y_i - \bar{y}$  to get the total variation:

$$TV = \sum_{i=1}^n (y_i - \bar{y})^2 = 166,513,967.$$

Next, we sum up the squared deviations of the fitted values. For each observation, the fitted value is just the group-wise mean for the corresponding month, given by the blue dots at right:

$$PV = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 50,262,962.$$

Finally, we sum up the squared residuals from the model:

$$UV = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 116,251,005.$$

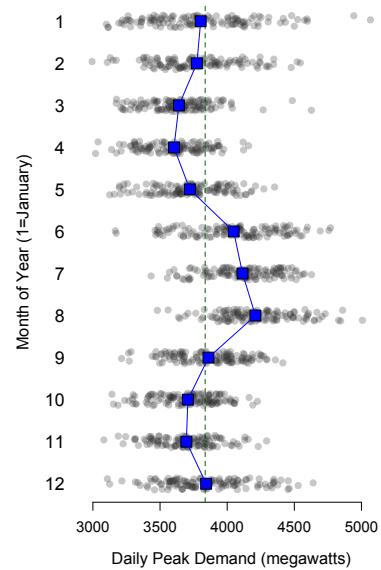
Sure enough:  $166,513,967 = 50,262,962 + 116,251,005$ . The same "TV = PV + UV" statement holds when using sums of squares, just as for the previous data set.

And if we try sums of absolute values?

$$\sum_{i=1}^n |y_i - \bar{y}| = 397,887.7$$

$$\sum_{i=1}^n |\hat{y}_i - \bar{y}| = 220,382.1$$

$$\sum_{i=1}^n |y_i - \hat{y}_i| = 325,409.0.$$



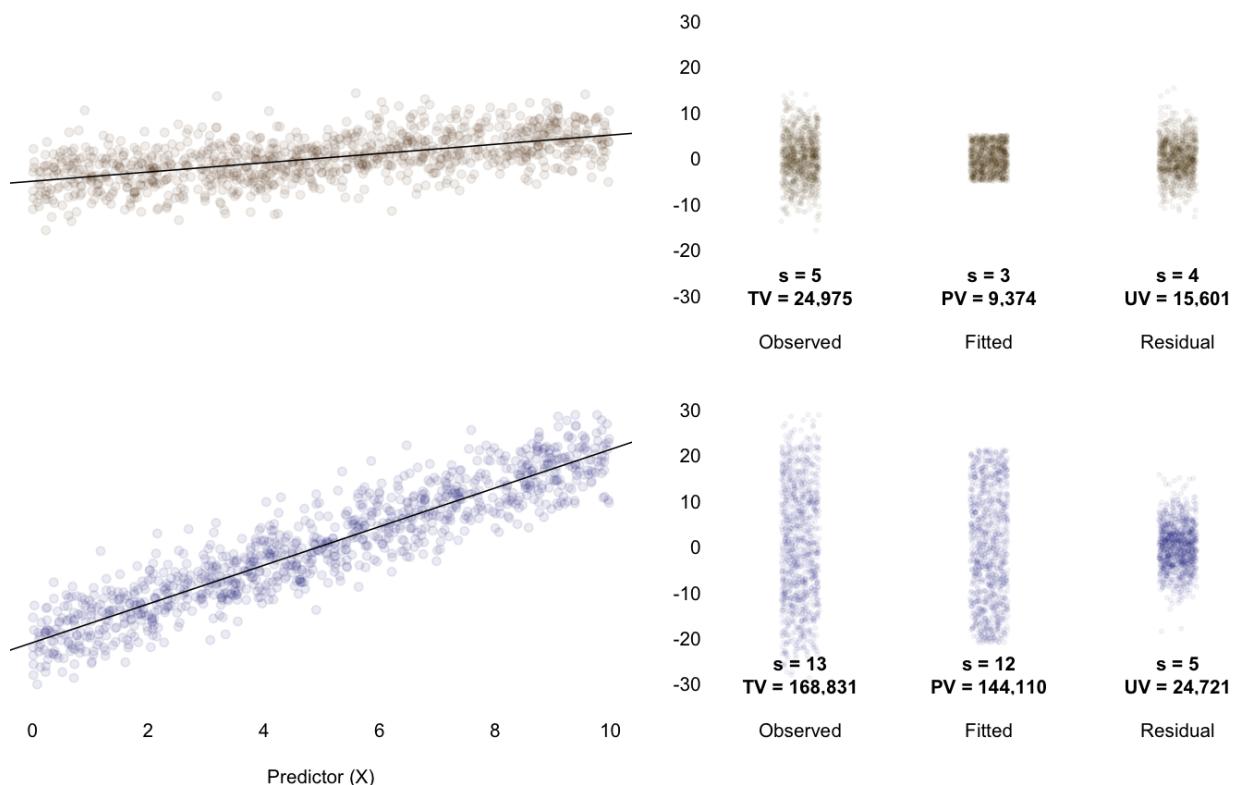


Figure 3.7: Two imaginary data sets, along with their least squares lines.

Clearly,  $397,887.7 \neq 220,382.1 + 325,409.0$ . Just like the mammalian sleep-pattern data, the peak-demand data exhibits no partitioning-of-variation effect using sums of absolute deviations.

The same decomposition also holds for linear regression models. In Figure 3.7 we see two scatter plots of two simulated data sets, both measured on the same  $X$  and  $Y$  scales. Next to each are dot plots of the original  $Y$  variable, the fitted values, and the residuals. In each case,  $TV = PV + UV$ , and therefore the three standard deviations form Pythagorean triples.

## The analysis of variance: a first look

MEASURING variation using sums of squares is not at all an obvious thing to start out doing. But obvious or not, we do it for a very good reason: sums of squares follow the lovely, clean decomposition that we happened upon in the previous section:

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ TV &= PV + UV. \end{aligned} \quad (3.1)$$

This is true both for group-wise models and for linear models. TV and UV tell us much variation we started with, and how much we have left over after fitting the model, respectively. PV tells us where the missing variation went—into the fitted values!

As we've repeatedly mentioned, it would be perfectly sensible to measure variation using sums of absolute values  $|y_i - \hat{y}_i|$  instead, or even something else entirely. But if we were to do this, the analogous "TV = PV + UV" decomposition would not hold as a general rule:

$$\sum_{i=1}^n |y_i - \bar{y}| \neq \sum_{i=1}^n |\hat{y}_i - \bar{y}| + \sum_{i=1}^n |y_i - \hat{y}_i|.$$

In fact, a stronger statement is true: there is literally no power other than 2 that we could have chosen that would have led to a decomposition like Equation 3.1. Sums of squares are special because they, and they alone, can be partitioned cleanly into predictable and unpredictable components.

This partitioning effect is something of a mystery—most things in everyday life simply don't work this way. For example, imagine that you and your sibling are trying to divide up a group of 100

stuffed animals that you own in common. It makes no sense to say: "Well, there are 10,000 ( $100^2$ ) squared-stuffed-animalss in total, so I'll take 3,600 ( $60^2$ ) squared stuffed animals, and you take the remaining 1,600 ( $40^2$ )."<sup>1</sup> Not only is the statement itself barely interpretable—what the heck is a squared stuffed animal?—but the math doesn't even work out ( $100^2 \neq 60^2 + 40^2$ ).

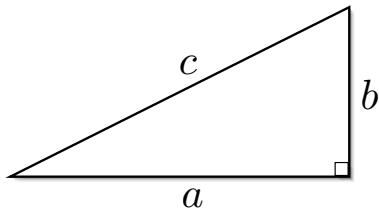
Is there a deeper reason why this partitioning effect occurs for sums of squares in statistical models, and not for some other measure of variation? The figure at right should jog your memory, for this isn't the first time you've seen a similar result before. Pythagoras' famous theorem says that  $c^2 = a^2 + b^2$ , where  $c$  is the hypotenuse of a right triangle, and  $a$  and  $b$  are the legs. Notice that Pythagoras *doesn't* have anything interesting to say about the actual numbers:  $c \neq a + b$ . It's the squares of the numbers that matter.

This way of partitioning a whole into parts makes no sense for DVDs, but it does occur in real life—namely, every time you traverse a city or campus laid out on a grid. In Figure 3.8, for example, you see part of a 1930 map of the University of Texas. Both then and now, any student who wanted to make her way from the University Methodist Church (upper left star) to the football stadium (lower right star) would need to travel about 870 meters as the crow flies. She would probably do so in two stages: first by going 440 meters south on Guadalupe, and then by going 750 meters east on 21st Street.

Notice how the total distance gets partitioned:  $870 \neq 440 + 750$ , but  $870^2 = 440^2 + 750^2$ . North–south and east–west are perpendicular directions, and if you stay along these axes, total distances will add in the Pythagorean way, rather than in the usual way of everyday arithmetic.

So it is with a statistical model. You can think of the fitted values  $\hat{y}_i$  and the residuals  $e_i$  as pointing in two different directions that are, mathematically speaking, perpendicular to one another: one direction that can be predicted by the model, and one direction that can't. The total variation is then like the hypotenuse of the right triangle so formed:

This business of partitioning sums of squares into components is called the *analysis of variance*, or ANOVA. (Analysis, as in splitting apart.) So far we've only split TV into two components, PV and UV. Later on, we'll learn that the same partitioning effect still holds even when we have more than one X variable, and that we



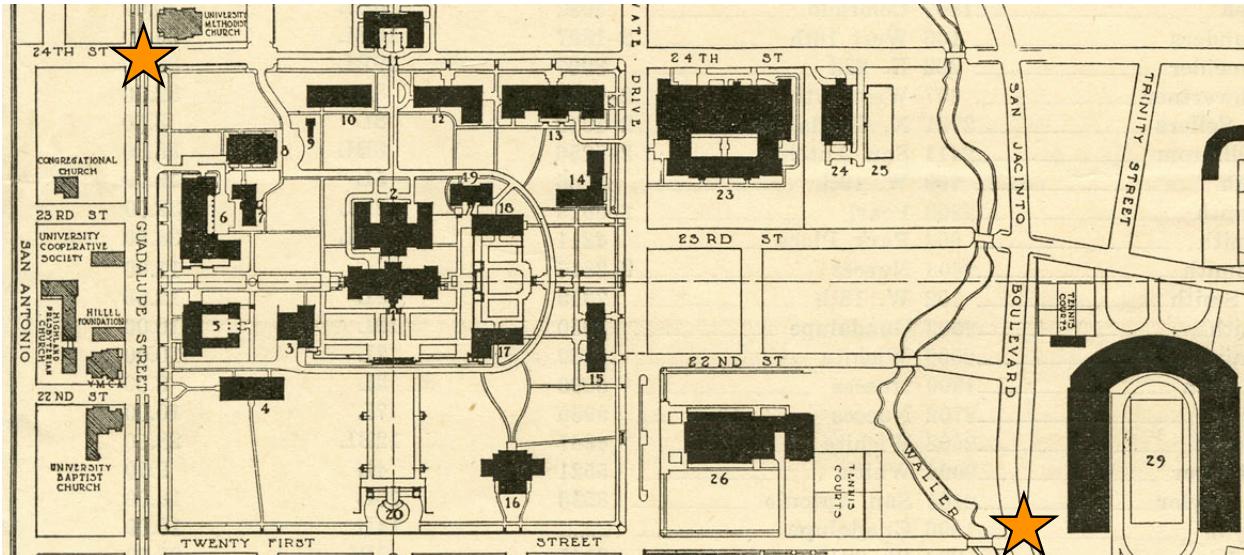
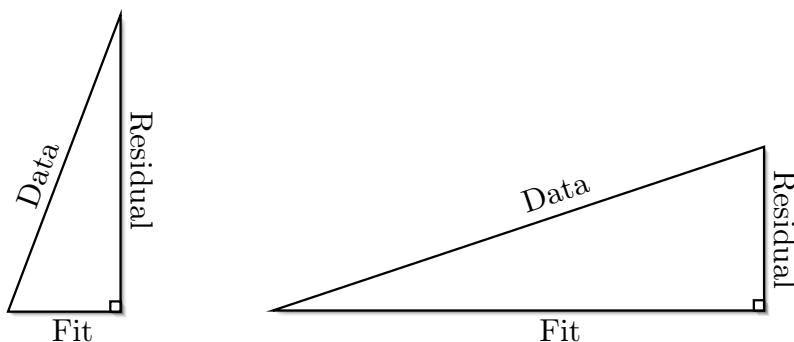


Figure 3.8: A map of the University of Texas in 1930, with two houses of worship highlighted: the University Methodist Church (upper left) and the football stadium (lower right).



can actually sub-partition PV into different components corresponding to the different predictors.

One final note on sums of squares: I've been vague about one crucial point. It turns out that this story about the fitted values and residuals pointing in perpendicular directions isn't a metaphor. It's a genuine mathematical reality—a deep consequence, in fact, of the geometry of vectors in high-dimensional Euclidean space. We'll leave it at the metaphorical level for now, though; it's not that the math is all that hard, but it does require some extra notation that is best deferred to a more advanced treatment of regression. Just be aware that the standard deviations of the three main quantities—the residuals, the fitted values, and the  $y$  values—will always form a Pythagorean triple.

### The coefficient of determination: $R^2$

BY THEMSELVES, sums of squares are hard to interpret, because they are measured in squared units of the  $Y$  variable. But their ratios are highly meaningful. In fact, the ratio of PV to TV—or what fraction of the total variation has been predicted by the model—is one of the most frequently quoted summary measures in all of statistical modeling. This ratio is called the *coefficient of determination*, and is usually denoted by the symbol  $R^2$ :

$$R^2 = \frac{PV}{TV} = 1 - \frac{UV}{TV}.$$

Dividing by TV simultaneously cancels the units of PV and standardizes it by the original scale of the data.

The value of  $R^2$  is a property of a model and a data set considered jointly, and not of either one considered on its own. In analyzing the mammalian sleep-pattern data, for example, we started out with  $TV = 102.1$  squared hours in total variation, and were left with  $UV = 65.7$  squared hours in unpredictable variation after fitting the group-wise model based on the predation index. Therefore  $R^2 = PV/TV \approx 0.36$ , meaning that the model predicts 36% of the total variation in dreaming hours.

The correct interpretation of  $R^2$  sometimes trips people up, and is therefore worth repeating: it is the proportion of variance in the data that can be predicted using the statistical model in question. Here are three common mistakes of interpretation to look out for, both in your own work and in that of others.

*Mistake 1: Confusing  $R^2$  with the slope of a regression line.* We've now encountered three ways of summarizing the dependence between a predictor  $X$  and response  $Y$ :

$r$ , the sample correlation coefficient between  $Y$  and  $X$ .

$\hat{\beta}_1$ , the slope from the least-squares fit of  $Y$  on  $X$ . This describes the average rate of change of the  $Y$  variable as the  $X$  variable changes.

$R^2$ , the coefficient of determination from the least-squares fit of  $Y$  on  $X$ . This measures how much of the variation in  $Y$  can be predicted using the least-squares regression line of  $Y$  on  $X$ :

$$R^2 = 1 - \frac{UV}{TV} = \frac{PV}{TV},$$

An interesting fact is that, for a linear regression model,  $R^2 = r^2$ . That is, the coefficient of determination is precisely equal to the square of the sample correlation coefficient between  $X$  and  $Y$ . This is yet another reason to use correlation only for measuring linear relationships.

or predictable variation divided by total variation.

These are different quantities: the slope  $\beta_1$  quantifies the trend in  $Y$  as a function of  $X$ , while both  $r$  and  $R^2$  quantify the amount of variability in the data that is predictable using the trend.

Another difference is that both  $r$  and  $R^2$  are unit-free quantities, while  $\beta_1$  is not. No matter how  $Y$  is measured, its units cancel out when you churn through the formulas for  $r$  and  $R^2$ —you should try the algebra yourself. This is as it should be:  $r$  and  $R^2$  are meant to provide a measure of dependence that can be compared across different data sets. They must not, therefore, be contingent upon the units of measure for a particular problem.

On the other hand,  $\beta_1$  is measured as a ratio of the units of  $Y$  to units of  $X$ , and is inescapably problem-specific. The slope, after all, is a rate of change:

- If  $X$  is years of higher education and  $Y$  is future salary in dollars, then  $\beta_1$  is dollars per year of education.
- If  $X$  is seconds and  $Y$  is meters, then  $\beta_1$  is meters per second.
- If  $X$  is bits and  $Y$  is druthers, then  $\beta_1$  is druthers per bit.

And so forth.

These quantities are also related to each other. We already know that  $R^2$  is also the square of the sample correlation between  $X$  and  $Y$ . What may come as more of a surprise is that  $R^2$  is *also* the square of the correlation coefficient between  $y_i$  and  $\hat{y}_i$ , the fitted values from the regression line.<sup>2</sup> Intuitively, this is because the least-squares line absorbs all the correlation between  $X$  and  $Y$  into the fitted values  $\hat{y}$ , leaving us with  $r(\hat{y}, x) = r(y, x)$  and  $r(e, x) = 0$ . Remember:  $TV = PV + UV$ , and the  $PV$  is precisely the variation we can explain by taking the “ $X$ -ness” out of  $Y$ .

The upshot is that all three of our summary quantities— $r$ ,  $\hat{\beta}_1$ , and  $R^2$ —can be related to each other in a single line of equations:

$$\{r(y, x)\}^2 = \{r(y, \hat{y})\}^2 = R^2.$$

That is: the squared correlation between  $y$  and  $x$  equals the squared correlation between  $y$  and the fitted values of the model ( $\hat{y}$ ), which also equals the  $R^2$  of the model.

*Mistake 2: Quoting  $R^2$  while ignoring the story in the residuals.* We have seen that the residuals from the least-squares line are uncorrelated with the predictor  $X$ . Uncorrelated, yes—but not necessarily independent. Take the four plots from Figure 1.9, shown

<sup>2</sup> To see this algebraically, note that

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{(n-1)s_y s_{\hat{y}}}.$$

Plug in the fitted values  $\hat{y}_i = \hat{\beta}_0 + x_i \hat{\beta}_1$ , and by churning through the algebra you will be able to recover  $r(y, x)$  at the end.

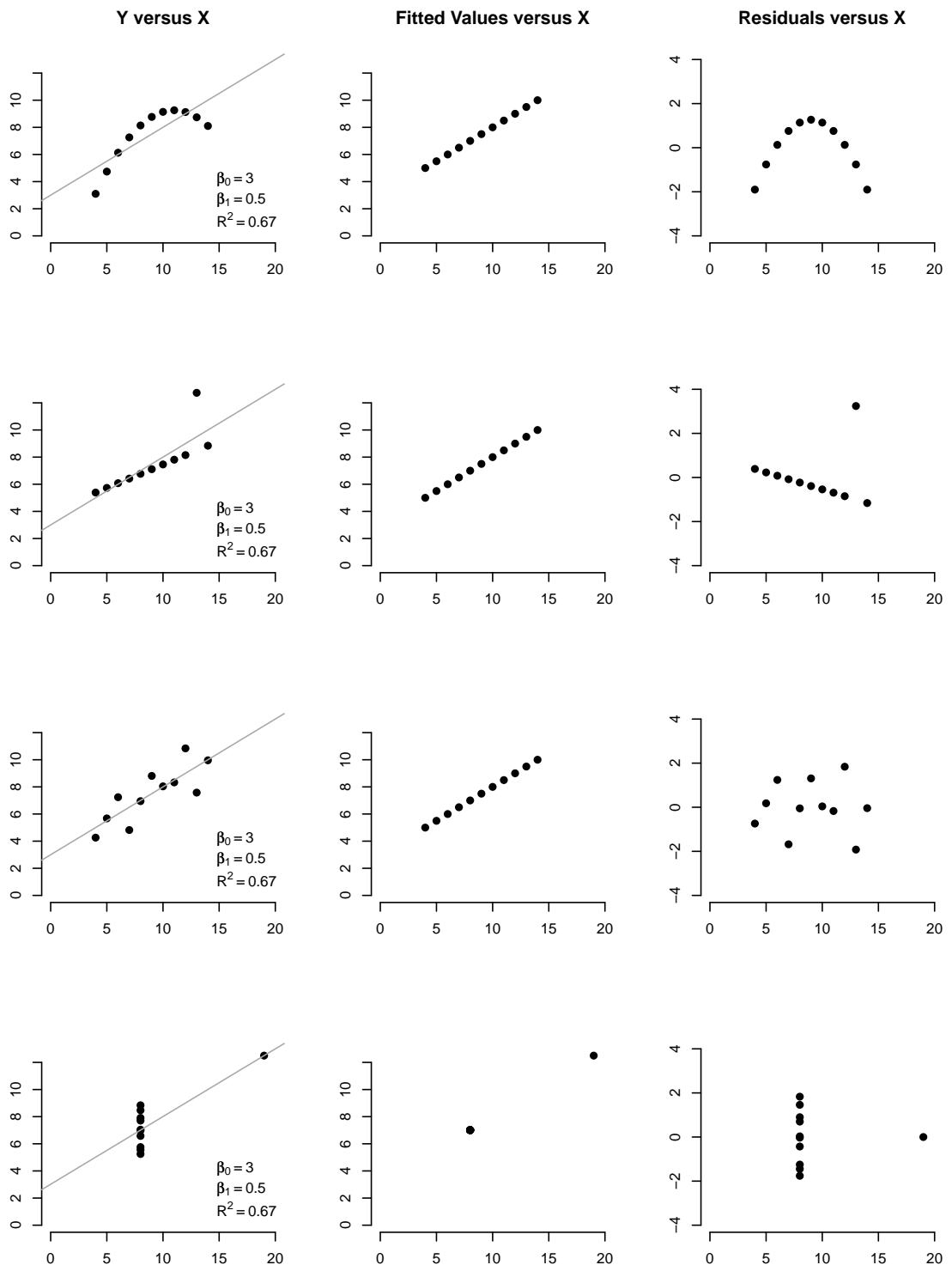


Figure 3.9: These four data sets have the same least-squares line.

again on page 73. These four data sets have the same correlation coefficient,  $r = 0.816$ , despite having very different patterns of dependence between the  $X$  and  $Y$  variable.

The disturbing similarity runs even deeper: the four data sets all have the same least-squares line and the same value of  $R^2$ , too. In Figure 3.9 we see the same set of three plots for each data set: the data plus the least-squares line; the fitted values versus  $X$ ; and the residuals versus  $X$ . Note that in each case, despite appearances, the residuals and the predictor variable have zero sample correlation; this is an inescapable property of least squares.

Despite being equivalent according to just about every standard numerical summary, these data sets are obviously very different from one another. In particular, only in the third case do the residuals seem truly *independent* of  $X$ . In the other three cases, there is clearly still some  $X$ -ness left in  $Y$  that we can see in the residuals. Said another way, there is still information in  $X$  left on the table that we can use for predicting  $Y$ , even if that information cannot be measured using the crude tool of sample correlation. It will necessarily be true that  $r(e, x) = 0$ . But sometimes this will be a truth that lies, and if you plot your data, your eyes will pick up the lie immediately.

The moral of the story is: like the correlation coefficient,  $R^2$  is just a single number, and can only tell you so much. Therefore when you fit a regression, always plot the residuals versus  $X$ . Ideally you will see a random cloud, and no  $X$ -ness left in  $Y$ . But you should watch out for systematic nonlinear trends—for example, groups of nearby points that are all above or below zero together. This certainly describes the first data set, where the real regression function looks to be a parabola, and where we can see a clear trend left over in the residuals. You should also be on the lookout for obvious outliers, with the second and fourth data sets providing good examples. These outliers can be very influential in a standard least-squares fit.

We will soon turn to the question of how to remedy these problems. For now, though, it's important to be able to diagnose them in the residuals.

### *Mistake 3: Confusing statistical explanations with real explanations.*

You will often hear  $R^2$  described as the proportion of variance in  $Y$  “explained” by the statistical model. Do not confuse this usage of the word “explain” with the ordinary English usage of the word,

which inevitably has something to do with causality. This is an insidious ambiguity. As Edward Tufte writes:

A big  $R^2$  means that  $X$  is relatively successful in predicting the value of  $Y$ —not necessarily that  $X$  causes  $Y$  or even that  $X$  is a meaningful explanation of  $Y$ . As you might imagine, some researchers, in presenting their results, tend to play on the ambiguity of the word “explain” in this context to avoid the risk of making an out-and-out assertion of causality while creating the appearance that something really was explained substantively as well as statistically.<sup>3</sup>

You’ll notice that, for precisely this reason, we’ve avoided describing  $R^2$  in terms of “explanation” at all, and have instead referred to it as the “ratio of predictable variation to total variation.”

We know that correlation and causality are not the same thing, and  $R^2$  quantifies the former, not the latter. Consider the data set in the table at right. Regressing the number of patent applications on the number of letters in the vice president’s first name yields  $\hat{\beta}_1 = -26,920$  applications per letter, suggesting a negative trend. Moreover, the regression produces an impressive-looking  $R^2$  of 0.71, meaning that over two-thirds of the variability in patent applications can be predicted using the length of the vice president’s first name alone.

Nothing has been “explained” here at all, the high  $R^2$  notwithstanding. The least-squares fit is capable of answering the question: *if X has a causal linear effect on Y, then what is the best estimate of this effect, and how much variation does this effect account for?* This question assumes a causal hypothesis, and therefore patently cannot be used to test this hypothesis. In particular, calling one variable the “predictor” and the other variable the “response” simply does not decide the issue of causation.

<sup>3</sup> *Data Analysis for Politics and Policy*, p. 72.

Year	Letters in first name of U.S. vice president	Number of U.S. patent applications
2000	2	315,015
1999	2	288,811
1998	2	260,889
1997	2	232,424
1996	2	211,013
1995	2	228,238
1994	2	206,090
1993	2	188,739
1992	3	186,507
1991	3	177,830
1990	3	176,264
1989	3	165,748
1988	6	151,491
1987	6	139,455
1986	6	132,665
1985	6	126,788
1984	6	120,276
1983	6	112,040
1982	6	117,987
1981	6	113,966

Table 3.1: Patent-application data available from the United States Patent and Trademark Office, Electronic Information Products Division.



# 4

## *Grouping variables in regression*

### **Grouping variables and aggregation paradoxes**

THE previous chapters have taught us to fit equations to data involving a numerical response and a numerical predictor. In this chapter, we'll generalize these ideas to incorporate grouping variables as predictors, too.

It's very common in real-world systems for one variable to modulate the effect of another. For example, a person's overall size and weight modulate the relationship between alcohol and cognitive impairment. A single glass of wine might make a small person feel drunk, but have a negligible effect on a big person.

This phenomenon is easiest to visualize in data when the variable that does the modulating is categorical. To see an example of this, we'll revisit the data set on college GPA versus high-school SAT scores. You'll recall that this data set catalogues all 5,191 students at the University of Texas who matriculated in the fall semester of 2000, and who went on to graduate within five years. In Figure 4.2, we notice the expected positive relationship between combined SAT score and final GPA. We also notice the fact that SAT scores and graduating GPAs tend to differ substantially from one college to the next. Figure 4.1 shows boxplots of SAT and GPA stratified by the ten undergraduate colleges at UT.

What we see in Figures 4.2 and 4.1 is an example of an *aggregation paradox*, where the same trend that holds for individuals does not hold for groupings of individuals. Why is this a paradox? Look carefully at the data: Figure 4.1 says that students with higher SAT scores tend to have higher GPAs. Yet this trend does not hold at the college level, even broadly. For example, Engineering students (as a group) have among the highest average SAT scores, and among the lowest average GPAs. Thus we have a paradox: it looks as though high SAT scores predict high GPAs, but being in a college with high SAT scores does not predict being in a

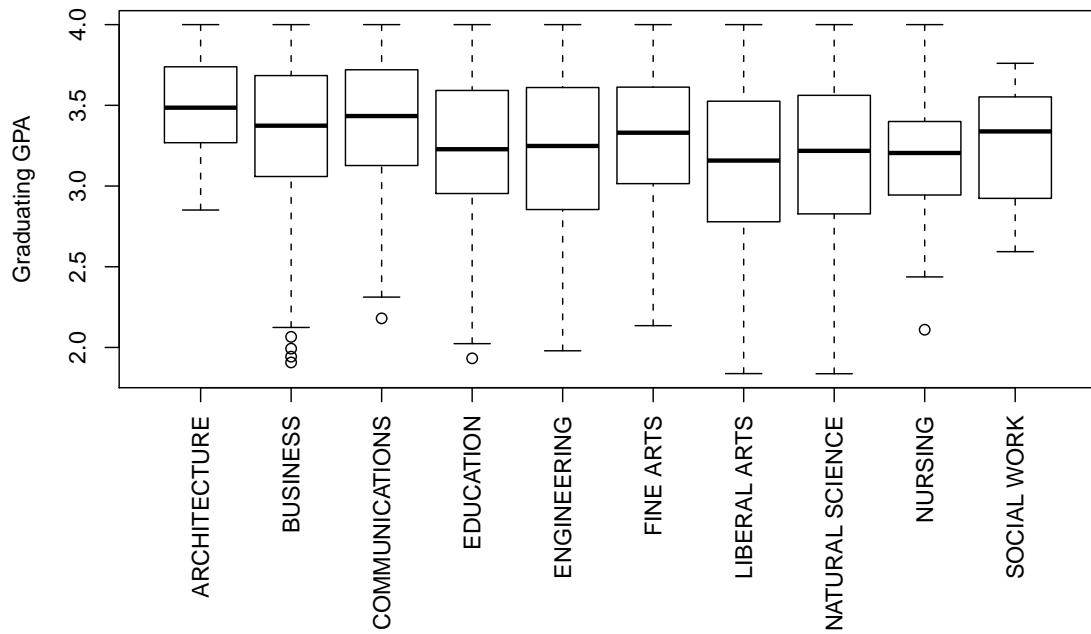
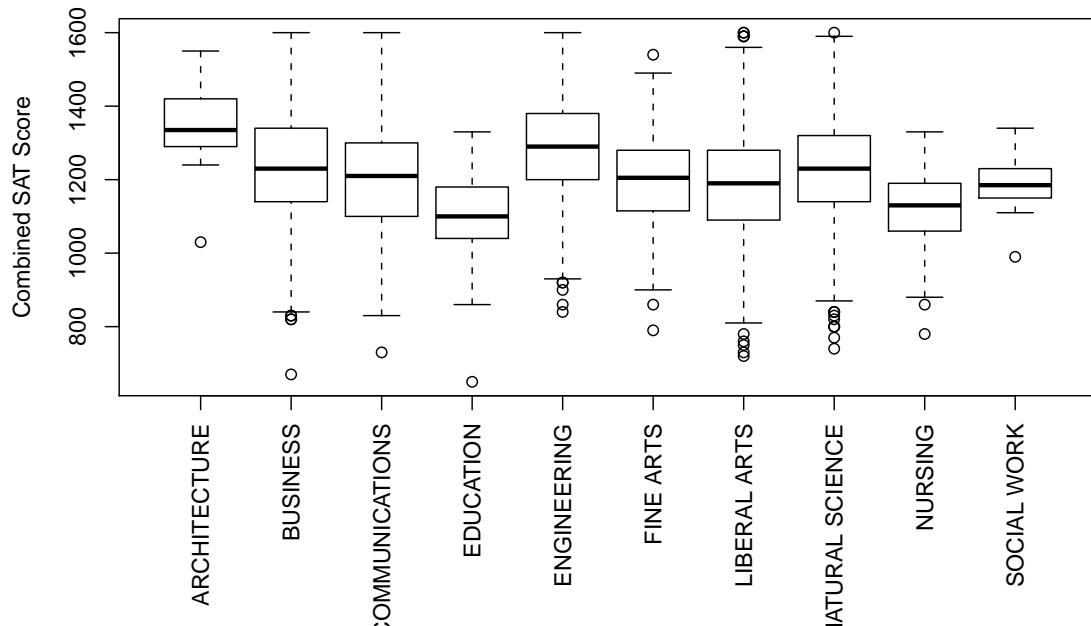
**Graduating GPA for Entering Class of 2000, by College****SAT Scores for Entering Class of 2000, by College**

Figure 4.1: GPA and SAT scores stratified by the ten undergraduate colleges at UT.

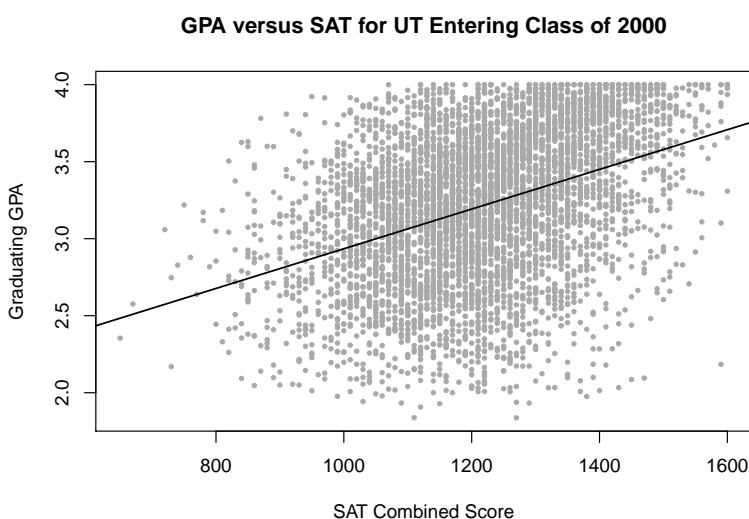


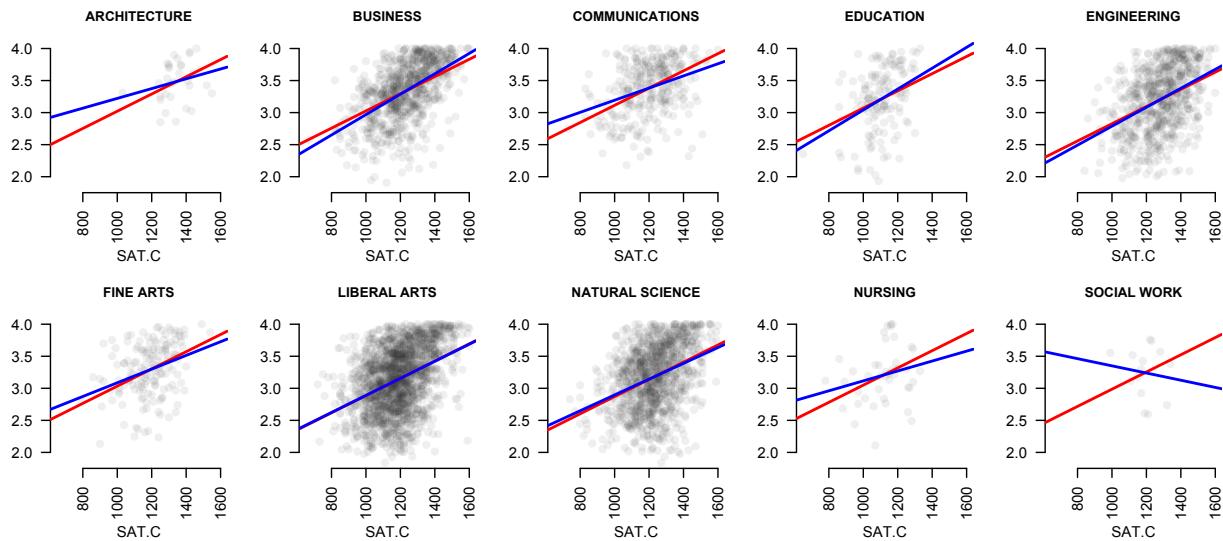
Figure 4.2: Combined SAT scores versus graduating GPA for the entering fall class of 2000 at the University of Texas.

college with high GPAs.

The paradox disappears when we realize the the “College” variable modulates the relationship between SAT score and GPA. A student’s college is systematically associated with both SAT and GPA: some degrees are harder than others, and these degrees tend to enroll students with higher test scores.

The right way to proceed here is to disaggregate the data and fit a different regression line within each of the ten colleges, to account for the effect of the modulating variable. There are two different ways to do this:

1. We could fit ten different lines, each with a different intercept ( $\beta_0^{(k)}$ ), but all with the same slope ( $\beta_1$ ). This would make sense if we thought that the same SAT–GPA relationship ought to hold within each college, but that each college had a systematically higher or lower intercept (average GPA). These are the red lines in Figure 4.3. You can see the differences among the red lines if you look carefully at where they hit the  $y$  axis in relation to a GPA of 2.5—for example, compare Communications and Engineering.
2. We could fit ten different lines, allowing both the slope and the intercept to differ for each college. We would do this if we thought that the SAT–GPA relationship differed fun-



damentally across the colleges. These are the blue lines in Figure 4.3.

But which strategy should we take? And how would we even accomplish strategy 1 using ordinary least squares?

Things get even more complex in the presence of more than one grouping variable. For example, we might want to look at these relationships separately for different years, for men versus women, and for in-state and out-of-state students. To be able to model the effect of all these variables on GPA simultaneously, we will need to introduce some new notation and a few new concepts.

## Models for a single grouping variable

### Dummy variables

LET's return to a simple scenario where we have numerical data that falls into two groups, and we want to compare the variation between the groups. The dotplot in Figure 4.4 shows the weekly sales volume of package sliced cheese over 61 weeks at a Dallas-area Kroger's grocery store. In 38 of these weeks, the store set up a prominent display near the entrance, calling shoppers' attention to the various culinary adventures they might undertake with the

Figure 4.3: Separate regression models fit for GPA versus SAT within each college. The red lines all have the same slope, but a different intercept for each college. The blue lines all have different intercepts and different slopes.

cheese. The data show that, in these 38 weeks, sales were higher overall than when no display was present.

How much higher? The average sales volume in display weeks was 5,577 units (the blue dotted line in Figure 4.4), versus an average of 2341 units in non-display weeks (the red dotted line). Thus sales were 3236 units higher in the display weeks. This difference is depicted in Figure 4.4 as the difference or offset between the dotted lines.

This example emphasizes that in many data sets, we care less about the absolute magnitude of a response under different conditions, and more about the differences between those conditions. We therefore often build our model in such a way that these differences are estimated directly, rather than indirectly (i.e. by calculating means and then subtracting them).

We do this using *indicator* or *dummy* variables. To understand this idea, take the simple case of a single grouping variable  $x$  with two levels: “on” ( $x = 1$ ) and “off” ( $x = 0$ ). We can write this model in “baseline/offset” form:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_i=1\}} + e_i .$$

The quantity  $\mathbf{1}_{\{x_i=1\}}$  is called a dummy variable; it takes the value 1 when  $x_i = 1$ , and the value 0 otherwise. Just as in an ordinary linear model, we call  $\beta_0$  and  $\beta_1$  the *coefficients* of the model. This way of expressing the model implies the following.

$$\begin{aligned} \text{Group mean for case where } x \text{ is off} &= \beta_0 \\ \text{Group mean for case where } x \text{ is on} &= \beta_0 + \beta_1 . \end{aligned}$$

Therefore, we can think of  $\beta_0$  as the baseline (or *intercept*), and  $\beta_1$  as the offset. To see this in action, consult Figure 4.4 again. Here the dummy variable encodes the presence of an in-store display. The red dot at 2341, in the non-display weeks, is  $\beta_0$ . This is the baseline case, when the dummy variable  $x$  is “off.” The coefficient for the dummy variable,  $\beta_1 = 3236$ , is the vertical distance between the two means. Thus if we wanted to reconstruct the mean for the with-display weeks, we would just add the baseline and the offset, to arrive at  $2341 + 3236 = 5577$ , where the blue dot sits.

As before, we estimate the values of  $\beta_0$  and  $\beta_1$  using the least-squares criterion: that is, make the sum of squared errors,  $\sum_{i=1}^n e_i^2$ , as small as possible. This is mathematically equivalent to computing the group-wise means separately, and then calculating the difference between the means.

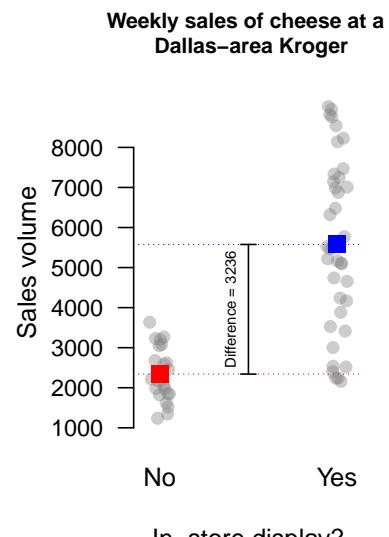
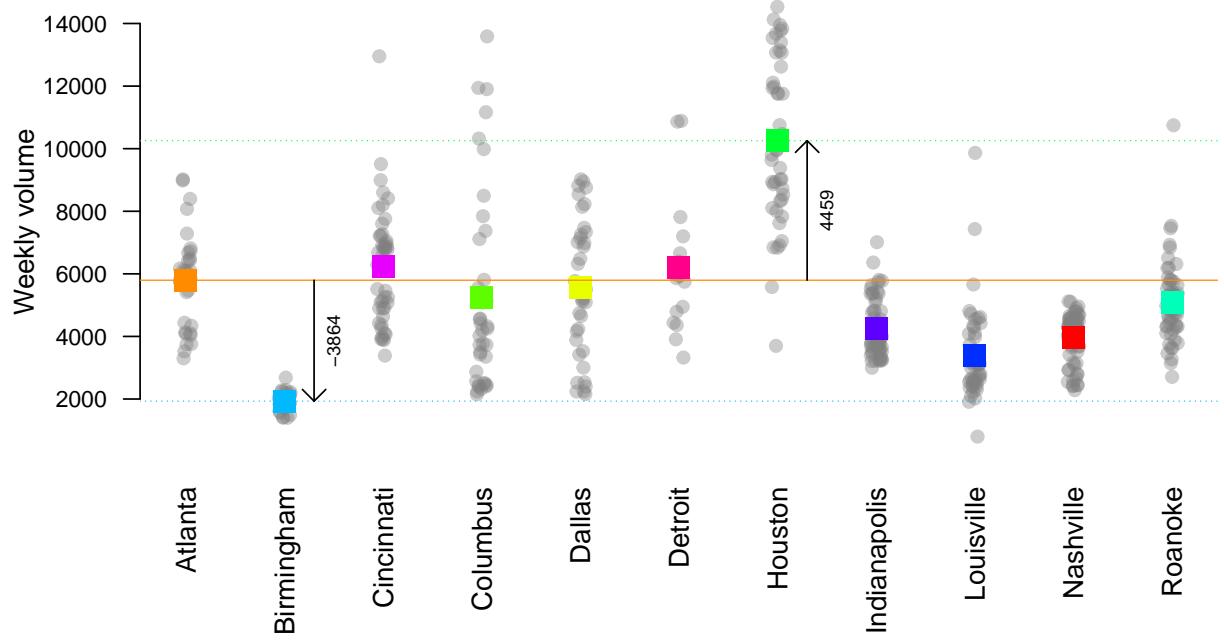


Figure 4.4: Weekly sales of packaged cheese slices at a Dallas-area Kroger’s grocery store, both with and without the presence of an in-store display ad for the cheese. The red dot shows the mean of the no-display weeks, and the blue dot shows the mean of the with-display weeks. The estimated coefficient for the dummy variable that encodes the presence of a display ad is 3236, which is the vertical distance between the two dots.

## Weekly cheese sales at 11 Kroger's stores



*More than two levels*

If the categorical predictor  $x$  has more than two levels, we represent it in terms of more than one dummy variable. Suppose that  $x$  can take three levels, labeled arbitrarily as 0 through 2. Then our model is

$$y_i = \beta_0 + \beta_1^{(1)} \mathbf{1}_{\{x_i=1\}} + \beta_1^{(2)} \mathbf{1}_{\{x_i=2\}} + e_i.$$

The dummy variables  $\mathbf{1}_{\{x_i=1\}}$  and  $\mathbf{1}_{\{x_i=2\}}$  tell you which of the levels is active for the  $i$ th case in the data set.<sup>1</sup>

More generally, suppose we have a grouping variable with  $K$  levels. Then  $\beta_1^{(k)}$  is the coefficient associated with the  $k$ th level of the grouping variable, and we write the full model as a sum of  $K - 1$  dummy-variable effects, like this:

$$y_i = \beta_0 + \sum_{k=1}^{K-1} \beta_1^{(k)} \mathbf{1}_{\{x_i=k\}} + e_i \quad (4.1)$$

Figure 4.5: Weekly sales of packaged cheese slices during weeks with an advertising display at 11 Kroger's grocery stores across the country.

<sup>1</sup> Normal people count starting at 1. Therefore you might find it strange that we start counting levels of a categorical variable at 0. The rationale here is that this makes the notation for group-wise models a lot cleaner compared to starting at 1.

We call this a *group-wise model*. Notice that there is no dummy variable for the case  $x = 0$ . This is the baseline level, whose group mean is the intercept  $\beta_0$ . In general, for a categorical variable with  $K$  levels, we will need  $K - 1$  dummy variables, and at most one of these  $K - 1$  dummy variables is ever active for a single observation. The coefficient on each dummy variable ( $\beta_1^{(k)}$ ) is the differences between the baseline and the mean of group  $k$ :

$$\begin{aligned}\text{Group mean for case where } (x_i = 0) &= \beta_0 \\ \text{Group mean for case where } (x_i = k) &= \beta_0 + \beta_1^{(k)}.\end{aligned}$$

In Figure 4.5, we see an example of a single categorical variable with more than two levels. The figure shows weekly cheese sales (during display-present weeks only) at 11 different Kroger stores in 11 different markets across the country. The grouping variable here is the market: Atlanta, Birmingham, Cincinnati, and so forth. If we fit a model like Equation 4.1 to the data in this figure, choosing Atlanta to be the baseline, we get the set of estimated coefficients in the second column (“Coefficient”) of the table below:

Variable	Coefficient	Group mean
Intercept	5796	—
Birmingham	-3864	1932
Cincinnati	427	6223
Columbus	-543	5253
Dallas	-219	5577
Detroit	400	6196
Houston	4459	10255
Indianapolis	-1542	4254
Louisville	-2409	3387
Nashville	-1838	3958
Roanoke	-717	5079

Atlanta is the baseline, and so the intercept is the group mean for Atlanta: 5796 packages of crappy cheese. To get the group mean for an individual market, we add that market’s offset to the baseline. For example, the mean weekly sales volume in Houston is  $5796 + 4459 = 10255$  units. Group mean = baseline + offset.

The figure also shows you two of the offsets as arrows, to give you a visual sense of what these numbers in the above table represent. The coefficient for Houston is  $\beta_6^{(1)} = 4459$ , because the group

mean for Houston (10255) is 4459 units *higher* than the baseline group mean for Atlanta (a positive offset). Similarly, the coefficient for Birmingham is  $\beta_1^{(1)} = -3864$ , because the group mean for Birmingham (1932) is 3864 units *lower* than the baseline group mean for Atlanta (a negative offset).

*The choice of baseline.* In the above analysis, we chose Atlanta as the baseline level of the grouping variable. This was arbitrary.

We could have chosen any city as a baseline, measuring the other cities as offsets from there instead.

A natural question is: does the model change depending on what level of the grouping variable we choose to call the baseline? The answer is: yes and no. Yes, the estimated model coefficients will change when a different baseline is used; but no, the underlying group means do not change. To see this, consider what happens when we fit another model like Equation 4.1 to the Kroger cheese-sales data, now choosing the Dallas store to be the baseline:

Variable	Coefficient	Group mean
Intercept	5577	—
Atlanta	219	5796
Birmingham	-3644	1932
Cincinnati	646	6223
Columbus	-324	5253
Detroit	619	6196
Houston	4678	10255
Indianapolis	-1323	4254
Louisville	-2190	3387
Nashville	-1619	3958
Roanoke	-498	5079

The intercept is the Dallas group mean of 5577, and the other market-level coefficients have changed from the previous table, since these now represent offsets compared to a different baseline. But the group means themselves do not change. The moral of the story is that the coefficients in a model involving dummy variables *do* depend upon the choice of baseline, but that the information these coefficients encode—the means of the underlying groups—does not. Different choices of the baseline just lead to different ways of expressing this information.

## Multiple grouping variables

WE BEGAN our discussion of dummy variables by looking at a simple group-wise model with a binary predictor, meaning that  $x_i$  is either 0 or 1. Such a model takes the form

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_i=1\}} + e_i.$$

We learned something important about this model: that the coefficient  $\beta_1$  can be interpreted as the differential effect of being in group 1, as opposed to the baseline (group 0).<sup>2</sup> That's a nice feature of using dummy variables: if we care primarily about the difference in the average response between conditions, we get an estimate of that difference ( $\hat{\beta}_1$ ) directly from the fitted model.

This approach of using dummy variables to encode the grouping structure of our data really comes into its own when we encounter data sets with more than one grouping variable. To see why, we'll spend some time with the data in Figure 4.6.

<sup>2</sup> Remember, we start counting groups/levels at 0.

### Main effects

Making a best-selling video game is hard. Not only do you need a lot of cash, a good story, and a deep roster of artists, but you also need to make the game fun to play. Take Mario Kart for the Super Nintendo, my favorite video game from childhood. In Mario Kart, you had to react quickly to dodge banana peels and Koopa shells launched by your opponents as you all raced virtual go-karts around a track. The game was calibrated just right. If the required reaction time had been just a little slower, the game would have been too easy, and therefore boring. But if the required reaction time had been a little bit faster, the game would have been too hard, and therefore also boring.

Human reaction time to visual stimuli is a big deal to video game makers. They spend a lot of time studying it and adjusting their games according to what they find. Figure 4.6 shows the results of one such study. Participants were presented with a natural scene on a computer monitor, and asked to react (by pressing a button) when they saw an animated figure appear in the scene.<sup>3</sup>

The experimenters varied the conditions of the natural scene: some were cluttered, while others were relatively open; in some, the figure appeared far away in the scene, while in others it appeared close up. They presented all combinations of these conditions to each participant many times over. The top two panels of

<sup>3</sup> Essentially the company was measuring how quickly people could react to a bad guy popping up on the screen in a video game.

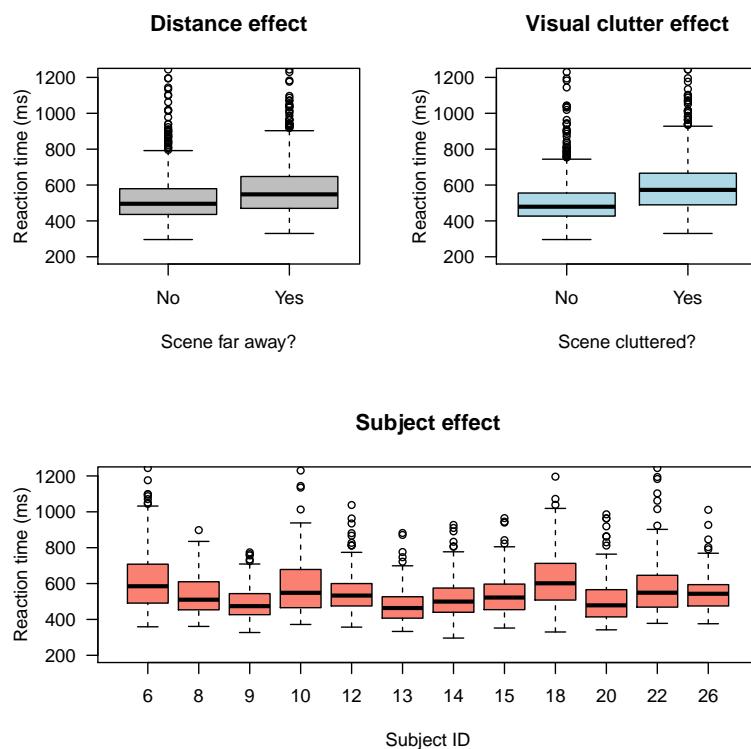


Figure 4.6: Reaction time to visual stimuli in a controlled experiment run by a major video-game maker. Top-left: participants reacted more slowly, on average, when the stimulus was far away within the scene. Top-right: participants reacted more slowly, on average, in a scene with significant visual clutter. Bottom: systematic differences in reaction time across participants in the trial.

Figure 4.6 show boxplots of all participants' reaction times across all trials under these varying conditions. On average, participants reacted more slowly to scenes that were far away (top left panel) and that were cluttered (top right panel).

We'll return to the bottom panel of Figure 4.6 shortly. For now, let's focus on the "distance effect" and the "clutter effect" in the top two panels. This presents us with the case of two grouping variables,  $x_1$  and  $x_2$ , each of which affects the response variable, and each of which can take the value 0 ("off") or 1 ("on"). To account for this, we need to build a model that is capable of describing the joint effect of both variables at once.

*Strategy 1: slice and dice.* One approach to modeling the joint effect of  $x_1$  and  $x_2$  on the response  $y$  is to slice and dice the data. In other words: take subsets of the data for each of the four combinations of  $x_1$  and  $x_2$ , and compute the mean within each subset. For our video-game data, we get the result in Table 4.1. Clearly the

Table 4.1: Mean reaction time across all trials and participants for the four combinations of the two experimental factors in the video game data.

Cluttered	Far away	Time (ms)
No	No	491
	Yes	522
Yes	No	559
	Yes	629

“cluttered + far away” scenes are the hardest, on average.

This slice-and-dice approach is intuitively reasonable, but combinatorially explosive. With only two binary grouping variables, we have four possible combinations—not a big deal. But suppose we had 10 binary grouping variables instead. Then there would be  $2^{10} = 1024$  possible subsets of the data, and thus 1024 group-wise means to estimate. For a scenario like this, if you were to take the slice-and-dice approach, you would need a lot of data—and not merely a lot of data overall, but a lot of data for each combination separately.

*Strategy 2: use dummy variables.* A second strategy is to estimate the effect of  $x_1$  and  $x_2$  by building a model that uses dummy variables. Intuitively, the model we'll fit assumes that the response can be expressed as:

$$y_i = \hat{y}_i + e_i = \text{Baseline} + (\text{Effect if } x_{i1} \text{ on}) + (\text{Effect if } x_{i2} \text{ on}) + \text{Residual}.$$

Notice that we need two subscripts on the predictors  $x_{i1}$  and  $x_{i2}$ :  $i$ , to index which case in the data set is being referred to; and 1 or 2, to indicate which categorical predictor is being referred to (e.g. far away versus cluttered).

This notation gets cumbersome quickly. We can write it more concisely in terms of dummy variables, just as we learned to do in the case of a single grouping variable:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_{i1}=1\}} + \beta_2 \mathbf{1}_{\{x_{i2}=1\}} + e_i.$$

Notice how the dummy variables affect the expected value of  $y_i$  by being either present or absent, depending on the case. For example, if  $x_{i2} = 0$ , then the  $\beta_2 \mathbf{1}_{\{x_{i2}\}}$  term falls away, and we're left with the baseline, plus the effect of  $x_1$  being on, plus the residual. We refer to  $\beta_1$  and  $\beta_2$  as the *main effects* of the model, for reasons that will become clear in a moment.

If we fit this model to the video-game data in Figure 4.1, we get the equation

$$\text{Reaction} = 482 + 87 \cdot \mathbf{1}_{\{x_{i1}=1\}} + 50 \cdot \mathbf{1}_{\{x_{i2}=1\}} + \text{Residual}, \quad (4.2)$$

where  $x_{i1} = 1$  means that the scene was cluttered, and  $x_{i2} = 1$  means that the scene was far away. This equation says that if the scene was cluttered, the average reaction time became 87 milliseconds slower; while if the scene was far away, the average reaction time became 50 milliseconds slower.

### *Interactions*

A key assumption of the model in Equation 4.2 is that the effects of clutter and distance on reaction time are separable. That is, if we want to compute the joint effect of both conditions, we simply add the individual effects together.

But what if the effects of  $x_1$  and  $x_2$  aren't separable? We might instead believe a model like this:

$$y_i = \text{Baseline} + (\text{Effect if } x_1 \text{ on}) + (\text{Effect if } x_2 \text{ on}) + (\text{Extra effect if both } x_1 \text{ and } x_2 \text{ on}) + \text{Residual}.$$

In the context of our video-games data, this would imply that there's something different about scenes that are both cluttered *and* far away that cannot be described by just summing the two individual effects.

The world is full of situations like this, where the whole is different than the sum of the parts. The ancient Greeks referred to this idea as *συνέργεια*, or synergia. This roughly means "working together," and it's the origin of the English word "synergy." Synergies abound:

- Neither an actor nor a cameraman can do much individually, but together they can make a film.
- Two hydrogens and an oxygen make water, something completely unlike either of its constituent parts.
- Biking up a hill is hard. Biking in a big gear is hard. Biking up a hill in a big gear is impossible, unless you take drugs.

Examples of the whole being worse than the sum of the parts also abound—groupthink on committees, ill-conceived corporate mergers, Tylenol and alcohol, and so forth.<sup>4</sup>

In statistics, we operationalize the idea of synergy using *interactions among variables*. An interaction is what we get when we multiply two variables together. In the case of two binary categorical predictors, a model with an interaction looks like this:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_1=1\}} + \beta_2 \mathbf{1}_{\{x_2=1\}} + \beta_{12} \mathbf{1}_{\{x_1=1\}} \mathbf{1}_{\{x_2=1\}} + e_i.$$

We call  $\beta_{12}$  an *interaction term*; this term disappears from the model unless  $x_1$  and  $x_2$  are both equal to 1. Fitting this model to the video-games data gives the following estimates:

$$\text{Reaction} = 491 + 68 \cdot \mathbf{1}_{\{x_{i1}=1\}} + 31 \cdot \mathbf{1}_{\{x_{i2}=1\}} + 39 \cdot \mathbf{1}_{\{x_{i1}=1\}} \mathbf{1}_{\{x_{i2}=1\}} + \text{Residual},$$

We interpret this model as follows:

- The baseline reaction time for scenes that are neither cluttered nor far away is 491 milliseconds (ms).

<sup>4</sup> Don't take Tylenol and alcohol together or you'll risk liver damage.

- The main effect for the “cluttered” variable is 68 ms.
  - The main effect for the “far away” variable is 31 ms.
  - The interaction effect for “cluttered” and “far away” is 39 ms.
- In other words, scenes that are both cluttered and far away yield average reaction times that are 39 milliseconds slower than what you would expect from summing the individual effects of the two variables.

From these main effects and the interaction we can use the model to summarize the expected reaction time under any combination of experimental variables:

- $(x_1 = 0, x_2 = 0)$ :  $\hat{y} = 491$  (neither cluttered nor far).
- $(x_1 = 1, x_2 = 0)$ :  $\hat{y} = 491 + 68 = 559$  (cluttered, near).
- $(x_1 = 0, x_2 = 1)$ :  $\hat{y} = 491 + 31 = 522$  (not cluttered, far).
- $(x_1 = 1, x_2 = 1)$ :  $\hat{y} = 491 + 68 + 31 + 39 = 629$  (cluttered, far).

A key point regarding the fourth case in the list is that, when a scene is both cluttered and far away, both the main effects *and* the interaction term enter the prediction. You should also notice that these predictions exactly match up with the group means in Table 4.1 on page 86.

### *Incorporating still more categorical predictors*

Once you understand the basic recipe for incorporating two categorical predictors, you can easily extend that recipe to build a model involving more than two. For example, let’s return one last time to the video-game data in Figure 4.6 on page 86. So far, we’ve been ignoring the bottom panel, which shows systematic differences in reaction times across different subjects in the study. But we can also incorporate subject-level dummy variables to account for these differences. The actual model equation starts to get ugly with this many dummy variables, so we often use a shorthand that describes our model intuitively rather than mathematically:

$$\begin{aligned} \text{Time} \sim & \text{ Clutter effect} + (\text{Distance effect}) \\ & + (\text{Interaction of distance/clutter}) + (\text{Subject effects}). \end{aligned} \quad (4.3)$$

Here the  $\sim$  symbol means “is modeled by” or “is predicted by.”

There are 12 subjects in the data set. Thus to model the subject-level effects, we introduce 11 dummy variables, in a manner similar to what was done in Equation 4.1. The estimated coefficients for this model are in Table 4.2.

Table 4.2: Fitted coefficients for the model incorporating subject-level dummy variables into the video-game data. Remember,  $K$  levels of a factor require  $K - 1$  dummy variables, because one level—in this case, the subject labeled “Subject 6” in Figure 4.6—is the baseline.

Variable	$\hat{\beta}$
Intercept	570
Cluttered	68
FarAway	31
Subject 8	-90
Subject 9	-136
Subject 10	-44
Subject 12	-76
Subject 13	-147
Subject 14	-112
Subject 15	-93
Subject 18	-8
Subject 20	-118
Subject 22	-34
Subject 26	-79
Cluttered:FarAway	39

*When to include interactions.* In the model above, we're assuming that clutter and distance affect all subjects in the same way. Thus we have 15 parameters to estimate: an intercept/baseline, two main effects for Littered and FarAway, one interaction term, and 11 subject-level dummy variables. If instead we were to compute the groupwise means for all possible combinations of subject, clutter, and distance, we'd have 48 parameters to estimate: the group mean for each combination of 12 subjects and 4 experimental conditions. Moreover, we'd be implicitly assuming an interaction between the experimental conditions and the subject, allowing clutter and distance to affect each person's average reaction time in a different way, rather than all people in the same way.

This example should convey the power of using dummy variables and interactions to express how a response variable changes as a function of several grouping variables. This framework forces us to be explicit about our assumptions, but it also allows us to be selective about the complexity of our models. Compare estimating 15 parameters versus estimating 48 parameters in the video-games example—that's a big difference in what we're asking of our data.

The essence of the choice is this:

- If a variable affects the response in a similar way under a broad range of conditions, regardless of what the other variables are doing, then that variable warrants only a main effect in our model.
- But if a variable's effect is modulated by some other variable, we should describe that using an interaction between those two variables.

The choice of which variables interact with which other ones should ideally be guided by knowledge of the problem at hand. For example, in a rowing race, a strong headwind makes all crews slower. But wind affects lighter crews more than heavier crews: weight modulates the effect of wind. Thus if we want to build a model to predict the winner of an important race, like [the one between Oxford and Cambridge every spring](#) on the Thames, we should strongly consider including an interaction between wind speed and crew weight. This is something that anyone with knowledge of rowing could suggest, even before seeing any data. But the choice of whether to include an interaction term in a model can also be guided by the data itself. We will now learn about a process called the analysis of variance that can help us

address this important modeling question.

Before we get there, however, here's one final generic guideline about interactions: it is highly unusual to include an interaction in a regression model without also including both corresponding main effects. There are various technical math reasons why most textbooks warn you about this, and why I'm doing so now. But the most important concern is that it is very difficult to interpret a model having interaction terms but no main effects. You should fit such a model only if you have a very good reason.

## ANOVA: the analysis of variance

THE model in Equation 4 postulates four effects on the reaction time for the video-game data: (1) an effect due to visual clutter; (2) an effect due to distance of the stimulus in the scene; (3) an interaction effect (synergy) of distance and clutter; and (4) effects due to differences among experimental subjects. The  $R^2$  for this model is about 0.23, and the residual standard deviation is about 126 milliseconds. This tells us something about the overall predictive abilities of the model. But can we say something about the predictive abilities of the individual variables within this model?

Yes, we can, by conducting an analysis of variance (ANOVA). An analysis of variance is just a simple book-keeping exercise aimed at attributing credit to individual variables in a model. To run an ANOVA, we build a model one step at time, adding one new variable (or one new interaction among variables) at each step. Every time we do this, we ask two questions:

- (1) How many parameters did we have to add to the model to account for the effects of this variable?<sup>5</sup> This is usually called the *degrees of freedom* associated with that parameter.
- (2) By how much did we improve the predictive power of the model when we added this variable? There are a couple of ways to measure this. First, remember the variance decomposition:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{TV} = \text{PV} + \text{UV}.$$

Every time we add a new variable to a model, the total variation in the response variable (TV) stays the same, but we move

<sup>5</sup> For example, we needed to add 11 parameters to account for the "Subject" variable in the video-games data, because we needed to represent this information in terms of 11 dummy variables.

some of this variation out of the “unpredictable” column (UV) and into the “predictable” (PV) column. As a result,  $R^2$  will always go up as a result of adding a variable to a model. In ANOVA, we keep track of the precise numerical value of this change in  $R^2$ .

We could also measure the improvement in the model’s predictive power using the residual standard deviation, which we recall is calculated using the formula

$$s_e = \sqrt{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

There’s an important difference with  $R^2$  here, in that  $s_e$  can actually get worse (i.e. go up) when we add a variable to a model. If this happens, it is generally a good indication of overfitting.

The final result of an analysis of variance is a table—called the ANOVA table—that shows the answers to these two questions at each model-building step.

Let’s take the specific example of our model for the video-games data, for which  $TV = 39,190,224$ . We’ll add one variable at a time and track how  $TV$  is partitioned among PV and UV.<sup>6</sup>

*Step 1.* First, we add an effect due to visual clutter ( $Time \sim Clutter$ ). The variance decomposition for this model is

$$\begin{array}{rcl} 39,190,224 & = & 3,671,938 + 35,518,285 \\ TV & & PV \\ & & UV \end{array}$$

Thus the clutter effect gets credit for predicting 3,671,938 (out of a possible 39,190,224) units of total variation, at the cost of adding one parameter to the model.

*Step 2.* Next, we add the distance effect to the model already containing the clutter variable ( $Time \sim Clutter + Distance$ ). The new variance decomposition is:

$$\begin{array}{rcl} 39,190,224 & = & 4,878,397 + 34,311,827 \\ TV & & PV \\ & & UV \end{array}$$

The previous PV was 3,671,938, and the new one is 4,878,397. Thus the distance effect gets credit for  $4,878,397 - 3,671,938 = 1,206,459$  units of total variation.

<sup>6</sup> The quantity  $TV = 39,190,224$  highlights one feature that makes ANOVA tricky at first: the units are non-intuitive, since we measure improvement using sums of squares. Here the units are squared milliseconds; when you square a quantity like 1000 ms (1 second), you get 1,000,000 ms<sup>2</sup>, which is why we’re seeing numbers in the millions here.

Variable added	# Pars (DF)	$\Delta PV$	$R^2$	$\Delta R^2$	$s_e$	$\Delta s_e$
Intercept only	1		0.000		142.9	
Clutter	1	3671938	0.094	0.094	136.1	6.8
Distance	1	1206459	0.125	0.031	133.8	2.3
Clutter:Distance	1	183633	0.129	0.005	133.5	0.3
Subject	11	4060822	0.223	0.104	125.6	7.9
Predictable Variation		9122852				
Unpredictable Variation		30067371				
Total Variation		39190224				

*Step 3.* Third, we add the interaction of distance and clutter to the previous model ( $\text{Time} \sim \text{Clutter} + \text{Distance} + \text{Clutter:Distance}$ ). The new variance decomposition is:

$$39,190,224 = 5,062,030 + 34,128,194. \quad \begin{matrix} TV \\ PV \\ UV \end{matrix}$$

The previous PV was 4,878,397, and the new one is only slightly better at 5,062,030. Thus the interaction effect gets credit for a measly  $5,062,030 - 4,878,397 = 183,633$  units of total variation.

*Step 4.* Finally—almost done here—we add the 11 subject-level dummy variables to the previous model ( $\text{Time} \sim \text{Clutter} + \text{Distance} + \text{Clutter:Distance} + \text{Subject}$ ). The new variance decomposition reveals a big bump in PV:

$$39,190,224 = 9,122,852 + 30,067,371. \quad \begin{matrix} TV \\ PV \\ UV \end{matrix}$$

The previous PV was 5,062,030, and the new one is better at 9,122,852. Thus the subject effects get credit for  $9,122,852 - 5,062,030 = 4,060,822$  units of total variation.

*Interpreting the ANOVA table.* As you've now seen, the analysis of variance really is just bookkeeping! The ANOVA table for the final model ( $\text{Time} \sim \text{Clutter} + \text{Distance} + \text{Clutter:Distance} + \text{Subject}$ ) is shown in Table 4.3. The change in predictable variation at each stage gives us a more nuanced picture of the model, compared with simply quoting  $R^2$ , because it allows us to partition credit among the individual predictor variables in the model.

The most intuitive way to summarize this information is to track the change in  $R^2$  and residual standard deviation ( $s_e$ ) at

Table 4.3: The analysis of variance (ANOVA) table for the model incorporating effects due to clutter, distance, and subject, along with an interaction between clutter and distance. In an ANOVA table, we add each variable in stages, one at a time. “# Pars” refers to the number of new parameters added to the model at each stage.  $\Delta PV$  refers to the change in predictable variation at each stage.  $R^2$  is the coefficient of determination for the model at each stage, and  $s_e$  is the residual standard deviation. Remember that  $R^2$  always goes up when we add a variable, while  $s_e$  usually (but not always) goes down.

each step. For example, in Table 4.3, it's clear that accounting for subject-level variation improves our predictions the most, followed by clutter and then distance. The distance–clutter interaction contributes a small amount to the predictive ability of the model, relatively speaking: it improves  $R^2$  by only half a percentage point. In fact, the distance/clutter interaction looks so negligible that we might even consider removing this effect from the model, just to simplify. We'll revisit this question later in the book, when we learn some more advanced tools for statistical hypothesis testing and predictive model building.

Finally, always remember that the construction of an ANOVA table is inherently sequential. For example, first we add the clutter variable, which remains in the model at every subsequent step; then we add the distance variable, which remains in the model at every subsequent step; and so forth. Thus the actual question being answered at each stage of an analysis of variance is: how much variation in the response can this new variable predict, in the context of what has already been predicted by other variables in the model? This point—the importance of context in interpreting an ANOVA table—is subtle, but important. We'll revisit it soon, when we discuss the issues posed by correlation among the predictor variables in a regression model.

## Numerical and grouping variables together

Now we are ready to add a continuous predictor into the mix. Start with the simplest case of two predictors for each observation: a grouping variable  $x_{i,1}$  that can take levels 0 to  $K$ , and a numerical predictor  $x_{i,2}$ . We start with the regression equation involving a set of  $K$  dummy variables, and add the effect of the continuous predictor onto the right-hand side of the regression equation:

$$y_i = \beta_0 + \beta_1^{(1)} \mathbf{1}_{\{x_{i1}=1\}} + \beta_1^{(2)} \mathbf{1}_{\{x_{i1}=2\}} + \cdots + \beta_1^{(K)} \mathbf{1}_{\{x_{i1}=K\}} + \beta_2 x_{i2} + e_i .$$

Now each group has its own regression equation:

$$\text{Regression equation for case where } (x_i = 0): \quad y_i = \beta_0 + \beta_2 x_{i2} + e_i$$

$$\text{Regression equation for case where } (x_i = k): \quad y_i = (\beta_0 + \beta_1^{(k)}) + \beta_2 x_{i2} + e_i .$$

Each line has a different intercept, but they all have the same slope. These are the red lines in Figure 4.3 back on page 80.

The coefficients  $\beta_1^{(k)}$  are associated with the dummy variables that encode which college a student is in. Notice that only one of these dummy variables will be 1 for each person, and the rest will be zero, since a person is only in one college. Here's the regression output when we ask for a model of  $\text{GPA} \sim \text{SAT.C} + \text{School}$ :

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.678365	0.096062	17.472	<2e-16 ***
SAT.C	0.001343	0.000043	31.235	<2e-16 ***
SchoolBUSINESS	0.004676	0.078285	0.060	0.9524
SchoolCOMMUNICATIONS	0.092682	0.080817	1.147	0.2515
SchooledUCATION	0.048688	0.085520	0.569	0.5692
SchoolENGINEERING	-0.195433	0.078460	-2.491	0.0128 *
SchoolFINE ARTS	0.012366	0.084427	0.146	0.8836
SchoolLIBERAL ARTS	-0.134092	0.077629	-1.727	0.0842 .
SchoolNATURAL SCIENCE	-0.150631	0.077908	-1.933	0.0532 .
SchoolNURSING	0.028273	0.102243	0.277	0.7822
SchoolSOCIAL WORK	-0.035320	0.139128	-0.254	0.7996

There is no dummy variable associated with Architecture, because it is the baseline case, against which the other colleges are compared. The regression coefficients associated with the "School" dummy variables then shift the line systematically up or down relative to the global intercept, but they do not change the slope of the line. As the math above shows, we are fitting a model where all colleges share a common slope, but have unique intercepts (11 parameters total). This is clearly a compromise solution between two extremes: fitting a single model, with one slope and one intercept common to all colleges (2 parameters); versus fitting ten distinct models for the ten individual colleges, each with their slope and intercept (20 parameters).

### *Interactions between grouping and numerical variables*

We can also have modulating effects between numerical and grouping predictors. For example, we might expect that, for students in Liberal Arts, GPA's will vary more sharply with SAT Verbal scores, and less sharply with Math scores, than for students in Engineering. Mathematically, this means that College modulates the slope of the linear relationship between GPA and SAT scores.

If this is the case, then we should include an interaction term in the model. Remember, in statistical models, interactions are

formed by multiplying two predictors together—in this case, a numerical predictor and a dummy (0–1) variable. When the dummy variable is 0, the interaction term disappears. But when the dummy is 1, the interaction is equal to the original quantitative predictor, whose effective partial slope then changes.

Let's take a simple example involving baseball salaries, plotted in Figure 4.7 on page 97. On the  $y$ -axis are the log salaries of 142 baseball players. On the  $x$ -axis are their corresponding batting averages. The kind of mark indicates whether the player is in the Major League, AAA (the highest minor league), or AA (the next-highest minor league). The straight lines reflect the least-squares fit of a model that regresses log salary upon batting average and dummy variables for a player's league. The corresponding model equation looks like this:

$$\hat{y}_i = \beta_0 + \underbrace{\beta_1^{(AAA)} \cdot 1_{AAA} + \beta_1^{(MLB)} \cdot 1_{MLB}}_{\text{Dummy variables}} + \beta_1 \cdot AVG$$

The three lines are parallel: the coefficients on the dummy variables shift the line up or down as a function of a player's league.

But if we want the slope to change with league as well—that is, if we want league to modulate the relationship between salary and batting average—then we must fit a model like this:

$$\hat{y}_i = \beta_0 + \underbrace{\beta_1^{(AAA)} \cdot 1_{AAA} + \beta_1^{(MLB)} \cdot 1_{MLB}}_{\text{Dummy variables}} + \beta_2 \cdot AVG + \underbrace{\beta_3^{(AAA)} \cdot AVG \cdot 1_{AAA} + \beta_3^{(MLB)} \cdot AVG \cdot 1_{MLB}}_{\text{Interaction terms}}$$

The  $y$  variable depends on  $\beta_0$  and  $\beta_2$  for all players, regardless of league. But when a player is in AAA, the corresponding dummy variable ( $1_{AAA}$ ) fires. Before, when a dummy variable fired, the entire line was merely shifted up for down (as in Figure 4.7). Now, an offset to the intercept ( $\beta_1^{(AAA)}$ ) and an offset to slope ( $\beta_3^{(AAA)}$ ) are activated. Ditto for players in the Major League: then the MLB dummy variable ( $1_{MLB}$ ) fires, and both an offset to the intercept ( $\beta_1^{(MLB)}$ ) and an offset to the slope ( $\beta_3^{(MLB)}$ ) are activated:

$$\text{Regression equation for AA: } y_i = (\beta_0) + (\beta_2) \cdot AVG + e_i$$

$$\text{Regression equation for AAA: } y_i = (\beta_0 + \beta_1^{(AAA)}) + (\beta_2 + \beta_3^{(AAA)}) \cdot AVG + e_i$$

$$\text{Regression equation for MLB: } y_i = (\beta_0 + \beta_1^{(MLB)}) + (\beta_2 + \beta_3^{(MLB)}) \cdot AVG + e_i.$$

Fitting such model produces a picture like the one in Figure 4.8.

Without any interaction terms, the fitted model is:

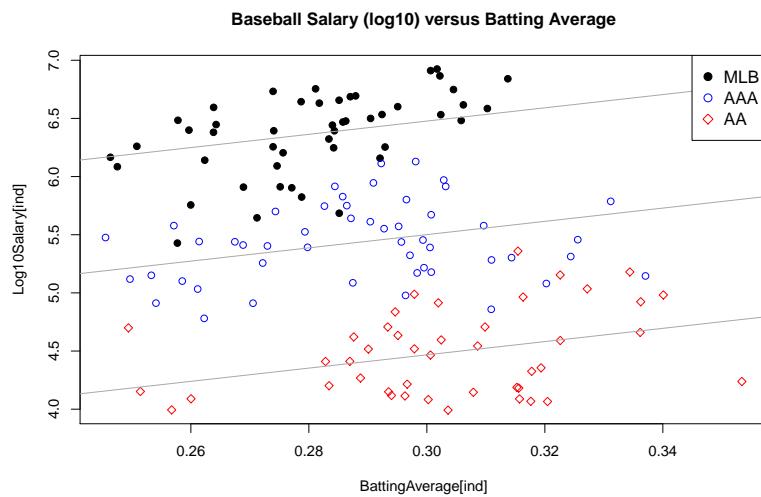


Figure 4.7: Baseball salaries versus batting average for Major League, AAA, and AA players.

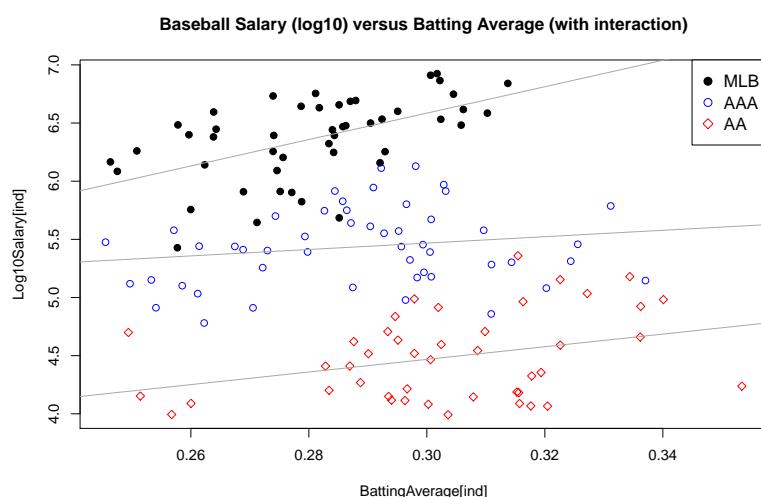


Figure 4.8: Baseball salaries versus batting average for Major League, AAA, and AA players. The fitted lines show the model with an interaction term between batting average and league.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.75795	0.41893	6.583	8.88e-10 ***
BattingAverage	5.69745	1.37000	4.159	5.59e-05 ***
ClassAAA	1.03370	0.07166	14.426	< 2e-16 ***
ClassMLB	2.00990	0.07603	26.436	< 2e-16 ***

---

Multiple R-squared: 0.845, Adjusted R-squared: 0.8416

With the interaction terms, we get:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.8392	0.6718	4.227	4.33e-05 ***
BattingAverage	5.4297	2.2067	2.461	0.0151 *
ClassAAA	1.8024	0.9135	1.973	0.0505 .
ClassMLB	0.3393	1.0450	0.325	0.7459
BattingAverage:ClassAAA	-2.6758	3.0724	-0.871	0.3853
BattingAverage:ClassMLB	5.9258	3.6005	1.646	0.1021

---

Multiple R-squared: 0.8514, Adjusted R-squared: 0.846

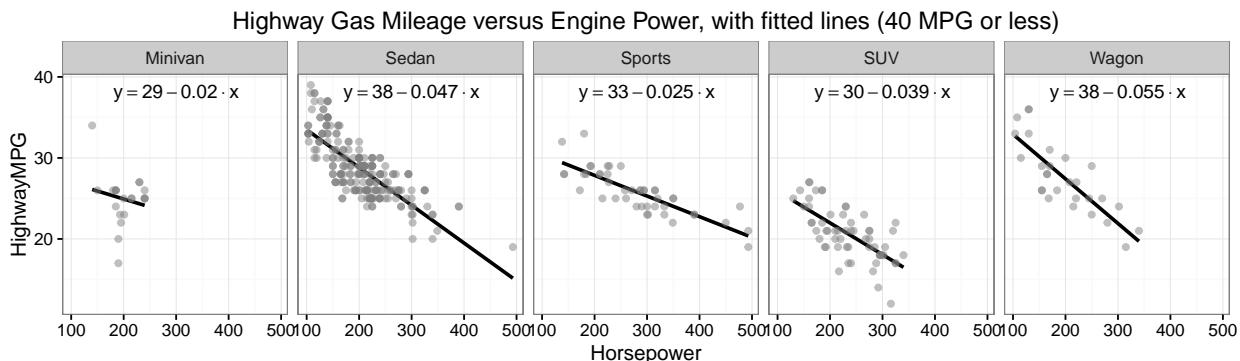


Figure 4.9: A model for the car-mileage data involving an interaction between class and horsepower. Here we've focused only on cars whose gas mileage is less than 40 miles per gallon. For this subset of the data, linearity looks like a reasonable, if imperfect, assumption.

## Dependence among predictors

IN THIS section, we'll discuss the issue of how to interpret an analysis of variance for a model where the predictors themselves are correlated with each other. (Another term for correlation among predictors is *collinearity*.) This discussion will expand upon a point raised before—but only briefly—about the importance of context in the sequential construction of an ANOVA table.

Let's briefly review the analysis of variance (ANOVA). You'll recall that, in our look at the data on reaction time in video games, we ran an ANOVA (Table 4.3) of a regression model that predicted variation in human reaction time in terms of distance, visual clutter, subject-level variation, and a distance/clutter interaction. Our goal was to apportion credit among the individual parts of the model, where “credit” was measured by each variable’s contribution to the predictable variation in the model’s variance decomposition ( $TV = PV + UV$ ). This led us, for example, to the conclusions that subject-level variation was large relative to the other effects, and that the distance/clutter interaction contributed only a modest amount to the predictive abilities of the model.

We can also run an analysis of variance on models containing numerical predictors. To see this in action, let's revisit the data on the gas mileage of cars from Figure 1.11, back on page 27. Recall that this data set involved 387 cars and three variables: gas mileage, engine horsepower, and vehicle class (minivan, sedan, sports car, SUV, or wagon). We can see this data once more in Figure 4.9, which shows a lattice plot of mileage versus horsepower, stratified by vehicle class.

In our earlier discussion of this data, we noted two facts:

- (1) The classes exhibit systematic differences in their typical mileages. For example, sedans have better gas mileage, on average, than SUVs or minivans.
- (2) Vehicle class seems to modulate the relationship between MPG and engine power. As engine power increases, mileage gets worse on average, regardless of vehicle class. But this drop-off is steeper for wagons than for sports cars.

Previously, we described these facts only informally. But we now have the right tools—dummy variables and interactions—that allow us to quantify them in the context of a regression model. Specifically: point (1) suggests that we need class-level dummy variables, to move the intercepts up and down as appropriate for each class; while point (2) suggests that we need an interaction between class and horsepower, to make the slope of the regression line get steeper or shallower as appropriate for each class. Using our informal notation from earlier, our regression model should look like this:

$$\text{MPG} \sim \text{Horsepower} + \text{Class} + \text{Class: Horsepower}.$$

Upon fitting this model by least squares, we get the coefficients in Table 4.4, at right. The corresponding fitted lines within each class are also shown in Figure 4.9. The parameters of this fitted model confirm our earlier informal observations based on the lattice plot: that both the average mileage and the steepness of the mileage/horsepower relationship are affected by vehicle class.

An analysis of variance table for this model looks like this.

Variable added	# Pars	$R^2$	$\Delta R^2$	$s_e$	$\Delta s_e$
Intercept only	1	0		4.59	
Horsepower	1	0.426	0.426	3.48	1.11
Class	4	0.725	0.299	2.42	1.06
Horsepower:Class	4	0.743	0.018	2.36	0.07

According to this table, we can attribute most of the credit for predicting fuel economy to the horsepower variable ( $\Delta R^2 = 0.426$ ). Most of the remaining credit goes to vehicle class ( $\Delta R^2 = 0.299$ ). The interaction produces a modest change in  $R^2$ ; this bears out the

Table 4.4: Fitted coefficients (rounded to the nearest hundredth) for the model that predicts car gas mileage in terms of engine horsepower, vehicle class, and a class/horsepower interaction.

Variable	$\hat{\beta}$
Intercept	28.86
Horsepower	-0.02
Sedan	9.28
Sports	4.08
SUV	0.94
Wagon	9.55
Horsepower:Sedan	-0.03
Horsepower:Sports	-0.01
Horsepower:SUV	-0.02
Horsepower:Wagon	-0.04

Table 4.5: An analysis of variance (ANOVA) table for the model that predicts highway gas mileage in terms of a car's engine power and vehicle class, including both main effects and an interaction term. In this ANOVA table, the horsepower variable has been added first, followed by vehicle class.

visual impression conveyed by Figure 4.9, in which the slopes in each panel are clearly different, but not dramatically so.

But this conclusion about the relative importance of horsepower and vehicle class involves a major, even deal-breaking, caveat. Remember that an analysis of variance is inherently sequential: first we add the horsepower variable, then we add vehicle class, and then we add the interaction, tracking the variance decomposition at each stage. What happens if we build an ANOVA table by adding vehicle class before we add horsepower?

Variable added	# Pars	$R^2$	$\Delta R^2$	$s_e$	$\Delta s_e$
Intercept only	1	0		4.59	
Class	4	0.397	0.397	3.58	1.01
Horsepower	1	0.725	0.328	2.42	1.16
Class:Horsepower	4	0.743	0.018	2.36	0.07

Now we reach the opposite conclusion: that vehicle class contributes more ( $\Delta R^2 = .397$ ) to the predictable variation than does horsepower ( $\Delta R^2 = .328$ ). Why does this happen? How could our conclusion about the relative importance of the variables depend upon something so arbitrary as the order in which we decide to add them?

#### *Shared versus unique information*

Figure 4.10 provides some intuition why this is so. In our data on gas mileage, the two predictors (horsepower and vehicle class) are correlated with each other: vehicles in certain classes, like SUVs and sports cars, have more powerful engines on average than sedans, wagons, and minivans.

To understand why this correlation between predictors would matter so much in an analysis of variance, let's consider the information provided by each variable. First, a vehicle's class tells us at least two important things relevant for predicting gas mileage.

- 1) *Weight*: for example, SUVs tend to be heavier than sedans, and heavier vehicles will get poorer gas mileage.
- 2) *Aerodynamics*: for example, minivans tend to be boxier than sports cars, and boxier cars will get poorer gas mileage due to increased drag at highway speeds.

Similarly, the horsepower of a vehicle's engine also tells us at

Table 4.6: A second analysis of variance (ANOVA) table for the model that predicts highway gas mileage in terms of a car's engine power and vehicle class, including both main effects and an interaction term. In this ANOVA table, vehicle class has been added first, followed by horsepower.

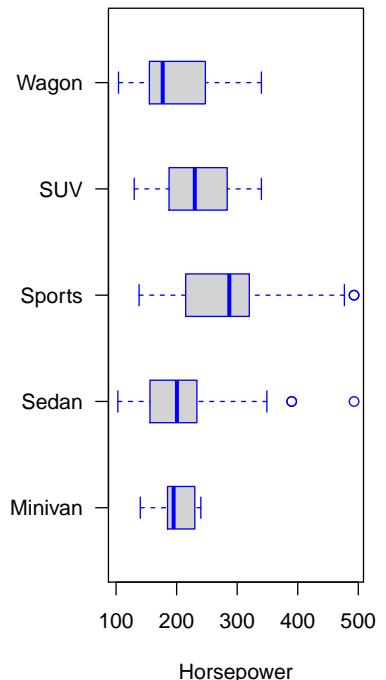


Figure 4.10: Correlation between vehicle class and horsepower.

least two important things relevant for predicting gas mileage.

- 1) *Weight*: more powerful engines are themselves heavier, and tend to come in cars that are heavier in other ways, too.
- 2) *Fuel consumption*: a smaller engine consumes less fuel and typically has better mileage than a bigger engine.

Notice that both variables provide information about a vehicle's weight; let's call this the shared information. But each also provides information on something else specific to that variable; let's call this the unique information. The shared information between the predictors manifests itself as correlation: bigger cars tend to have both bigger engines, and they also tend to be in certain classes. We can use a Venn diagram to represent both the shared and the unique information provided by the predictors in a stylized (i.e. non-mathematical) way:

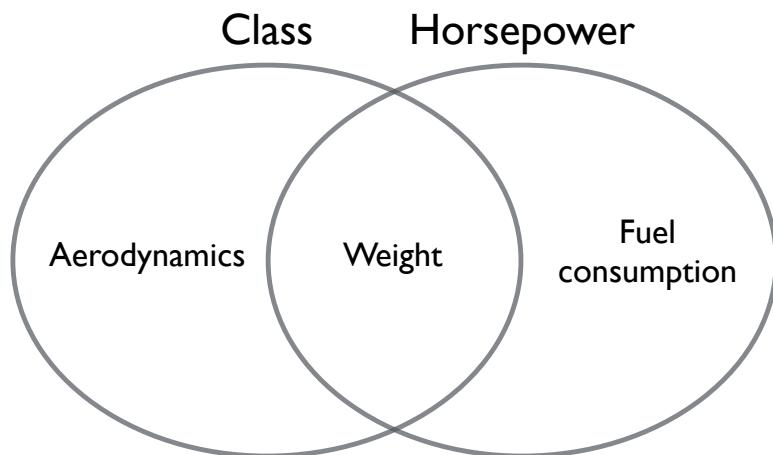


Figure 4.11: The two predictors in the gas-mileage data set provide some information content that is shared between them, in addition to some information that is unique to each one.

In the first analysis of variance (Table 4.5), we added horsepower first. When we did so, the regression model greedily used all the information it could from this predictor, including both the “shared” and “unique” information. As a result, when we added the class variable second, the shared information is redundant—it was already accounted for by the model. We therefore end up giving the class variable credit only for its unique information content; all the information content it shares with horsepower was already counted in step 1. This is illustrated in Figure 4.12.

But when we flip things around and add vehicle class to the

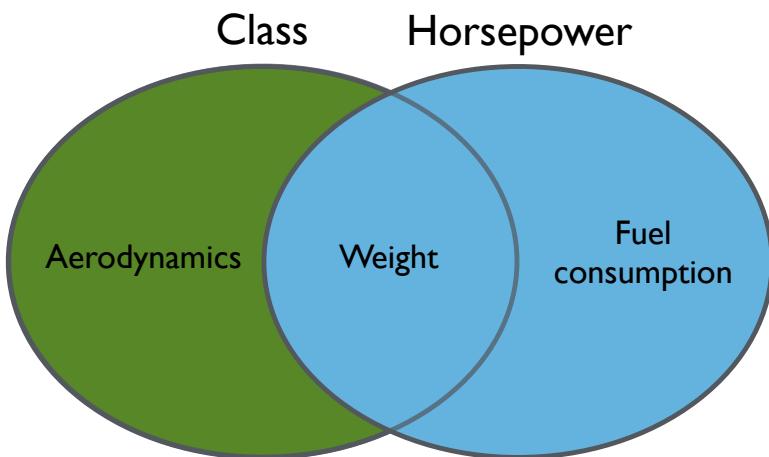


Figure 4.12: Our model for gas mileage includes two variables: engine horsepower and vehicle class. These variables both convey information about a vehicle's size, in addition to some unique information (e.g. class tells us about aerodynamics, while horsepower tells us about fuel consumption). When we add the Horsepower variable first in an analysis of variance (Table 4.5), we attribute all of the shared information content to Horsepower, and none to Vehicle class, in our ANOVA table.

model first (Table 4.6), this picture changes. We end up giving the class variable credit both for its unique information content *and* for the information it shares with Horsepower. This leaves less overall credit for Horsepower when we add it in step 2 of the ANOVA. This is illustrated in Figure 4.13.

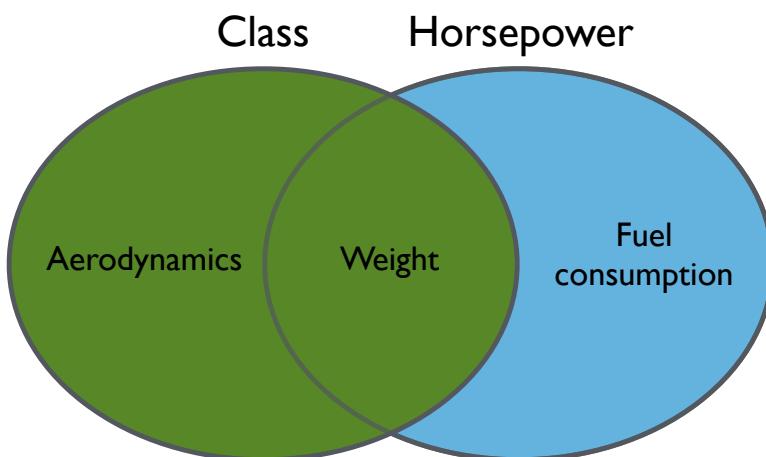


Figure 4.13: (Continued from Figure 4.12.) But when we add the Class variable first in an analysis of variance (Table 4.5), we attribute all of the shared information content to Class, and none to Horsepower, in our ANOVA table.

This example highlights an unsatisfying but true feature of the analysis of variance: when the variables are correlated, *their ordering matters* when you build the ANOVA table.

This feature of an ANOVA table at first seems counterintuitive, even disturbing. Yet similar phenomena occur all the time in everyday life. A good analogy here is the dessert buffet at Thanksgiving dinner. Imagine two different versions of dessert.

*Version 1:* After dinner, your aunt offers you apple pie, and you eat your fill. The apple pie is delicious—you were really looking forward to something sweet after a big Thanksgiving meal. It makes you very happy.

Next, after you've eaten your fill of apple pie, your aunt offers you pumpkin pie. Pumpkin pie is also delicious—you love it just as much as apple. But your dessert tummy is pretty full already. You eat a few bites, and you enjoy it; that spicy pumpkin flavor is a little different to what you get from an apple pie. But of course, pumpkin pie is still a dessert, and you don't enjoy it as much as you might have if you hadn't eaten so much apple pie first.

*Version 2:* After dinner, your aunt offers you pumpkin pie, and you eat your fill. The pumpkin pie is delicious—all that whipped cream on top goes so well with the nutmeg and earthy pumpkin flavor. It makes you very happy.

Next, after you've eaten your fill of pumpkin pie, your aunt offers you apple pie. Apple pie is also delicious—you love it just as much as pumpkin. But your dessert tummy is pretty full already. You eat a few bites, and you enjoy it; those tart apples with all the cloves and cinnamon give a little different flavor to what you get from a pumpkin pie. But apple pie is still a dessert, and you don't enjoy it as much as you might have if you hadn't eaten so much pumpkin pie first.

That evening, which pie are you going to remember? In version 1, you'll attribute most of your Thanksgiving dessert afterglow to the apple pie; while in version 2, you'll attribute most of it to pumpkin pie. *Context matters*, even if in the abstract you like both pies the same amount.

An analysis of variance is like the one-at-a-time dessert eater at Thanksgiving. Whatever variable we add to the model first, the model greedily eats its fill of that, before turning to the second variable. This affects how credit gets attributed. In our ANOVA tables for the gas mileage data, our two variables (horsepower and vehicle class) are like apple and pumpkin pie. Yes, they each offer

something unique, but they also share a lot of their information content (just like the pies are both desserts). Because of this, the order in which they are added to the ANOVA table—or equivalently, the context in which each variable's marginal contribution to the model is evaluated—matters a lot.

The moral of the story is that it rarely makes sense to speak of “the” ANOVA table for a model—only “an” ANOVA table. Thus there is no unique way to partition credit among multiple variables for their shared information content in a regression model. We must make an arbitrary choice, and in an ANOVA table, that choice is “winner take all” to the first variable added to the model.

*Final thoughts on ANOVA.* There are two further points to bear in mind about the analysis of variance. First, the ANOVA table is not the model itself, only an attempt to partition credit for predicting the outcome among the variables in the model by adding those variables one at a time. And while the ANOVA table is order-dependent, the model itself isn't. Regardless of the order in which you add variables, you will always get the same model coefficients, fitted values, and residuals at the end.

Second, we've discussed the subtleties of interpreting an ANOVA table in the presence of correlation among the predictors. However, if the variables in the model are independent of one another, then they have no shared information content, and the ANOVA table does not depend upon the ordering of the variables.

This is why we ignored the issue of variable ordering when building an ANOVA table for our model of reaction time in video games versus distance, clutter, and subject-level effects. For that data set, the predictor variables were independent with each other: the experimental design was perfectly balanced, with each subject sitting for exactly 40 trials for each pairwise combination of the cluttered and distance variables. Regardless of the order in which we add the variables, we will always get the same  $\Delta PV$  for each one. Thus in the absence of dependence among the predictors, we can uniquely assign credit for predicting the outcome to each one.<sup>7</sup>

Regression models, just like Thanksgiving guests, thrive on variety—that is, on multiple independent sources of information.

<sup>7</sup> For this reason, ANOVA is a commonly used tool in the analysis of designed experiments, when we can ensure that the predictors are independent of one another. It is less common in the analysis of observational studies, where the inevitable presence of collinearity significantly weakens the conclusions that we can draw from an ANOVA.



# 5

## *Quantifying uncertainty using the bootstrap*

### **Quantifying parameter uncertainty**

IN COMING this far through the book, you've already learned many valuable skills: how to summarize evidence both graphically and numerically; how to fit basic group-wise and linear statistical models to data; how to combine grouping and numerical variables; and how to use these models to explore trends and predict new outcomes.

But we're missing a crucial piece of the puzzle. Earlier we defined statistical modeling as the structured quantification of uncertainty. We've focused a lot so far on the "structure" part; now we'll begin to focus on the "uncertainty" part.

A question that almost always arises in statistical modeling is: how confident are we in our estimate of an effect size? Take the following study of a new therapeutic regime for esophageal cancer, from the New England Journal of Medicine in 2006:

We randomly assigned patients with resectable adenocarcinoma of the stomach, esophagogastric junction, or lower esophagus to either perioperative chemotherapy and surgery (250 patients) or surgery alone (253 patients). . . . With a median follow-up of four years, 149 patients in the perioperative-chemotherapy group and 170 in the surgery group had died. As compared with the surgery group, the perioperative-chemotherapy group had a higher likelihood of overall survival (five-year survival rate, 36 percent vs. 23 percent).<sup>1</sup>

Thus the chemotherapy regime appears to save lives: the relative risk of survival under chemo is 36/23, or about 1.6. But 1.6 plus-or-minus what? What if the physicians running the trial had enrolled a different sample of patients? Might the relative risk have looked more like 1.3 (a smaller effect) or even 1.0 (no effect)? Chemotherapy has nasty side effects and is very expensive. If you're a cancer patient or a Medicare administrator, uncertainty about the effect size matters.

<sup>1</sup> Cunningham, et. al. "Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer." *New England Journal of Medicine*, 2006 July 6; 355(1):11-20.

We use the phrase *statistical inference* to describe the framework and procedures we use to address uncertainty in statistical models. In this chapter, we'll approach statistical inference using a technique called the bootstrap.

## Sampling distributions, estimators, and alternate universes

In fitting statistical models, we typically equate the trustworthiness of a procedure with its stability under the influence of luck, and we seek to measure the degree to which that procedure might have given a different answer if the forces of randomness had made the world look a bit different. Specifically, the question we seek to answer is: “if our data set had been different merely due to chance, would our answer have been different, too?”

$$\text{Confidence in your estimates} \iff \text{Stability of those estimates under the influence of chance}$$

You can see why it makes sense to equate stability with trustworthiness if you imagine a suspect who gives the police three different answers to the question, “Where were you last Tuesday night?” If the story keeps changing, there is little basis for trust.

*Sources of instability.* One obvious source of instability in our estimates is when our observations are subject to random forces. For example, suppose we wish to characterize the relationship between SAT score and graduating GPA for the entering class of 2000 at the University of Texas. Figure 5.1 shows the entire relevant population, yet there is still randomness to worry about—for, as the teacher in Ecclesiastes puts it, “time and chance happeneth to them all.” If any of these 5,191 students had taken the SAT on a different day, or eaten a healthier breakfast on the day of their chemistry finals, we would be looking at a slightly different data set, and thus a slightly different least-squares line—even if the underlying SAT–GPA relationship had stayed the same.

Another source of instability is the effect of sampling variability, which arises when we’re unable to study the entire population of interest. The key insight here is that a different sample would have led to different estimates of the model parameters. Consider the example above, about the study of a new chemotherapy regime for esophageal cancer. If doctors had taken a different sample of

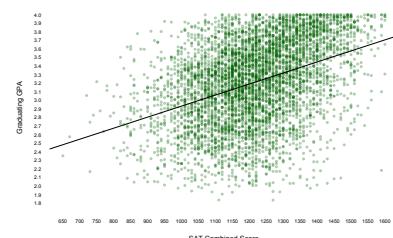
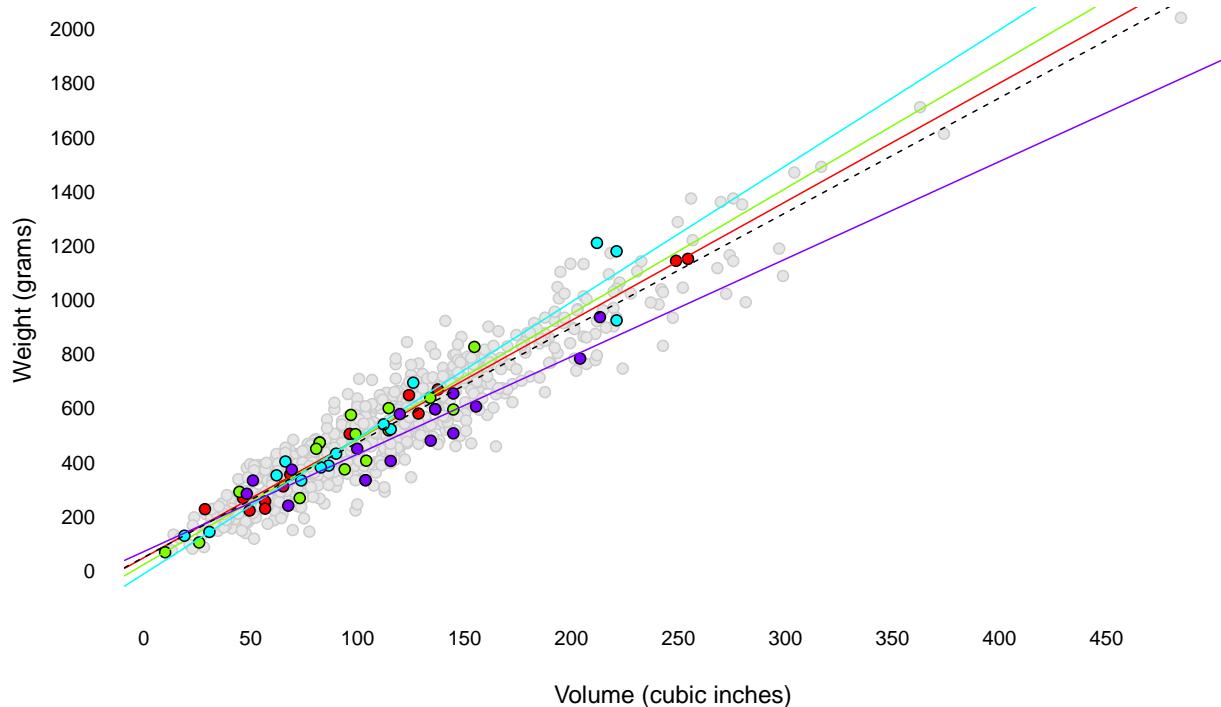


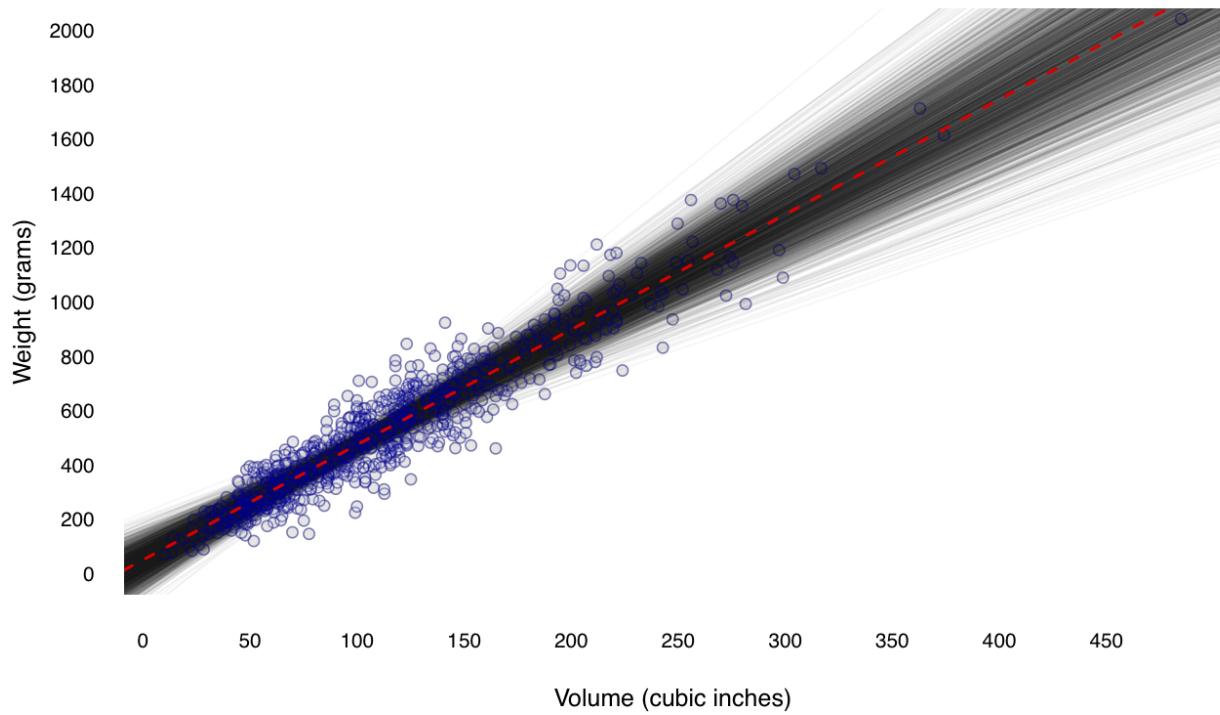
Figure 5.1: Graduating GPA versus high-school SAT score for all students who entered UT–Austin in the fall of 2000 and went on to earn a bachelor’s degree within 6 years. The black line shows the least-squares fit.



503 cancer patients and gotten a drastically different estimate of the new treatment's effect, then the original estimate isn't very trustworthy. If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for *this particular* subset of 503 is likely to be accurate.

*An example: simulating a sampling distribution by Monte Carlo.* To get some intuition for this way of thinking, imagine that you go on a four-day fishing trip to a lovely small lake out the woods. The lake is home to a population of 800 fish of varying size and weight, depicted in Figure 5.2. On each day, you take a random sample from this population—that is, you catch (and subsequently release) 15 fish, recording the weight of each one, along with its length, height, and width (which multiply together to give a rough estimate of volume). You then use the day's catch to compute a different estimate of the volume–weight relationship for the entire population of fish in the lake. These four different days—and the four different least-squares fits—show up in different colors in Figure 5.2.

Figure 5.2: Four different days of fishing, coded by color, on an imaginary lake home to a population of 800 fish. On each day's fishing trip, you catch 15 fish, and end up estimating a slightly different weight–volume relationship. The dashed black line is the true relationship for the entire population.

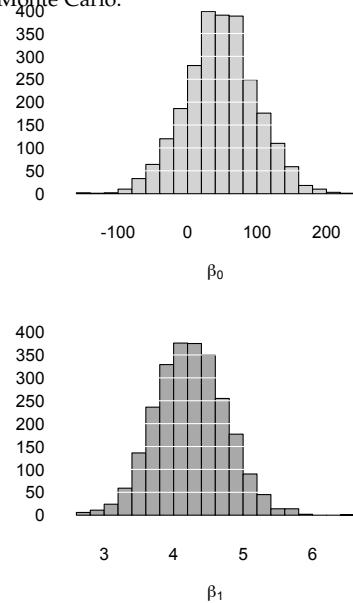


Four days of fishing give us some idea of how the estimates for  $\beta_0$  and  $\beta_1$  vary from sample to sample. But 2500 days of fishing, simulated by computer, give us a better idea. Figure 5.3 shows just this: 2500 different samples of size 15 from the population, together with 2500 different least-squares estimates of the weight-volume relationship. This is an example of a *Monte Carlo simulation*, in which we run a computer program to repeatedly simulate a random process (in this case, sampling from a population).

These pictures show the *sampling distribution* of the least-squares line—that is, how the estimates for  $\beta_0$  and  $\beta_1$  change from sample to sample, shown in histograms in the right margin. In theory, to know the sampling distributions exactly, we'd need to take an infinite number of samples, but 2500 gives us a rough idea.

*The sampling distribution.* To understand the concept of a sampling distribution, it helps to distinguish between an *estimator* and an *estimate*. A good analogy here is that an estimator is to a court trial as an estimate is to a verdict. Just like a trial is a procedure for reaching a verdict about guilt or innocence, an estimator is

Figure 5.3: 2500 days of fishing, together with the 2500 different estimates of  $\beta_0$  and  $\beta_1$  (below), simulated by Monte Carlo.



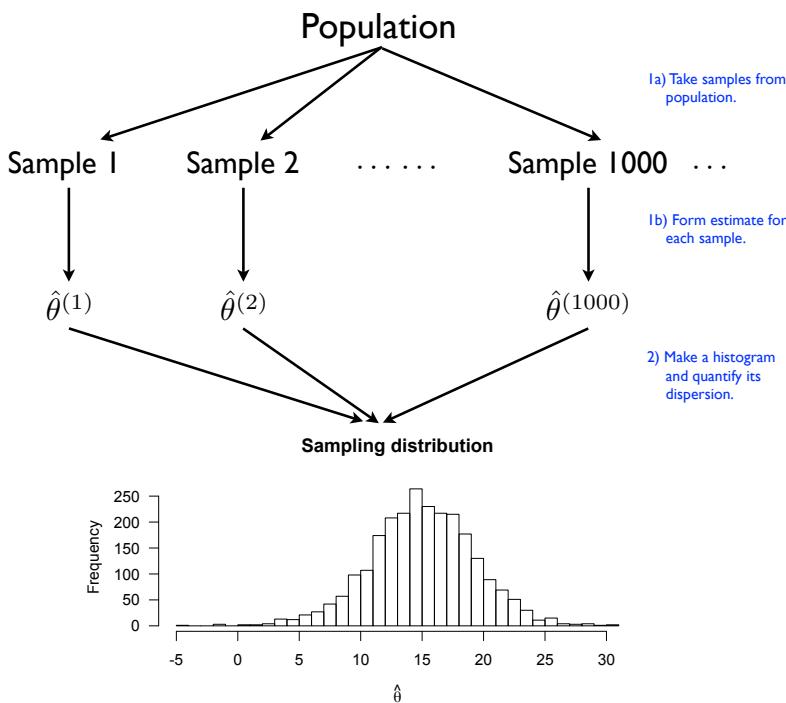


Figure 5.4: A stylized depiction of a sampling distribution of an estimator  $\hat{\theta}$ . To construct this distribution, we must imagine the following thought experiment. We repeatedly take many samples (say, 1000) from the population (step 1a). For each sample, we apply our estimator to compute the estimate  $\hat{\theta}^{(r)}$  (step 1b). At the end, we combine all the estimates  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(1000)}$  into a histogram, and we summarize the dispersion of that histogram (step 2). Technically, the sampling distribution is the distribution of estimates we'd get with an infinite number of samples, and the histogram is an approximation of this distribution. The difference between the true distribution and the approximation generated by Monte Carlo is called *Monte Carlo error*.

a procedure for reaching an estimate of some population-level quantity on the basis of a sample. The least-squares procedure is a specific set of steps (i.e. equations) that one applies to a data set. The procedure yields estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for the slope and intercept of a population-wide linear trend; while the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  you get for a specific data set are the estimates. An estimator's sampling distribution is the distribution of results (that is, the estimates) that one obtains from that estimator under repeated sampling from a population. Figure 5.4 shows graphically how, in principle, this distribution is constructed. Concrete examples of an estimator include the sample mean, the least squares procedure, and the residual standard deviation.

Good estimators are those that usually yield estimates close to the truth, with minimal variation. Therefore, we typically summarize a sampling distribution using its standard deviation, which we refer to as the *standard error*.<sup>2</sup> In quoting the standard error of an estimator's sampling distribution, you are saying: "If I were to take repeated samples from the population and use this estimator for every sample, my estimate is typically off from the truth by

<sup>2</sup> We are also sometimes interested in the mean of a sampling distribution. If the mean of an estimator's sampling distribution is equal to the true population value, we say that the estimator is *unbiased*. This term has a precise mathematical meaning, but also an unwarranted connotation of universal desirability that many statisticians find problematic. Alas, for historical reasons, we're basically stuck with the term. It turns out that unbiasedness is not always a good property of an estimator. There can be very good reasons to use estimators that we know to be biased. But that's for another book.

about this much.” Notice again that this is a claim about a procedure, not a particular estimate. The bigger the standard error, the less stable the estimator across different samples, and the less you can trust the estimate for any particular sample. To give a specific example, for the 2500 samples in Figure 5.3, the standard error of  $\hat{\beta}_0$  is about 50, while the standard error of  $\hat{\beta}_1$  is about 0.5.

Of course, if you really could take repeated samples from the population, life would be easy. You could simply peer into all of those alternate universes, tap each version of yourself on the shoulder, and ask, “What slope and intercept did you get for *your* sample?” By tallying up these estimates and seeing how much they differed from one another, you could discover precisely how much confidence you should place in your own estimates of  $\beta_0$  and  $\beta_1$ , and report appropriate error bars based on the standard error of your estimator.<sup>3</sup>

Most of the time, however, we’re stuck with one sample, and one version of reality. We cannot know the actual sampling distribution of our estimator, for the same reason that we cannot peer into all those other lives we might have lived, but didn’t:

Two roads diverged in a yellow wood,  
And sorry I could not travel both  
And be one traveler, long I stood  
And looked down one as far as I could  
To where it bent in the undergrowth. . . .<sup>4</sup>

<sup>3</sup> Let’s ignore the obvious fact that, if you had access to all those alternate universes, you’d also have more data. The presence of sample-to-sample variability is the important thing to focus on here.

<sup>4</sup> Robert Frost, *The Road Not Taken*, 1916.

Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.

Surprisingly, we can come close to performing the impossible. There are two ways of feasibly constructing something like the histogram in Figure 5.4, thereby approximating an estimator’s sampling distribution without ever taking repeated samples from the population.

- 1) *Resampling*: that is, by pretending that the sample itself is the population, which allows one to approximate the effect of sampling variability by resampling from the sample.
- 2) *Parametric probability modeling*: that is, by assuming that the forces of randomness obey certain mathematical regularities, and by drawing conclusions about these regularities using probability theory.

In this chapter, we’ll discuss the resampling approach, deferring the probability-modeling approach to a later chapter.

## Bootstrapping: standard errors through resampling

AT THE core of the resampling approach to statistical inference lies a simple idea. Most of the time, we can't feasibly take repeated samples of size  $n$  from the population, to see how our estimate changes from one sample to the next. But we can repeatedly take samples of size  $n$  from the sample itself, and apply our estimator afresh to each notional sample. The idea is that the variability of the estimates across all these samples can be used to approximate our estimator's true sampling distribution.

This process—pretending that our sample is the whole population, and taking repeated samples of size  $n$  with replacement from our original sample of size  $n$ —is called *bootstrap resampling*, or just *bootstrapping*.<sup>5</sup> Each block of  $n$  resampled data points is called a bootstrapped sample. To bootstrap, we write a computer program that repeatedly resamples our original sample and recomputes our estimate for each bootstrapped sample. Modern software makes a non-issue of the calculational tedium involved.

You may be puzzled by something here. There are  $n$  data points in the original sample. If we repeatedly resample  $n$  data points from our “pseudo-population” of size  $n$ , won’t each bootstrapped sample be identical to the original sample? If so, and every bootstrapped sample looks the same, then how can this process be used to simulate sampling variability?

This fact highlights a key requirement of bootstrapping: the resampling must be done *with replacement* from the original sample, so that each bootstrapped sample contains duplicates and omissions from the original sample.<sup>6</sup> These duplicates and omissions induce variation from one bootstrapped sample to the next, mimicking the variation you’d expect to see across the real repeated samples that you can’t take.

To summarize, let’s say we have a data set  $D$ , consisting of  $n$  cases. We want to understand how our estimator  $\hat{\theta}$  might have behaved differently with a different sample of size  $n$ . To answer this question using bootstrapping, we follow two main steps.

(1) Repeat the following substeps many times (e.g. 1000 or more):

- a. Generate a new bootstrapped sample  $D^{(r)}$  by taking  $n$  samples with replacement from  $D$ .
- b. Apply the estimator  $\hat{\theta}$  to the bootstrapped sample  $D^{(r)}$  and save the resulting estimate,  $\hat{\theta}^{(r)}$ .

<sup>5</sup> The term “bootstrapping” is a metaphor. It is an old-fashioned phrase that means performing a complex task starting from very limited resources. Imagine trying to climb over a tall fence. If you don’t have a rope, just “pull yourself up by your own bootstraps.”

<sup>6</sup> Imagine a lottery drawing, where there’s a big urn with 60 numbered balls in it. We want to choose a random sample of 6 numbers from the urn. After we choose a ball, we could do one of two things: 1) put the ball to the side, or 2) record the number on the ball and then throw it back into the urn. If you set the ball aside, it can be selected only once; this is sampling without replacement, and it’s what happens in a real lottery. But if instead you put the ball back into the urn, it has a chance of being selected more than once in the final sample; this is sampling with replacement, and it’s what we do when we bootstrap.

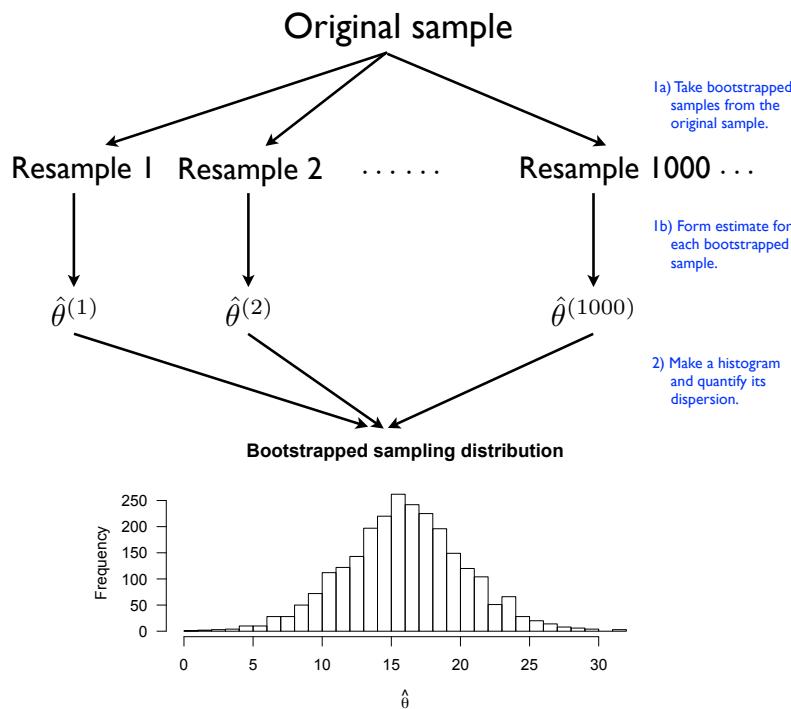


Figure 5.5: A stylized depiction of a bootstrapped sampling distribution of an estimator  $\hat{\theta}$ . We have a single original sample. We repeatedly take many bootstrapped samples (say, 1000) from the original sample (step 1a). For each resample, we compute the estimator  $\hat{\theta}$  (step 1b). At the end, we combine all the estimates  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(1000)}$  into a histogram of the bootstrapped sampling distribution, and we summarize the dispersion of that histogram (step 2). Compare with Figure 5.4.

- (2) Take all of the  $\hat{\theta}^{(r)}$ 's you've generated and make a histogram.  
This is your estimate of the sampling distribution.

See Figure 5.5, and compare with Figure 5.4.

Resampling won't yield the true sampling distribution of an estimator, but it is often good enough for approximating the standard error (which you'll remember is just the standard deviation of the sampling distribution). We use the term *bootstrapped standard error* for the standard deviation of the bootstrapped sampling distribution. The bootstrapped standard error is an estimate of the true standard error.

The quality of this estimate depends almost entirely on one thing: how closely the original sample resembles the wider population. This is a question of judgment best answered by someone with subject-area expertise relevant to the data set at hand. As a data analyst this often isn't under your control, and therefore it's almost worth remembering that the bootstrap is not entirely free of assumptions. You can't magic your way to sensible estimates of the true sampling distribution by bootstrapping a biased, woefully small, or otherwise poor sample.

The quality of the Monte Carlo approximation also depends to a lesser extent on how many bootstrapped samples you take from the original sample. Simulating more bootstrapped samples help to reduce the variability inherent in any Monte Carlo simulation—up to a point. But taking more bootstrapped samples is never a substitute for having more actual samples in the real data set. Fundamentally, it is the size of your original sample that governs the precision of your estimates.

A natural question is: how well does bootstrapping work in practice? To see the procedure in action, let's reconsider the least-squares estimator of the slope ( $\beta_1$ ) for the weight–volume line describing the fish in our hypothetical lake. The top row of Figure 5.6 shows three actual sampling distributions, corresponding to samples of size  $n = 15$ ,  $n = 50$ , and  $n = 100$  from the entire population. These were constructed using the Monte Carlo method described several pages ago, as depicted in Figures 5.3 and 5.4. For example, the top left panel (for  $n = 15$ ) was constructed by taking 2,500 Monte Carlo samples from the true population in Figure 5.3, and computing the least-squares estimate of the slope for each sample as in Figure 5.4.

Below each true sampling distribution, we have focused on four of these 2500 samples. For each of these real samples, we ran the bootstrapping procedure by 2500 bootstrapped samples from the original sample of size  $n$ , treating it as a pseudo-population. For each bootstrapped sample, we compute the least-squares line for weight versus volume. These 2500 estimates of  $\beta_1$  are what you see in each grey-colored panel of Figure 5.6. For example, the first grey panel in column 1 corresponds to the bootstrapped sampling distribution from the first sample of size 15; the second grey panel corresponds to the bootstrapped sampling distribution from the second sample of size 15; and so on for the rest of the grey panels.

If bootstrapping were perfect, each grey panel would look exactly like the corresponding orange panel above, regardless of the same size. But of course, bootstrapping isn't perfect. If you study these pictures closely, you'll notice a few things.

- (1) The bootstrapped sampling distribution can differ substantially from one original sample to the next (top to bottom). The sample-to-sample differences are larger when the original sample size is small.
- (2) The bootstrapped sampling distribution gets both closer to the truth, and less variable from one original sample

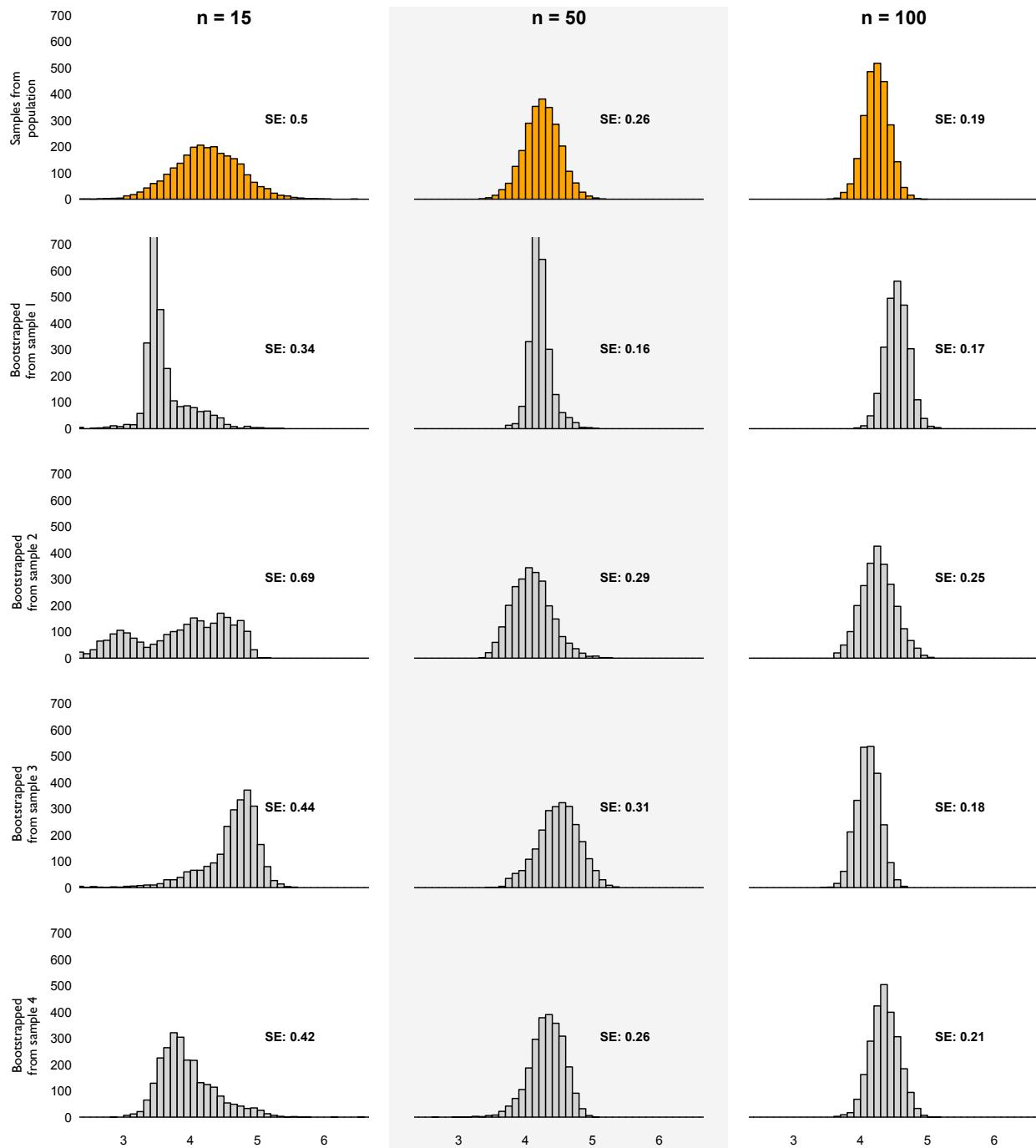


Figure 5.6: Actual (top, in orange) and bootstrapped sampling distributions (four replications) for the least-squares estimator of  $\beta_1$  from Figure 5.2.

- to the next, as the original sample size gets larger.
- (3) The bootstrapped standard errors (printed next to each histogram) are often closer to the true standard error than you might naïvely expect, based on the visual correspondence of the bootstrapped sampling distribution to the true one.

## Confidence intervals and coverage

Now that we've learned to approximate an estimator's sampling distribution via bootstrapping, what do we do with this information? The answer is: we quantify the uncertainty of our estimate via a *confidence interval*: a range of plausible values for the true value of a parameter, together with an associated *confidence level* between 0% and 100%. The width of a confidence interval conveys the precision with which the data have allowed you to estimate the underlying population parameter. If your interval actually contains the true population value, we say that the interval *covers* the truth. If it doesn't, the interval *fails to cover* the truth. In real life, you won't know whether your interval covers. The confidence level expresses how confident you are that it actually does.

There are many ways of generating confidence intervals from bootstrapped sampling distributions, ranging from the simple to the highly sophisticated (and mathematically daunting). We'll focus on two simple ways here, with the understanding that the more technical ways we don't discuss are a bit more accurate.<sup>7</sup>

First, there's the basic standard-error method. Here, you quote a symmetric error bar centered on the estimate from the original sample, plus-or-minus some multiple  $k$  of the bootstrapped standard error. To be precise, let's say that  $\theta$  is some population parameter you're trying to estimate; that  $\hat{\theta}$  is the estimate of  $\theta$  generated by your actual sample; and that you've run the bootstrapping procedure on your sample and found that the bootstrapped standard error is  $\hat{\sigma}$ . Your confidence interval would then be

$$\theta \in \hat{\theta} \pm t^* \hat{\sigma},$$

where  $t^*$  is a chosen multiple. This number  $t^*$  is called the *critical value*. It is the number of standard errors you must go out from the center to capture a certain percentage of the sampling distribution. Typical values are  $t^* = 1$  (for an approximate 68% confidence interval) and  $t^* = 2$  (for an approximate 95% confidence interval).

<sup>7</sup> If you want to get an introduction to the more technical ways of getting confidence intervals from the bootstrap, see the following article: "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians." James Carpenter and John Bithell. *Statistics in Medicine* 2000; 19:1141–64.

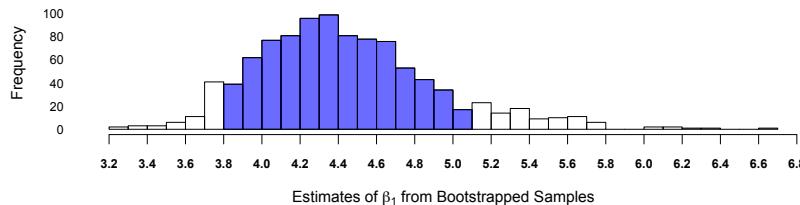


Figure 5.7: The estimated sampling distribution of  $\hat{\beta}_1$  that arises from bootstrapping one sample of size 30 from the full fish population. The blue area reflects an 80% confidence interval generated by the coverage method, with symmetric tail areas of 10% above and 10% below the blue area.

The answer to the question of *why*  $t^* = 1$  corresponds to 68% and  $t^* = 2$  to 95% is beyond the scope of this chapter. It has to do with the normal distribution and something called the central limit theorem. For now, it is fine if you accept this is an empirical rule of thumb that statisticians have found gives a good approximation in situations where your bootstrapped sampling distribution looks approximately bell-shaped. Some of the more sophisticated bootstrap techniques, mentioned in Footnote 7, are focused on improving the choice of  $t^*$  given by these simple guidelines.

Second, there's the coverage-interval method, in which you simply calculate a coverage interval using the quantiles of your bootstrapped sampling distribution. For example, Figure 5.7 shows the bootstrapped sampling distribution for the slope of the weight-volume relationship arising from a single sample of 30 fish from the same lake as before. If you wanted to compute an 80% confidence interval based on this data, you would calculate the 10th and 90th percentiles of this histogram, giving you an interval that contains 80% of the bootstrapped estimates of the slope. In Figure 5.7, this interval is (3.8, 5.1), shown in blue. This example highlights that, unlike the intervals generated by the standard-error method, the intervals generated by the coverage method need not be symmetric about the estimate  $\hat{\theta}$  derived from your actual sample.

Is one of these two methods better? Not as a general rule. The coverage-interval approach is more common in practice, and it's a fine default option. The most conservative thing to do, assuming you don't want to go the very technical<sup>8</sup> route, is to compute both and report the wider interval.

*What does “confidence” mean?*

The word “confidence,” as it is used in the phrase “confidence interval,” has a notoriously tricky interpretation. To put it con-

<sup>8</sup> See Footnote 7.

cisely but opaqely, confidence intervals are intervals generated by a method that satisfies the frequentist coverage principle.

*The frequentist coverage principle:* If you were to analyze one data set after another for the rest of your life, and you were to quote X% confidence intervals for every estimate you made, those intervals should cover their corresponding true values at least X% of the time. Here X can be any number between 0 and 100.

Let's unpack this a bit. Imagine that your interval was generated with a procedure that, under repeated use on one sample after the next, tends to yield intervals that cover the true value with a relative frequency of at least 80%. Then, and only then, may you claim a bona fide 80% confidence level for your specific interval. (You may, of course, aim for whatever coverage level you wish in lieu of 80%. Many people seem stuck on 95%, but it's entirely your choice.) Thus confidence intervals involve something of a bait-and-switch: they purport to answer a question about an individual interval, but instead give you information about some hypothetical assembly line that could be used to generate a whole batch of intervals. Nonetheless, there is an appealing "truth in advertising" property at play here: that if you're going to claim 80% confidence, you should be right 80% of the time over the long run.

An obvious question is: do bootstrapped confidence intervals satisfy the frequentist coverage property? If your sample is fairly representative of the population, then the answer is a qualified yes. That is, the bootstrapping procedure yields nominal X% intervals that cover the true value "approximately" X% of the time. Moreover, as the size of the original sample gets bigger, the quality of the approximation gets better. Alas, it is necessary to appeal to some very advanced probability theory to put both of these claims on firm footing. (This is best deferred to another, much more advanced book. For those that like fancy math, the relevant branch of probability theory is called empirical-process theory, which part of a wider area called stochastic processes.)

For our purposes, it is better to show the procedure in action. Figure 5.8, for example, depicts the results of running 100,000 regressions—1,000 bootstrapped samples for each of 100 different real samples from the population in Figure 5.2. The vertical black line shows the true population value of the weight–volume slope ( $\beta_1 = 4.24$ ) for our population of fish. Each row corresponds to a different actual sample of size  $n = 30$  from the population. Dots

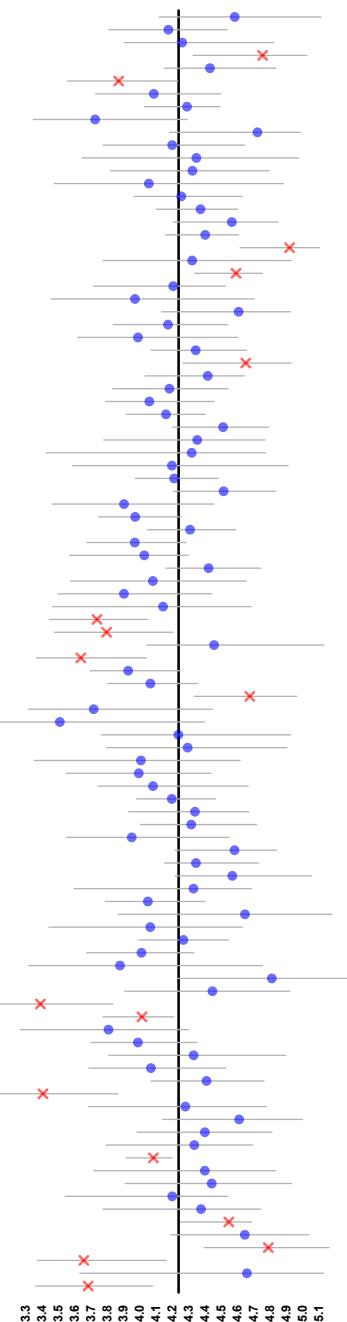


Figure 5.8: 100 different samples of size 30 from the population in Figure 5.2, along with each least-squares estimate of the weight–volume slope, and an 80% bootstrapped confidence interval, just like that at the top left. Blue dots show confidence intervals that cover; red crosses show those that don't.

and crosses indicate the least-squares estimate of the slope arising from that sample, while the grey bars show the corresponding 80% bootstrapped confidence intervals generated by the coverage method (just like the blue region in Figure 5.7).

The nominal confidence level of 80% for each individual interval must be construed as a claim about the *whole ensemble* of 100 intervals: 80% should cover, 20% shouldn't. In fact, 83 of these intervals cover and 17 don't, so the claim is approximately correct.

### *Gaussian versus bootstrapped confidence intervals*

Most statistical software packages have built-in routines for calculating standard errors and confidence intervals, and will show them as part of a routine summary output for a regression model. For example, in R, the `summary` and `confint` functions do just this.

Chances are, however, that the package you use is *not* using the bootstrap to calculate these confidence intervals. So what is it doing instead? The full answer to this question turns out to be rather long and drawn-out, and we'll return to it in a later chapter. But we can give a quick summary here.

The short answer is that your statistical software is calculating *Gaussian* standard errors and confidence intervals, which are based on the assumption that the residuals in the regression model follow a Gaussian, or normal, distribution:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i \\ e_i &\sim N(0, \sigma^2). \end{aligned} \quad (5.1)$$

The first equation is familiar: observation = fitted value + residual. But the second equation is new. It invokes an assumption that we never needed to make before: that the residuals  $e_i$  arise from a normal distribution with mean 0 and variance  $\sigma^2$ . In fact, this assumption long predicated the use of the bootstrap to calculate confidence intervals in regression modeling, and it is embedded in most statistical software today. Gaussian standard errors are sometimes numerically similar to bootstrapped standard errors, but they are not calculated in the same way.

There are three obvious questions that arise in conjunction with this assumption.

- (1) Huh? How does the assumption of normally distributed residuals let us calculate standard errors and confidence intervals?

- (2) This seems useless and kind of goofy. Why bother with this assumption? That is, under what circumstances would we use this assumption to calculate confidence intervals and standard errors, as opposed to the bootstrapping technique that we've already learned?
- (3) OK, fine. But how do we check whether the assumption of normally distributed residuals is satisfied for some particular data set?

Here are some very brief answers to these three questions.

- (1) *How does this even work?* Using probability theory, it is possible to mathematically derive formulas for standard errors and confidence intervals, based on the assumption of normally distributed residuals. The math, which exploits the nice properties of the normal distribution, isn't actually hard. But you do have to know a bit of probability theory to understand it. Moreover, the math is tedious, with lots of algebra; and it's just not that important, in the sense that it will add little to your conceptual understanding of regression. So we'll skip the math for now, and trust that our software has implemented it correctly. If you're really interested, turn to the chapter on the normal linear regression model, later in the book.
- (2) *Why bother with this assumption?* There are several possible answers here. The simplest one, and the one we'll go with for now, is that the Gaussian standard errors are often a good approximation to the bootstrapped standard errors—assuming the normality assumption is met (see point 2, above). Moreover, the Gaussian standard errors take our software a lot less time to calculate, because they don't require us to resample the data set and refit the model thousands of times. So if your data set is very large and bootstrapping would take a prohibitively long time—or even if bootstrapping is just giving you strange software bugs—then the Gaussian standard errors and confidence intervals might be your next-best option.
- (3) *How can we check the normality assumption?* Just make a histogram of your residuals. If they look like a normal distribution, then the normality assumption is probably reasonable. If they don't, then you should stick with bootstrapped standard errors if you can. For example, Figure 5.9 shows three examples of regression models, together with a histogram of the

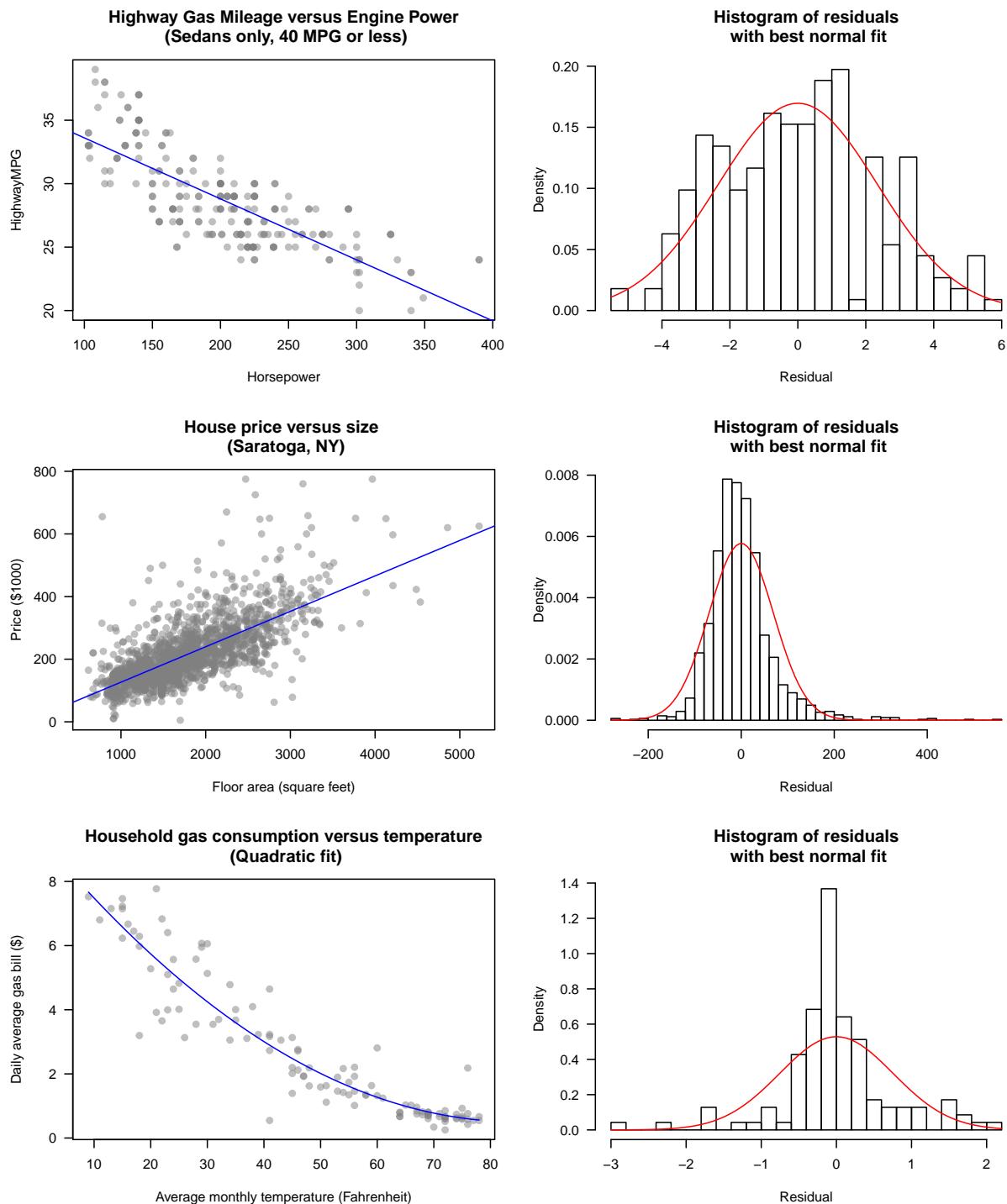


Figure 5.9: Three examples of regression models (left column), together with the best-fitting normal approximation to the histogram of each model's residuals (right column).

residuals. The top panel looks approximately normal, while the middle and bottom panels obviously don't. As a result, for the data sets in the middle and bottom panels, we can't necessarily trust the Gaussian confidence intervals; they may be a case of "garbage in, garbage out."<sup>9</sup>

We'll elaborate on these much more in a later chapter. For the time being, it's fine to think of the confidence intervals returned by regression software as just an approximation to the bootstrapped confidence intervals you've become familiar with.

<sup>9</sup> This is an oversimplification. Even if the residuals don't look Gaussian, the Gaussian confidence intervals can still be approximately correct, because of something called the central limit theorem. But this topic is for a much more advanced treatment of regression analysis.

## Bootstrapped prediction intervals (advanced topic)

Recall the problem of forecasting a future  $y^*$  corresponding to some predictor  $x^*$ , using past data as a guide. (For example, how much should a used truck with 80,000 miles cost? How much can an Austin restaurant with a food rating of 7.5 charge for a meal?) Previously, we were content to quote a prediction interval of the form

$$\hat{y}^* \in \hat{\beta}_0 + \hat{\beta}_1 x^* \pm s_e,$$

or the best guess, plus-or-minus one residual standard deviation. (We could, if we wish, also go out two residual standard deviations to get a wider interval that covered more of the data.)

These prediction intervals are good enough for most purposes. However, when we introduced them, we point that they were a bit naïve, because of how they ignore uncertainty in our estimates for  $\beta_0$  and  $\beta_1$ . For example, imagine that you work for a major metropolitan newspaper with a daily (Monday–Friday) circulation of 200,000 newspapers, and that your employer is contemplating a new weekend edition. You could certainly use the data in Figure 5.10, which correlates Sunday circulation with daily circulation for 34 major metropolitan newspapers, to inform your guess about the new Sunday edition's likely circulation. But the available data don't pin down  $\beta_0$  and  $\beta_1$  for sure; we have some uncertainty about the true values for these parameters. The kind of basic or naïve prediction interval that we've constructed until now will mask these sources of uncertainty, which may be large.

Luckily, now that we understand the logic of the bootstrap, we can try to account for this extra uncertainty. Suppose we have some value of the predictor  $x^*$ , and we want to form a prediction interval for the corresponding value of the response,  $y^*$ . The

idea is to break down our uncertainty about  $y^*$  into its constituent parts—uncertainty due to lack of perfect knowledge about the parameters, and uncertainty about the residual. The key equation is that  $y^* = \hat{y}^* + e^*$ , or future data point = point estimate + residual. We will use bootstrapping to approximate the uncertainty in each of these two terms individually.

To do so, we repeat the following steps a few thousand times.

- (1) Take a single bootstrapped sample from the original sample, and compute the least-squares estimates  $\hat{\beta}_0^{(r)}$  and  $\hat{\beta}_1^{(r)}$ . This gives you your best guess for the future  $y^*$ , given the information in the bootstrapped sample:

$$\hat{y}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x^*.$$

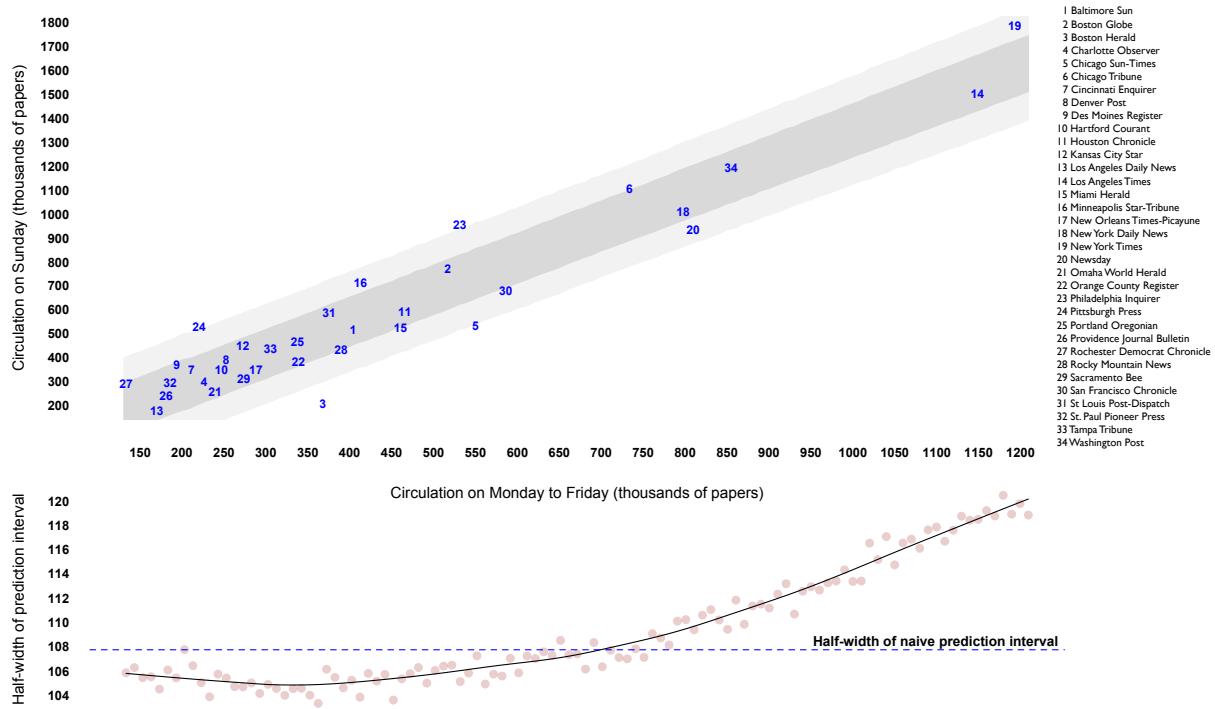
Here the superscript  $r$  denotes the  $r^{th}$  bootstrap sample.

- (2) Sample a residual  $e^{(r)}$  at random from the bootstrapped least-squares fit, to mimic the unpredictable variation in the model.
- (3) Set  $y^{(r)} = \hat{y}^{(r)} + e^{(r)}$ . This is your notional “future  $y$ ” for the  $r^{th}$  bootstrapped sample.

In step 1, we simulate the uncertainty in  $\hat{y}$  by using different parameter estimates  $\hat{\beta}_0^{(r)}$  and  $\hat{\beta}_1^{(r)}$  each time through the three-step loop. In step 2, we simulate the uncertainty in  $e$ , the future residual, by resampling the residuals from the model fit in step 1. Finally, in step 3, we combine these two sources of uncertainty to form the notional future data point,  $y^{(r)} = \hat{y}^{(r)} + e^{(r)}$ .

By repeating this process many thousands of times, we can build up a distribution of values for  $y^*$ . If you take the standard deviation of all those  $y^{(r)}$ 's, you can directly quantify the uncertainty in your prediction corresponding to  $x^*$ —for example, by quoting the dark- and light-grey prediction intervals in Figure 5.10, which stretch to one and two standard deviations (respectively) on either side of the least-squares line.

One noticeable feature of the bootstrapped prediction intervals is the way they bend outwards as they get further away from the center of the sample. This is a bit hard to see in the top panel of Figure 5.10. To show this effect more clearly, the bottom panel explicitly plots the half-width of the dark grey bootstrapped prediction intervals at 109 different hypothetical X points: every increment of 10,000 newspapers across the entire range of daily circulation, from 130,000 to 1.2 million.



The black curve shows an unmistakeable trend. Prediction uncertainty increases when you move away from the mean of  $X$ . Figure 5.3, several pages earlier, will give you some intuition for why this is so: small differences in the slope get magnified when you move further away from the middle of the sample. The naïve prediction interval fails to capture this effect entirely. On this problem, for example, the naïve interval understates prediction uncertainty by 10,000 newspapers or more for large values of  $X$ .

A final point worth noting: all of the previous warnings about bootstrapped standard errors also apply to bootstrapped prediction intervals. If the observed data is unrepresentative of the population, bootstrapping will mislead rather than inform.

Figure 5.10: Sunday circulation versus daily circulation for 34 major metropolitan newspapers, together with one- and two-standard-deviation bootstrapped prediction intervals across the range of the  $X$  variable (top panel). Also shown is the half-width of the darker-grey prediction interval across the range of  $X$  (bottom panel), versus the half-width of the naïve prediction interval, shown by the dotted blue line.

You'll notice that the pink dots marking the half-width of each bootstrapped prediction interval wiggle up and down a bit from the black curve. This happens because we only took 2,500 bootstrap samples, which produces a bit of unwanted noise. Taking more bootstrapped samples would make the pink points fall closer to the black curve, but it wouldn't shift the black curve up or down.



# 6

## *Multiple regression: the basics*

### **From lines to planes**

LINEAR regression, as we've learned, is a powerful tool for finding patterns in data. So far, we've only considered models that involve a single numerical predictor, together with as many grouping variables as we want. These grouping variables were allowed to modulate the intercept, or both the slope and intercept, of the underlying relationship between the numerical predictor (like SAT score) and the response (like GPA). This allowed us to fit different lines to different groups, all within the context of a single regression equation.

In this chapter, we learn how to build more complex models that incorporate two or more numerical predictors. For example, consider the data in Figure 6.1 on page 128, which shows the highway gas mileage versus engine displacement (in liters) and weight (in pounds) for 59 different sport-utility vehicles.<sup>1</sup> The data points in the first panel are arranged in a three-dimensional point cloud, where the three coordinates  $(x_{i1}, x_{i2}, y_i)$  for vehicle  $i$  are:

- $x_{i1}$ , engine displacement, increasing from left to right.
- $x_{i2}$ , weight, increasing from foreground to background.
- $y_i$ , highway gas mileage, increasing from bottom to top.

Since it can be hard to show a 3D cloud of points on a 2D page, a color scale has been added to encode the height of each point in the  $y$  direction.

Fitting a linear equation for  $y$  versus  $x_1$  and  $x_2$  results in a regression model of the following form:

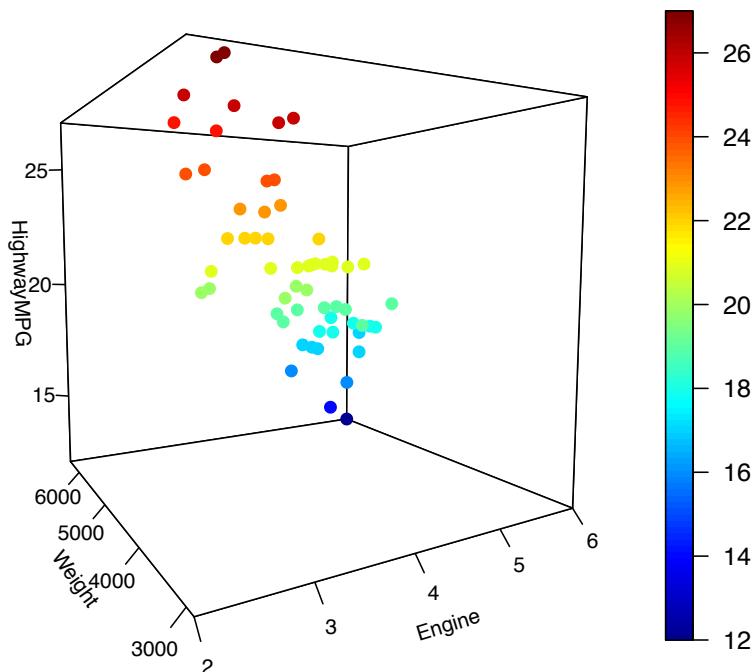
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i .$$

Just as before, we call the  $\beta$ 's the coefficients of the model and the  $e_i$ 's the residuals. In Figure 6.1, this fitted equation is

$$\text{MPG} = 33 - 1.35 \cdot \text{Displacement} - 0.00164 \cdot \text{Weight} + \text{Residual} .$$

<sup>1</sup> These are the same SUVs shown in the second-from-right panel in Figure 4.9, when we discussed ANOVA for models involving correlated predictors.

### Mileage versus weight and engine power



### With fitted plane

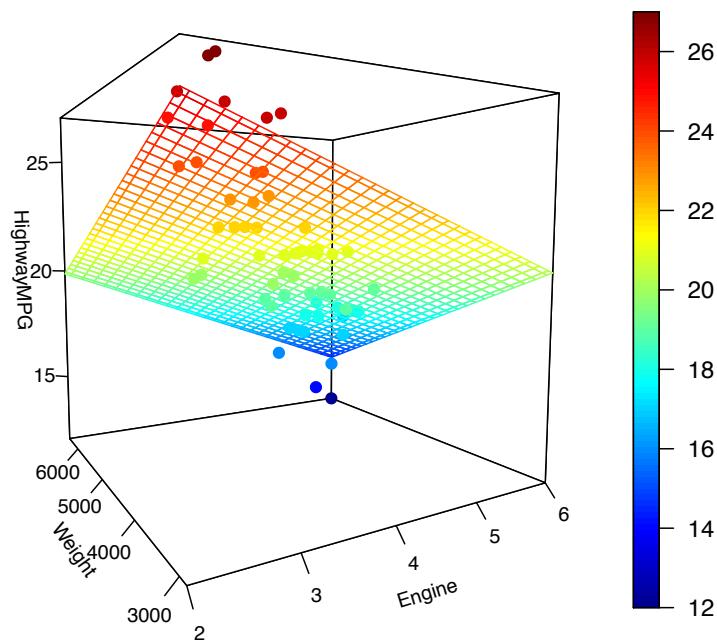


Figure 6.1: Highway gas mileage versus weight and engine displacement for 59 SUVs, with the least-squares fit shown in the bottom panel.

Both coefficients are negative, showing that gas mileage gets worse with increasing weight and engine displacement.

This equation is called a *multiple regression model*. In geometric terms, it describes a plane passing through a three-dimensional cloud of points, which we can see slicing roughly through the middle of the points in the bottom panel in Figure 6.1. This plane has a similar interpretation as the line did in a simple one-dimensional linear regression. If you read off the height of the plane along the  $y$  axis, then you know where the response variable is expected to be, on average, for a particular pair of values  $(x_1, x_2)$ .

*In more than two dimensions.* In principle, there's no reason to stop at two predictors. We can easily generalize this idea to fit regression equations using  $p$  different predictors  $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ :

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} = \beta_0 + \sum_{k=1}^p \beta_k x_{i,k}.$$

This is the equation of a  $p$ -dimensional plane embedded in  $(p+1)$ -dimensional space. This plane is nearly impossible to visualize beyond  $p = 2$ , but straightforward to describe mathematically.

*From simple to multiple regression: what stays the same.* In this jump from the familiar (straight lines in two dimensions) to the foreign (planes in arbitrary dimensions), it helps to start out by cataloguing several important features that don't change.

First, we still fit parameters of the model using the principle of least squares. As before, we will denote our estimates by  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ , and so on. For a given choice of these coefficients, and a given point in predictor space, the fitted value of  $y$  is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}.$$

This is a scalar quantity, even though the regression parameters describe a  $p$ -dimensional hyperplane. Therefore, we can define the residual sum of squares in the same way as before, as the sum of squared differences between fitted and observed values:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left\{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}) \right\}^2.$$

The principle of least squares prescribes that we should choose the estimates so as to make the residual sum of squares as small as possible, thereby distributing the “misses” among the observations

We use a bolded  $\mathbf{x}_i$  as shorthand to denote the whole vector of predictor values for observation  $i$ . That way we don't have to write out  $(x_{i,1}, x_{i,2}, \dots, x_{i,p})$  every time. When writing things out by hand, a little arrow can be used instead, since you obviously can't write things in bold:  $\vec{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ . By the same logic, we also write  $\vec{\beta}$  for the vector  $(\beta_0, \beta_1, \dots, \beta_p)$ .

in a roughly equal fashion. Just as before, the little  $e_i$  is the amount by which the fitted plane misses the actual observation  $y_i$ .

Second, these residuals still have the same interpretation as before: as the part of  $y$  that is unexplained by the predictors. For a least-squares fit, the residuals will be uncorrelated with each of the original predictors. Thus we can interpret  $e_i = y_i - \hat{y}_i$  as a statistically adjusted quantity: the  $y$  variable, adjusted for the systematic relationship between  $y$  and all of the  $x$ 's in the regression equation. Here, as before, statistical adjustment just means subtraction.

Third, we still summarize preciseness of fit using  $R^2$ , which has the same definition as before:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{UV}{TV} = \frac{PV}{TV}.$$

The only difference is that  $\hat{y}_i$  is now a function of more than just an intercept and a single slope. Also, just as before, it will still be the case  $R^2$  is the square of the correlation coefficient between  $y_i$  and  $\hat{y}_i$ . It will not, however, be expressible as the correlation between  $y$  and any of the original predictors, since we now have more than one predictor to account for. (Indeed,  $R^2$  is a natural generalization of Pearson's  $r$  for measuring correlation between one response and a whole basket of predictors.)

Finally, we still estimate the residual standard deviation using the same formula as before:

$$s_e = \sqrt{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

One slightly tricky thing to keep in mind is that  $p$  refers to the number of free parameters in the model. So in the model for mileage versus engine size and weight, we have  $p = 3$ : an intercept ( $\beta_0$ ), an engine-size coefficient ( $\beta_1$ ) and a weight coefficient ( $\beta_2$ ). Your regression software should keep track of this for you.

## Multiple regression and partial relationships

NOT everything about our inferential process stays the same when we move from lines to planes. We will focus more on some of the differences later, but for now, we'll mention a major one: the interpretation of each  $\beta$  coefficient is no longer quite so simple as the interpretation of the slope in one-variable linear regression.

The best way to think of  $\hat{\beta}_k$  is as an estimated *partial slope*: that is, the change in  $y$  associated with a one-unit change in  $x_k$ , holding all other variables constant. This is a subtle interpretation that is worth considering at length. To understand it, it helps to isolate the contribution of  $x_k$  on the right-hand side of the regression equation. For example, suppose we have two numerical predictors, and we want to interpret the coefficient associated with  $x_2$ . Our equation is

$$\underbrace{y_i}_{\text{Response}} = \beta_0 + \underbrace{\beta_1 x_{i1}}_{\text{Effect of } x_1} + \underbrace{\beta_2 x_{i2}}_{\text{Effect of } x_2} + \underbrace{e_i}_{\text{Residual}}.$$

To interpret the effect of the  $x_2$  variable, we isolate that part of the equation on the right-hand side, by subtracting the contribution of  $x_1$  from both sides:

$$\underbrace{y_i - \beta_1 x_{i1}}_{\text{Response, adjusted for } x_1} = \underbrace{\beta_0 + \beta_2 x_{i2}}_{\text{Regression on } x_2} + \underbrace{e_i}_{\text{Residual}}.$$

On the left-hand side, we have something familiar from one-variable linear regression: the  $y$  variable, adjusted for the effect of  $x_1$ . If it weren't for the  $x_2$  variable, this would just be the residual in a one-variable regression model. Thus we might call this term a *partial residual*.

On the right-hand side we also have something familiar: an ordinary one-dimensional regression equation with  $x_2$  as a predictor. We know how to interpret this as well: the slope of a linear regression quantifies the change of the left-hand side that we expect to see with a one-unit change in the predictor (here,  $x_2$ ). But here the left-hand side isn't  $y$ ; it is  $y$ , adjusted for  $x_1$ . We therefore conclude that  $\beta_2$  is the change in  $y$ , *once we adjust for the changes in  $y$  due to  $x_1$* , that we expect to see with a one-unit change in the  $x_2$  variable.

This same line of reasoning can allow us to interpret  $\beta_1$  as well:

$$\underbrace{y_i - \beta_2 x_{i2}}_{\text{Response, adjusted for } x_2} = \underbrace{\beta_0 + \beta_1 x_{i1}}_{\text{Regression on } x_1} + \underbrace{e_i}_{\text{Residual}}.$$

Thus  $\beta_1$  is the change in  $y$ , *once we adjust for the changes in  $y$  due to  $x_2$* , that we expect to see with a one-unit change in the  $x_1$  variable.

We can make the same argument in any multiple regression model involving two or more predictors, which we recall takes the form

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{i,k} + e_i.$$

To interpret the coefficient on the  $j$ th predictor, we isolate it on the right-hand side:

$$y_i - \underbrace{\sum_{k \neq j} \beta_k x_{i,k}}_{\text{Response adjusted for all other } x\text{'s}} = \underbrace{\beta_0 + \beta_j x_{ij}}_{\text{Regression on } x_j} + \underbrace{e_i}_{\text{Residual}}.$$

Thus  $\beta_j$  represents the rate of change in  $y$  associated with one-unit change in  $x_j$ , after adjusting for all the changes in  $y$  that can be predicted by the other predictor variables.

*Partial versus overall relationships.* A multiple regression equation isolates a set of *partial relationships* between  $y$  and each of the predictor variables. By a partial relationship, we mean the relationship between  $y$  and a single variable  $x$ , holding other variables constant. The partial relationship between  $y$  and  $x$  is very different than the *overall relationship* between  $y$  and  $x$ , because the latter ignores the effects of the other variables. When the two predictor variables are correlated, this difference matters a great deal.

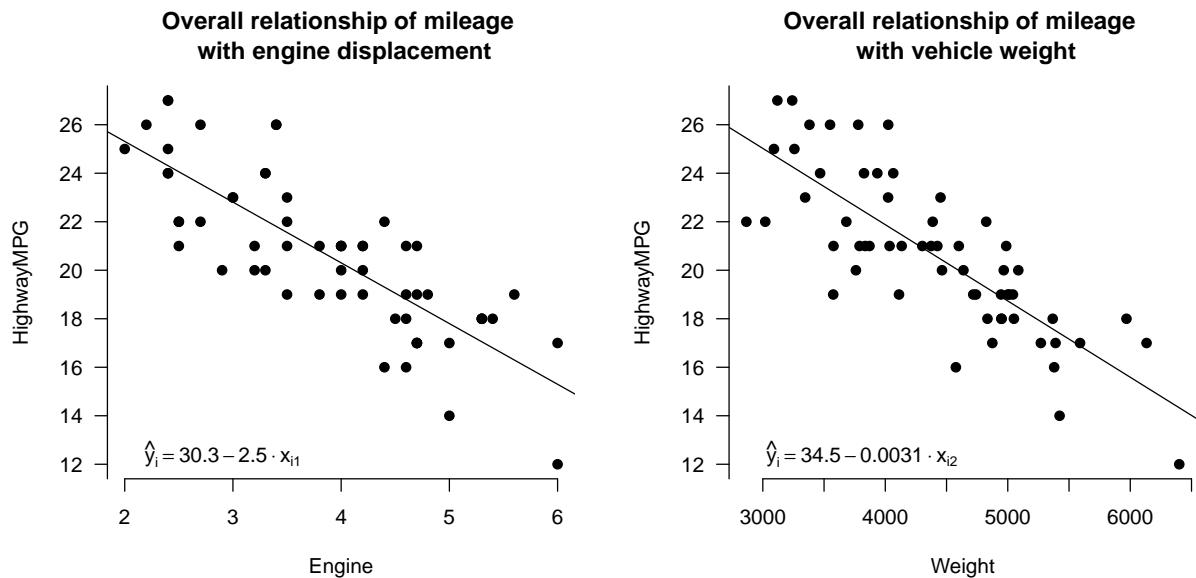
To compare these two types of relationships, let's take the multiple regression model we fit to the data on SUVs in Figure 6.1:

$$\text{MPG} = 33 - 1.35 \cdot \text{Displacement} - 0.00164 \cdot \text{Weight} + \text{Residual}.$$

This model isolates two partial relationships:

- We expect highway gas mileage to decrease by 1.35 MPG for every 1-liter increase in engine displacement, after adjusting for the simultaneous effect of vehicle weight on mileage. That is, if we held weight constant and increased the engine size by 1 liter, we'd expect mileage to go down by 1.35 MPG.
- We expect highway gas mileage to decrease by 1.64 MPG for every additional 1,000 pounds of vehicle weight, after adjusting for the simultaneous effect of engine displacement on gas mileage. That is, if we held engine displacement constant and added 1,000 pounds of weight to an SUV, we'd expect mileage to go down by 1.64 MPG.

Let's compare these partial relationships with the overall relationships depicted in Figure 6.2. Here we've fit two separate one-variable regression models: mileage versus engine displacement on the left, and mileage versus vehicle weight on the right.



Focus on the left panel of Figure 6.2 first. The least-squares fit to the data is

$$\text{MPG} = 30.3 - 2.5 \cdot \text{Displacement} + \text{Residual}.$$

Thus when displacement goes up by 1 liter, we expect mileage to go down by 2.5 MPG. This overall slope is quite different from the partial slope of  $-1.35$  isolated by the multiple regression equation. That's because this model doesn't attempt to adjust for the effects of vehicle weight. Because weight is correlated with engine displacement, we get a steeper estimate for the overall relationship than for the partial relationship: for cars where engine displacement is larger, weight also tends to be larger, and the corresponding effect on the  $y$  variable isn't controlled for in the left panel.

Similarly, the overall relationship between mileage and weight is

$$\text{MPG} = 34.5 - 0.0031 \cdot \text{Weight} + \text{Residual}.$$

The overall slope of  $-0.0031$  is nearly twice as steep as the partial slope of  $-0.00164$ . The one-variable regression model hasn't successfully isolated the marginal effect of increased weight from that of increased engine displacement. But the multiple regression model has—and once we hold engine displacement constant, the marginal effect of increased weight on mileage looks smaller.

Figure 6.2: Overall relationships for highway gas mileage versus weight and engine displacement individually.

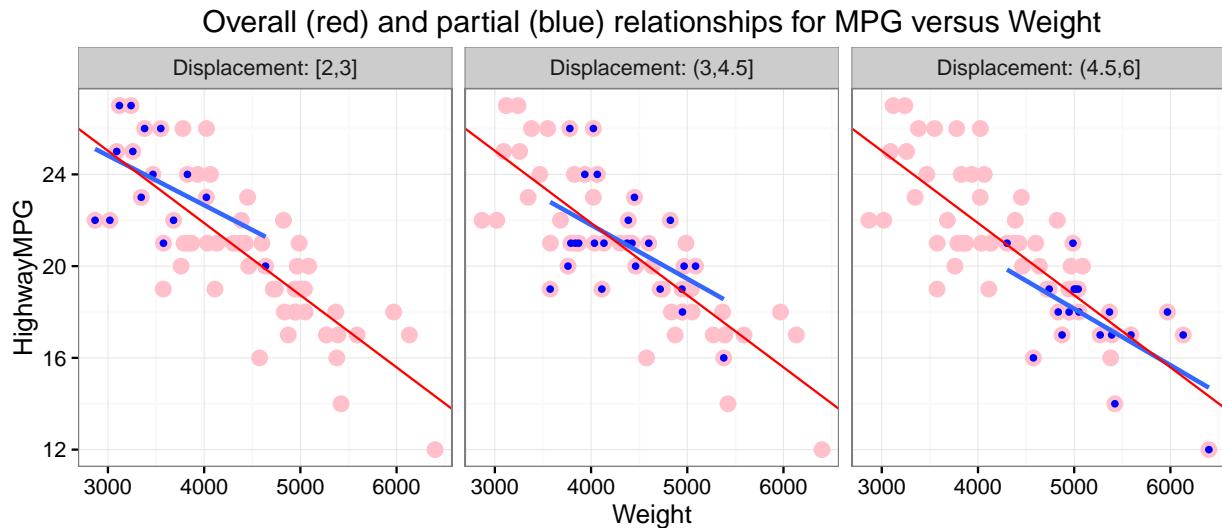


Figure 6.3 provides some intuition here about the difference between an overall and a partial relationship. The figure shows a lattice plot where the panels correspond to different strata of engine displacement: 2–3 liters, 3–4.5 liters, and 4.5–6 liters. Within each stratum, engine displacement doesn't vary by much—that is, it is approximately held constant. Each panel in the figure shows a straight line fit that is specific to the SUVs in each stratum (blue dots and line), together with the overall linear fit to the whole data set (red dots and line).

The two important things to notice here are the following.

- (1) The SUVs within each stratum of engine displacement are in systematically different parts of the  $x$ - $y$  plane. For the most part, the smaller engines are in the upper left, the middle-size engines are in the middle, and the bigger engines are in the bottom right. When weight varies, displacement also varies, and each of these variables have an effect on mileage. Another way of saying this is that engine displacement is a *confounding variable* for the relationship between mileage and weight. A confounder is something that is correlated with both the predictor and response.
- (2) In each panel, the blue line has a shallower slope than the red line. That is, when we compare SUVs that are similar in engine displacement, the mileage–weight relationship is not as steep

Figure 6.3: A lattice plot of mileage versus weight, stratified by engine displacement. The blue points within each panel show only the SUVs within a specific range of engine displacements:  $\leq 3$  liters on the left, 3–4.5 liters in the middle, and  $> 4.5$  liters on the right. The blue line shows the least-squares fit to the blue points alone within each panel. For reference, the entire data set is also shown in each panel (pink dots), together with the overall fit (red line) from the right-hand side of Figure 6.2. The blue lines are shallower than the red line, suggesting that once we hold engine displacement approximately (thought not perfectly) constant, we estimate a different (less steep) relationship between mileage and weight.

as it is when we compare SUVs with very different engine displacements.

This second point—that when we hold displacement roughly constant, we get a shallower slope for mileage versus weight—explains why the partial relationship estimated by the multiple regression model is different than the overall relationship from the left panel of Figure 6.2.<sup>2</sup> The slope of  $-1.64 \times 10^{-3}$  MPG per pound from the multiple regression model addresses the question: how fast should we expect mileage to change when we compare SUVs with different weights, but with the same engine displacement? This is similar to the question answered by the blue lines in Figure 6.3, but different than the question answer by the red line.

It is important to keep in mind that this “isolation” or “adjustment” is statistical in nature, rather than experimental. Most real-world systems simply don’t have isolated variables. Confounding tends to be the rule, rather than the exception. The only real way to isolate a single factor is to run an experiment that actively manipulates the value of one predictor, holding the others constant, and to see how these changes affect  $y$ . Still, using a multiple-regression model to perform a statistical adjustment is often the best we can do when facing questions about partial relationships that, for whatever reason, aren’t amenable to experimentation.

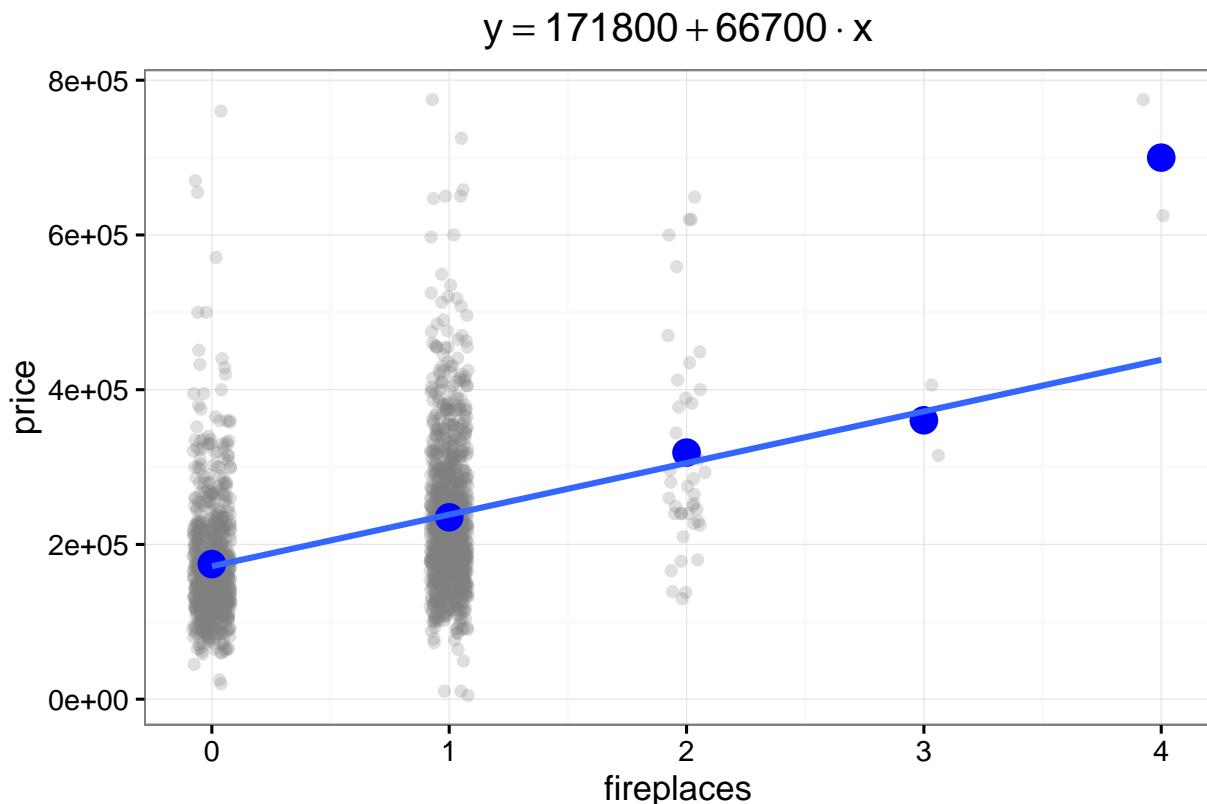
<sup>2</sup> This is a very general property of regression: if  $x_1$  and  $x_2$  are two correlated (collinear) predictors, then adding  $x_2$  to the model will change the coefficient on  $x_1$ , compared to a model with  $x_1$  alone.

## Using multiple regression to address real-world questions

While there are many possible uses of multiple regression, most applications will fall into one of two categories:

- (1) Isolating a partial relationship between the response and a predictor of interest, adjusting for possible confounders.
- (2) Building a predictive model for forecasting the response, using all available sources of information.

In the rest of this chapter, we’ll see examples in each category. As a case study, we’ll use a running example on house prices from Saratoga County, New York, distributed as part of the `mosaic` R package. We’ll show how, together with multiple regression, this data set can be used to address a few interesting questions of the kind that might be relevant to anyone buying, selling, or assessing the taxable value of a house.



*How much is a fireplace worth?*

Our first question is: how much does a fireplace improve the value of a house for sale? Figure 6.4 would seem to say: by about \$66,700 per fireplace. This dot plot shows the sale price of houses in Saratoga County, NY that were on the market in 2006.<sup>3</sup> We also see a linear regression model for house price versus number of fireplaces, leading to the equation

$$\text{Price} = \$171800 + 66,700 \cdot \text{Fireplaces} + \text{Residual},$$

This fitted equation is shown as a blue line in Figure 6.4. The means of the individual groups (1 fireplace, 2 fireplaces, etc) are also shown as blue dots. This helps us to verify that the assumption of linearity is reasonable here: the line passes almost right through the group means, except the one for houses with four fireplaces (which corresponds to just two houses).

But before you go knocking a hole in your ceiling and hiring a

Figure 6.4: The relationship between the price of a house and the number of fireplaces it has.

<sup>3</sup> Data from “House Price Capitalization of Education by Part Year Residents,” by Candice Corvetti. Williams College honors thesis, 2007, [available here](#), and in the mosaic R package.

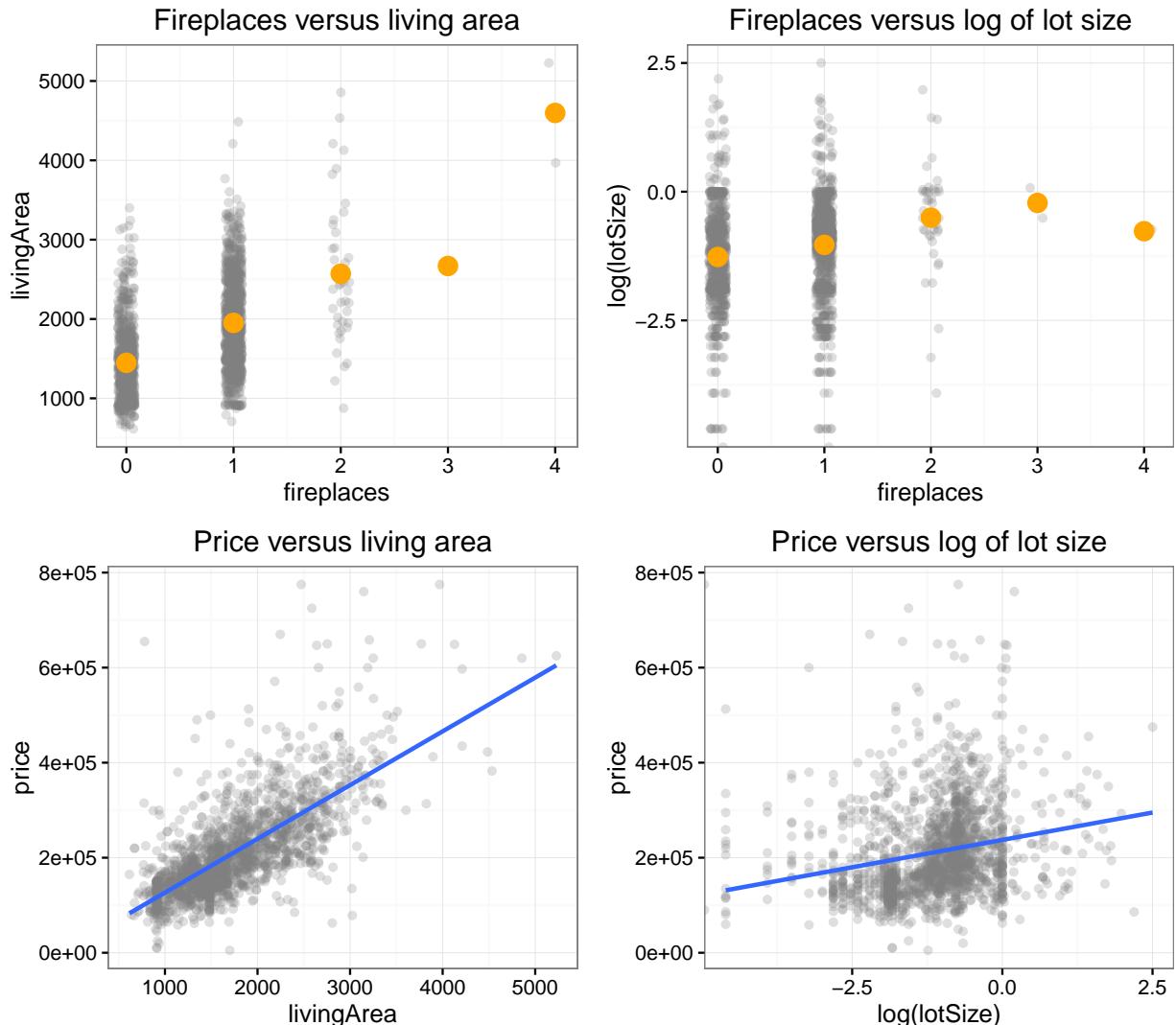


Figure 6.5: The relationship of house price with living area (bottom left) and with the logarithm of lot size in acres (bottom right). Both of these variables are potential confounders for the relationship between fireplaces and price, because they are also correlated with the number of fireplaces (top row).

bricklayer so that you might cash in on your new fireplace, consult Figure 6.5 on page 137. This figure shows that we should be careful in interpreting the figure of \$66,700 per fireplace arising from the simple one-variable model. Specifically, it shows that houses with more fireplaces also tend to be bigger (top left panel) and to sit on lots that have more land area (top right). These factors are also correlated with the price of a house.

Thus we have two possible explanations for the relationship we see in Figure 6.4. This correlation may happen because fireplaces are so valuable. On the other hand, it may instead (or also) happen because fireplaces happen to occur more frequently in houses that are desirable for other reasons (i.e. they are bigger). This is confounding again: when some third variable is correlated with both the response and the predictor of interest.

Disentangling these two possibilities requires estimating the partial relationship between fireplaces and prices, rather than the overall relationship shown in Figure 6.4. After all, when someone like a realtor or the county tax assessor asks how much a fireplace is worth, what they really want to know is: how much is a fireplace worth, holding other relevant features of the house constant?

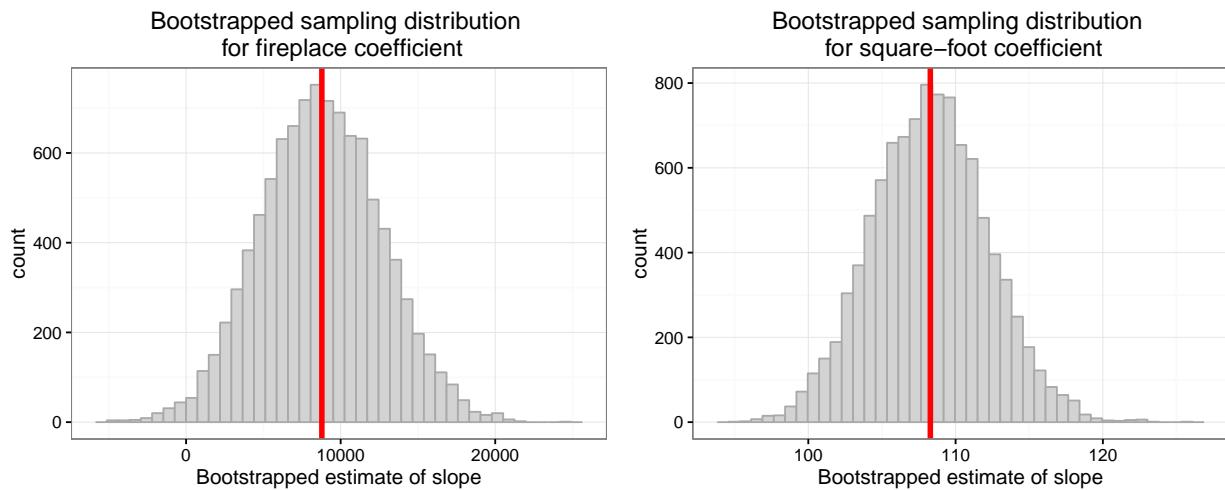
To address this question, we can fit a multiple regression model for price versus living area, lot size, and number of fireplaces. This will allow us to estimate the partial relationship between fireplaces and price, holding square footage and lot size constant. Such a model can tell us how much more we should expect a house with a fireplace to be worth, compared to a house that is identical in size and acreage but without a fireplace.

Fitting such a model to the data from Saratoga County yields the following equation:

$$\text{Price} = \$17787 + 108.3 \cdot \text{SqFt} + 1257 \cdot \log(\text{Acres}) + 8783 \cdot \text{Fireplaces} + \text{Residual}. \quad (6.1)$$

According to this model, the value of one extra fireplace is about \$8,783, holding square footage and lot size constant. This is a much lower figure than the \$66,700 fireplace premium that we would naively estimate from the overall relationship in Figure 6.4.

The example emphasizes the use of multiple regression to adjust statistically for the effect of confounders, by estimating a partial relationship between the response and the predictor of interest. This is one of the most useful real-world applications of regression modeling, and we'll see many similar examples. In general, the advice is: if you want to estimate a partial relationship, make sure



you include the potential confounders in the model.

### *Uncertainty quantification*

We can use bootstrapping to get confidence intervals for partial relationships in a multiple regression model, just as we do in a one-variable regression model.

The left panel of Figure 6.6 shows the bootstrapped estimate of the sampling distribution for the fireplace coefficient in our multiple regression model. The 95% confidence interval here is (1095, 16380). Thus while we do have some uncertainty we have about the value of a fireplace, we can definitively rule out the number estimated using the overall relationship from Figure 6.4. If the county tax assessor wanted to value your new fireplace at \$66,700 for property-tax purposes, Figure 6.6 would make a good argument in your appeal.<sup>4</sup>

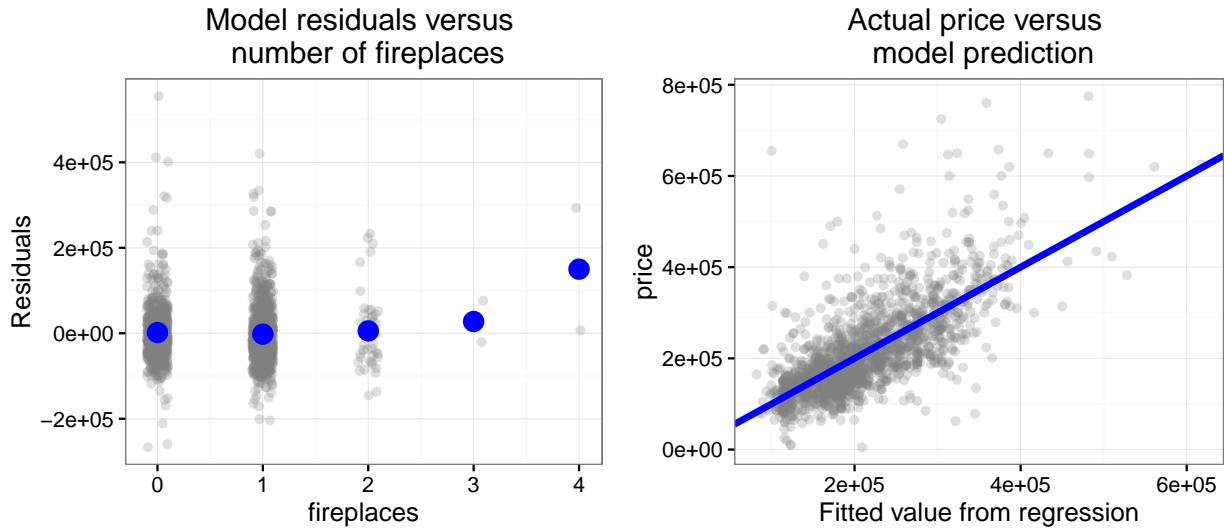
The right-hand side of Figure 6.6 shows the bootstrapped sampling distribution for the square-foot coefficient. While this wasn't the focus of our analysis here, it's interesting to know that an additional square foot improves the value of a property by about \$108, plus or minus about \$8.

### *Model checking*

However, before we put too much faith in the conclusions of your fitted model, it's important to check whether the assumption of a

Figure 6.6: Bootstrapped estimates for the sampling distributions of the partial slopes for number of fireplaces (left) and square footage (right) from the model in Equation 6.1 on page 138. The least-squares estimates are shown as vertical red lines.

<sup>4</sup> At a 2% property tax rate, this might save you over \$1000 a year in taxes.



linear regression model is appropriate in the first place. We call this step *model checking*. We'll learn a lot more about model checking later, but for now we'll cover the most basic step: validating that the response varies linearly with the predictors.

In one-variable regression models, we addressed this question using a plot of the residuals  $e_i$  versus the original predictor  $x_i$ . This allowed us to check whether there was still a pattern in the residuals that suggested a nonlinear relationship between the predictor and response. There are two ways to extend the idea of a residual plot to multiple regression models:

- plotting the residuals versus each of the predictors  $x_{ij}$  individually. This allows us to check whether the response changes linearly as a function of the  $j$ th predictor.
- plotting the actual values  $y_i$  versus the fitted values  $\hat{y}_i$  and looking for nonlinearities. This allows us to check whether the responses depart in a systematically nonlinear way from the model predictions.

Figure 6.7 shows an example of each plot. The left panel shows each the residual for each house versus the number of fireplaces it contains. Overall, this plot looks healthy: there are no obvious departures from linearity. The one caveat is that the predictions for houses with four fireplaces may be too low, which we can see from the fact that the mean residual for four-fireplace houses is positive. Then again, there are only two such houses, making it difficult to

Figure 6.7: Left: model residuals versus number of fireplaces. Right: observed house prices versus fitted house prices from the multiple regression model.

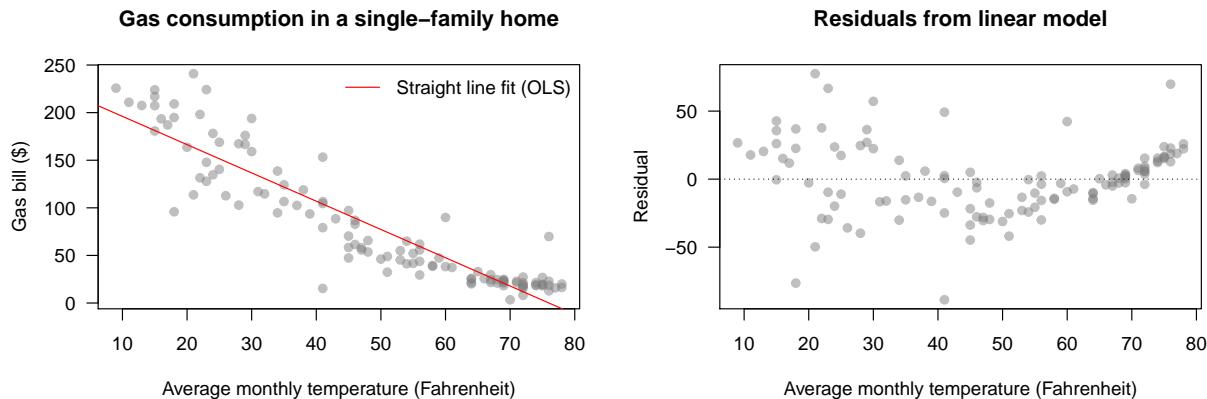


Figure 6.8: A copy of Figure 2.7 from the unit on nonlinear curve-fitting.

draw a firm conclusion here. We probably shouldn't change our model just to chase a better fit for two (very unusual) houses out of 1,726. But we should also recognize that our model might not be great at predicting the price for a house with four fireplaces, simply because it would involve extrapolation: we don't have a lot of data that can inform us about these houses.

The right panel of Figure 6.7 shows a plot of  $y_i$  versus  $\hat{y}_i$ . This also looks like a nice linear relationship, giving us further confidence that our model isn't severely distorting the true relationship between predictors and response. In a large multiple regression model with many predictors, it may be tedious to look at  $e_i$  versus each of those predictors individually. In such cases, a plot of  $y_i$  versus  $\hat{y}_i$  should be the first thing you examine to check for nonlinearities in the overall fit.

What would an unhealthy residual plot look like? To see an example, recall Figure 2.7 from the data set on gas consumption versus temperature, on page 44 (reproduced in Figure 6.8). Notice the pattern in the residual plot in the right panel:

- Below 20 degrees, the residuals are systematically above zero.
- Between 40 and 60 degrees, the residuals are below zero.
- Above 65 degrees, the residuals are again above zero.

The residuals *should* look like a random cloud centered around zero, and the fact that they don't suggests nonlinearity in the data.

In the case of the house-price model, imagine that we saw that the residuals for houses with no fireplace were systematically above zero, while the residuals for houses with one fireplace were

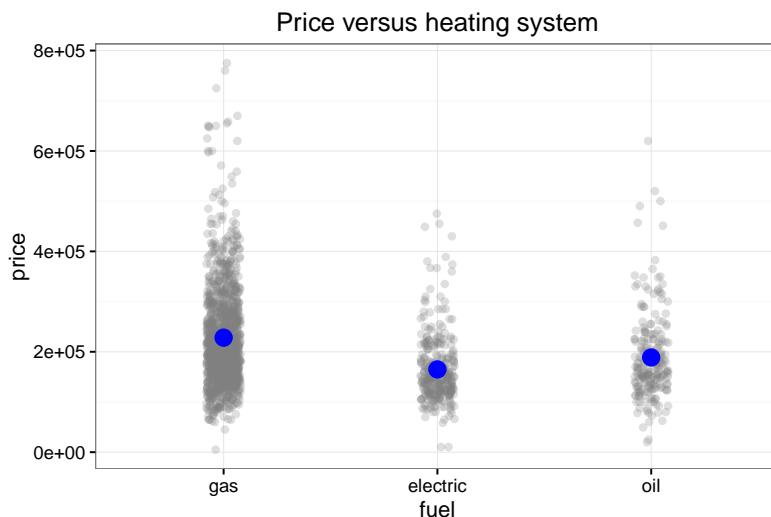


Figure 6.9: Prices of houses with gas, electric and fuel-oil heating systems.

systematically below zero. This would suggest a nonlinear effect that our model hasn't captured. Of course, we don't see these things, which gives credence to the linear model.

#### *How much is gas heating worth? Grouping variables in multiple regression*

Saratoga, NY is cold in the winter: the average January day has a low of  $13^{\circ}$  F and a high of  $31^{\circ}$  F. As you might imagine, residents spend a fair amount of money heating their homes, and are sensitive to the cost differences between gas, electric, and fuel-oil heaters. Figure 6.9 suggests that the Saratoga real-estate market puts a big premium for houses with gas heaters (mean price of \$228,000) versus those with electric or fuel-oil heaters (mean prices of \$165,000 and \$189,000, respectively). One possible reason is that gas heaters are cheaper to run and maintain.

But this figure shows an overall relationship. What does the story look like when we adjust for the effect of living area, lot size, and the number of fireplaces? There could be a confounding effect here. For example, maybe the bigger houses tend to have gas heaters more frequently than the small houses, or maybe fireplaces are used more in homes with expensive-to-use heating systems.

Remember: if you want to isolate a partial relationship, include potential confounders in the model. We'll do this here by including two sets of terms: (1) dummy variables for heating-system

Variable	Estimate	Std. Error	2.5%	97.5%
Intercept	29868	6743	16644	43093
livingArea	105	3	99	112
log(lotSize)	2705	1913	-1047	6457
fireplaces	7547	3348	980	14113
fuel=electric	-14010	4471	-22778	-5242
fuel=oil	-15879	5295	-26265	-5494

Table 6.1: Coefficients, standard errors, and 95% confidence intervals for the multiple regression model for house price ( $y$ ) versus living area, log of lot size, number of fireplaces, and heating system type.

type, to model the partial relationship of interest; and (2) all the possible confounding variables that we had in our previous regression equation (on page 138), which includes living area, lot size, and number of fireplaces. Fitting this model by least squares yields the following equation:

$$\text{Price} = \$29868 + 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{lotSize}) + 7546 \cdot \text{Fireplaces} \\ - 14010 \cdot \mathbf{1}_{\{\text{fuel} = \text{electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel} = \text{oil}\}} + \text{Residual}.$$

The full table of coefficients, standard errors, and 95% confidence intervals is in Table 6.1. The baseline here is gas heating, since it has no dummy variable.

Notice how the coefficients on the dummy variables for the other two types of heating systems shift the entire regression equation up or down. This model estimates the premium associated with gas heating to be about  $\$14,000 \pm 4500$  over electric heating (estimate, plus-or-minus one standard error), and about  $\$16,000 \pm 5300$  over fuel-oil heating. Because these are terms in a multiple regression model, these numbers represent partial relationships, adjusting for size, lot acreage, and number of fireplaces.

### Assessing statistical significance

A question that often comes up in multiple regression is whether a particular term in the model is “statistically significant” at some specified level (e.g. 5%). All this means is whether zero is a plausible value for that partial slope in the model. Remember, a coefficient of zero means that there is no partial relationship between the response and the corresponding predictor, adjusting for the other terms in the model. So when we say that a predictor is statistically significant, all we mean is that it we think it has a nonzero (partial) relationship with the response.

We'll take up the question of assessing statistical significance

in much more detail in the chapters to come. But here are a few quick observations and guidelines.

First, by convention, people express the statistical significance level as the opposite of the confidence level. So a confidence level of 95% means a significance level of 5%; a confidence level of 99% means a significance level of 1%; and so forth. This is confusing at first, but you'll get used to it. Just remember: the *lower* the significance level, the stronger the evidence that some variable has a nonzero relationship with the response.

Second, in regression models we can often<sup>5</sup> assess statistical significance just by looking at whether zero is included in the confidence interval. That's because "statistically significant" just means "zero is not a plausible value," and a confidence interval gives us a range of plausible values. For example, let's take the 95% confidence intervals for two terms in Table 6.1:

- The 95% confidence interval for the partial slope on fireplaces is (980, 14113). We can rule out zero as a plausible value at a 95% confidence level, and so we can say that the lot size variable is statistically significant at the 5% level.
- The 95% confidence interval for the partial slope on lot size is (-1047, 6457). We cannot rule out zero as a plausible value with 95% confidence, and so the lot size variable is not statistically significant at the 5% level.

Third, the fact that some variable is "statistically significant" does not mean that this variable is important in practical terms. A "significant" variable does not necessarily have a large effect on the response, nor is it automatically important for generating good predictions. Statistical significance means that we think the corresponding coefficient isn't zero. But it could still be very small. This is why, in most cases, it is better to focus on a variable's confidence interval, rather than on whether a variable is significant. The confidence interval carries a lot more information than a simplistic distinction between "significant" and "insignificant," because it gives you a range of plausible values for the coefficient.

Finally, the fact that some variable is *not* statistically significant does not imply that this variable has no relationship with the response, or that it should automatically be dropped from the model. A lack of statistical significance could just mean a big standard error—in other words, that we have a lot of uncertainty about the numerical magnitude of some variable's partial relationship

<sup>5</sup> But not always; see the next chapter.

with the response. There's an important but subtle distinction here: an insignificant coefficient means that we have an *absence of compelling evidence* for a nonzero effect. It does not mean that we have found *compelling evidence that the effect is absent*.

For example, the confidence interval for the log(acres) term in Table 6.1 is  $(-1047, 6457)$ . We therefore cannot rule out zero as a plausible value. But there are lots of large values, like 5000 or 6000, that we cannot rule out, either! There's a lot of uncertainty here. One symptom of this is a big standard error; another symptom is a lack of statistical significance at the 5% level. But it does not follow that lot size is irrelevant for predicting house price.<sup>6</sup>

#### *Prediction intervals from multiple regression models*

Suppose you have a house in Saratoga, NY that you're about to put up for sale. It's a 1900 square-foot house on a 0.7-acre lot. It has 3 bedrooms, 2.5 bathrooms,<sup>7</sup> 1 fireplace, gas heating, and central air conditioning. The house was built 16 years ago. (If you're counting, that's 8 possible predictors.) How much would you expect it to sell for?

A great way to assess the value of the house is to use the available data to fit a multiple regression model for its price, given its features. Building regression models for prediction is a rich, important topic that we'll consider in more detail later. For now, let's suppose we choose to fit a model for price versus all 8 variables mentioned above: bedrooms, bathrooms, living area, lot size, fireplaces, fuel system type, presence of central air conditioning, and the age of the home. Table 6.2 gives the coefficients, standard er-

<sup>6</sup> In this the large standard error is almost surely due to collinearity between lot size and other predictors, which we will discuss further in a later chapter.

<sup>7</sup> A half-bathroom has a toilet but no bath or shower.

Variable	Estimate	Std. Error	2.5%	97.5%
Intercept	48549	8190	32486	64611
bedrooms	-12263	2653	-17467	-7059
bathrooms	21330	3756	13965	28696
livingArea	98	5	89	107
lotSize	9514	2387	4832	14195
fireplaces	1017	3304	-5464	7498
fuel=electric	-14318	4467	-23080	-5556
fue=oil	-10465	5290	-20841	-89
centralAir=No	-19964	3665	-27151	-12776
age	28	63	-95	151

Table 6.2: Coefficients, standard errors, and 95% confidence intervals for our basic predictive model of house price.

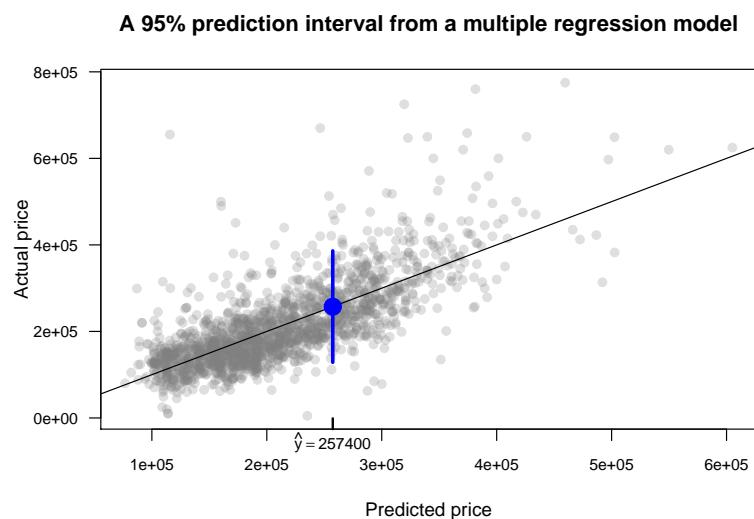


Figure 6.10: A 95% prediction interval (vertical blue line) for a house whose forecasted price is  $\hat{y} = \$257,400$ , using the model shown in Table 6.2.

rors, and 95% confidence intervals for this model. These, in turn, can be used to form a prediction interval for a “future” house with predictors  $(x_1^*, \dots, x_p^*)$ , just as we did back in the chapter on one-variable linear regression:

$$y^* \in \underbrace{\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j^*}_{\text{Best guess, } \hat{y}^*} \pm \underbrace{k \cdot s_e}_{\text{Uncertainty}} ,$$

where  $k$  is a chosen multiple, and where  $s_e$  is the standard deviation of the model residuals.<sup>8</sup>

We’re still using multiple regression here, but the goal here is slightly different than in the previous examples. Here, we don’t care so much about isolating and interpreting one partial relationship (like that between fireplaces and price). Instead, we just want to include any variables that will help us improve our predictions.

The model in Table 6.2 tells that us that, for your 1900-square-foot house in Saratoga on 0.7 beautiful acres, the expected price is  $\hat{y} = 257400$ ; that the residual standard deviation is  $s_e = 65600$ ; and that the 95% prediction interval is  $(128600, 386200)$ . See Figure 6.10; that’s a pretty wide range, reflecting the considerable variation in the price of different houses—even houses that look pretty similar on the page.

<sup>8</sup> As before, this is a slightly oversimplified formula, in that it ignores uncertainty due to lack of perfect knowledge about the parameters. Most regression software will use the correct (but much more complicated) formulas to calculate prediction intervals, e.g. the “predict” function in R.

## 7

## Testing hypotheses

### Assessing the evidence for a hypothesis

AMONG professional football fans, the New England Patriots are a polarizing team. Their fan base is hugely devoted, probably due to their long run of success over more than a decade. Many others, however, dislike the Patriots for their highly publicized cheating episodes, whether for deflating footballs or clandestinely filming the practice sessions of their opponents. This feeling is so common among football fans that sports websites often run images like the one at right (of the Patriots' be-hoodied head coach, Bill Belichick), or articles with titles like “[11 reasons why people hate the Patriots.](#)” Despite—or perhaps because of—their success, the Patriots always seem to be dogged by scandal and ill will.

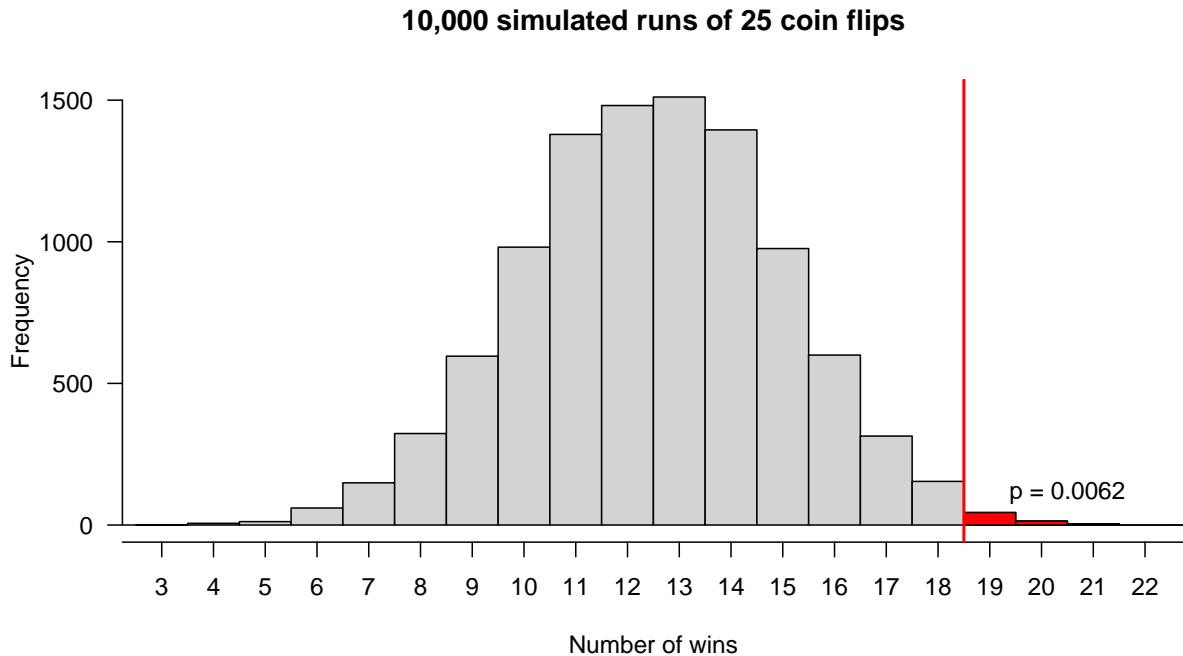
But could even the Patriots cheat at the pre-game *coin toss*?

Believe it or not, many people think so! That’s because, for a stretch of 25 games spanning the 2014–15 NFL seasons, the Patriots won 19 out of 25 coin tosses—that’s a 76% winning percentage. Needless to say, the Patriots’ detractors found this infuriating. As one TV commentator remarked when this unusual fact was brought to his attention: “This just proves that either God or the devil is a Patriots fan, and it sure can’t be God.”

But before turning to religion, let’s take a closer look at the evidence. Just how likely is it that one team could win the pre-game coin toss at least 19 out of 25 times, assuming that there’s no cheating going on?

This question is easy to answer using probability theory—specifically, something called the binomial distribution. But it’s also very easy to answer using the Monte Carlo method, in which we write a computer program that simulates a random process. In Figure 7.1, we see the results of a Monte Carlo simulation for pre-game NFL coin tosses, where the Patriots ought to have a 50% chance of winning each toss. Specifically, we have repeated the





following simple process 10,000 times:

1. Simulate 25 coin tosses in which the Patriots have a 50% chance of winning each toss.
2. Count how many times out of 25 that the Patriots won the toss.

If you're counting, that's 250,000 coin tosses: 10,000 simulations of 25 tosses each.

Figure 7.1 shows a histogram of the number of coin tosses won by the Patriots across 10,000 simulations. Clearly 19 wins is an unusual, although not impossible, number under this distribution: in our simulation, the Patriots won at least 19 tosses only 62 of 10,000 times ( $p = 0.0062$ ), shown as the red area in Figure 7.1.

So did the Patriots win 19 out of 25 coin tosses by chance? Well, nobody knows for sure—I report, you decide.<sup>1</sup> But unless you're a hard-core NFL conspiracy theorist, let me encourage you to forget the Patriots for a moment and focus instead on the process we've just gone through. This simple example has all the major elements of *hypothesis testing*, which is the subject of this chapter:

Figure 7.1: This histogram shows the results of a Monte Carlo simulation, in which we count the number of wins in 25 simulated coin flips over 10,000 different simulations. The red area (which has cumulative probability of 0.0062) approximates the probability of winning 19 or more flips, out of 25.

<sup>1</sup> Despite the small probability of such an extreme result, it's hard to believe that the Patriots cheated on the coin toss, for a few reasons. First, how could they? The coin toss would be extremely hard to manipulate, even if you were inclined to do so. Moreover, the Patriots are just one team, and this is just one 25-game stretch. There are 32 NFL teams, so the probability that *one* of them would go on an unusual coin-toss winning streak over *some* 25-game stretch over a long time period is a lot larger than the number we've calculated. Finally, after this 25-game stretch, the Patriots reverted back to a more typical coin-toss winning percentage, closer to 50%. The 25-game stretch was probably just luck.

- (1) We have a *null hypothesis*, that the pre-game coin toss in the Patriots' games was truly random.
- (2) We use a *test statistic*, number of Patriots' coin-toss wins, to measure the evidence against the null hypothesis.
- (3) There is a way of calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin.
- (4) Finally, we used this probability distribution to assess whether the null hypothesis looked believable in light of the data.

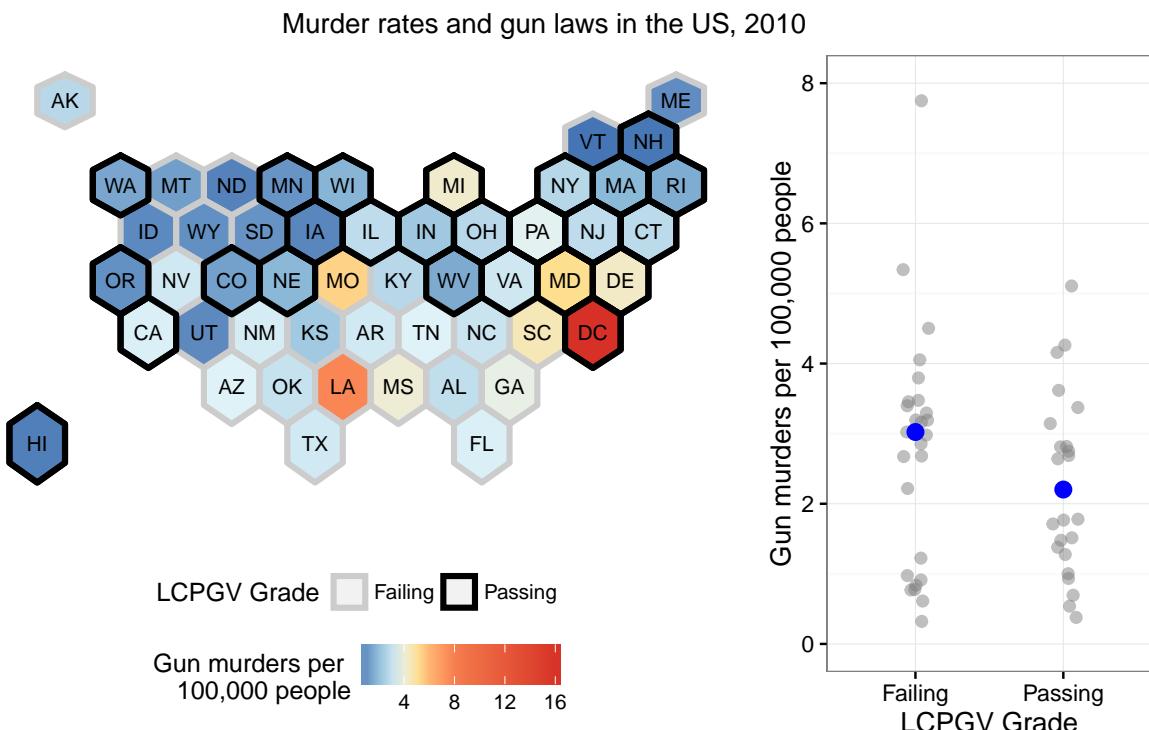
All hypothesis testing problems have these same four elements. Usually the difficult part is Step 3: calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. The essence of the problem is that, in most cases, we can't just run a simple simulation of coin flips. Luckily, there is a very general way of proceeding here, called the permutation test, which we will now learn about.

## Permutation tests

*Is gun violence correlated with gun policy?*

GUN policy is an important and emotionally charged topic in 21st-century America, where gun violence occurs with far higher frequency than it does in other rich countries. Many people feel strongly that certain types of guns, like military-style assault weapons, should be banned, and that all gun purchases should be subject to stronger background checks. Others view gun ownership as both an important part of their cultural heritage and a basic right protected by the U.S. Constitution. Like with many issues, there seems to be little prospect of a national consensus.

Both gun laws, and the likelihood of dying violently as a result of gun crime, vary significantly from state to state. Figure 7.2 shows some of this variation in a *chloropleth map*, where discrete areas on the map are shaded according to the value of some numerical variable. Notice that the states are shown as a gridded tile of equal-sized hexagons, rather than as an actual map of the United States. This is common technique used to avoid the visual imbalances due to large differences in the states' total area.



In the chloropleth map in Figure 7.2, the fill color indicates each state's gun-murder rate in 2010: blue is lower, red is higher. The outline color indicates whether a state's gun-control laws received a passing or failing grade from the Law Center to Prevent Gun Violence (LCPGV). The center graded each state's gun laws on an A–F letter-grade scale; here “failing” means a grade of F. In the figure, a black outline means a passing grade, while a grey outline means a failing grade.

The right panel of Figure 7.2 summarizes the relationship between gun laws and gun violence via a dot plot, together with the median for each group in blue. We use the median rather than the mean to estimate the center of each group, because the median is more robust to outliers; a clear example of an outlier here is Washington (D.C.), which at 16.2 gun murders per 100,000 people has a drastically higher rate than everywhere else in the country.

This dotplot shows that the median murder rate of states with a failing gun-laws grade is 3 murders per 100,000 people, while the median murder rate of states with a passing grade is 2.2 per

Figure 7.2: Left panel: a chloropleth map of murder rates versus gun laws across the U.S. states. The shaded color shows the state's gun-murder rate; blue is lower, and red is higher. The outline indicates whether a state's gun-control laws received a passing or a failing grade from the Law Center to Prevent Gun Violence (black for passing, grey for failing). The right panel shows a dot plot of the gun-murder rates across the two groups, together with the median for each group in blue. Washington (D.C.), at 16.2 gun murders per 100,000 people, is far off the top of the plot, but is still included in all calculations. According to its website, <http://smartgunlaws.org>, the LCPGV is “a national law center focused on providing comprehensive legal expertise in support of gun violence prevention and the promotion of smart gun laws that save lives.” You can read a full description of the methodology used to grade states at [this link](#).

100,000. On the face of it, it would seem as the states with stricter gun laws have lower murder rates.

Let's aside for a moment the fact that correlation does not establish causality. We will instead address the question: could this association have arisen due to chance? To make this idea more specific, imagine we took all 50 states and randomly divided them into two groups, arbitrarily labeled the "passing" states and the "failing" states. We would expect that the median murder rate would differ a little bit between the two groups, simply due to random variation (for the same reason that hands in a card game vary from deal to deal). But how big of a difference between these two groups could be explained by chance?

### *Null and alternative hypotheses*

Thus there are two hypotheses that can explain Figure 7.2:

- (1) There is no systematic relationship between murder rates and gun laws; the observed relationship between murder rates and gun laws is consistent with other unrelated sources of random variation.
- (2) The observed relationship between murder rates and gun laws is too large to be consistent with random variation.

We call hypothesis 1 the *null hypothesis*, often denoted  $H_0$ . Loosely, it states that nothing special is going on in our data, and that any relationship we thought might have existed isn't really there at all.<sup>2</sup> Meanwhile, hypothesis 2 is *alternative hypothesis*. In some cases the alternative hypothesis may just be the logical negation of the null hypothesis, but it can also be more specific.

In the approach to hypothesis testing that we'll learn here, we don't focus a whole lot on the alternative hypothesis.<sup>3</sup> Instead, we set out to check whether the null hypothesis looks plausible in light of the data—just as we did when we tried to check whether randomness could explain the Patriots' impressive run of 19 out of 25 coin flips won.

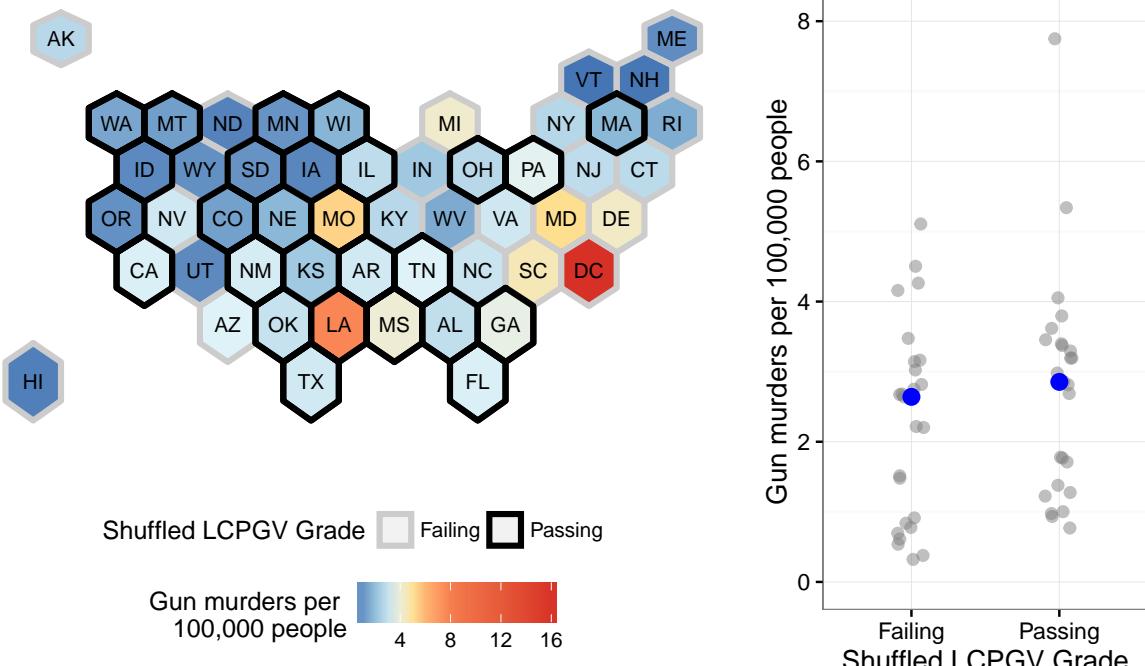
### *A permutation test: shuffling the cards*

In the Patriots' coin-flipping example, we could easily simulate data under the null hypothesis, by programming a computer to repeatedly flip a virtual coin and keep track of the winner. But of course, most real-life hypothesis-testing situations don't involve

<sup>2</sup> "Null hypothesis" is a term coined in the early twentieth century, back when "null" was a common synonym for "zero" or "lacking in distinctive qualities." So if the term sounds dated, that's because it is.

<sup>3</sup> Specifically, this approach is called the *Fisherian* approach, named after the English statistician Ronald Fisher. There are more nuanced approaches to hypothesis testing in which the alternative hypothesis plays a major role. These include the Neyman–Pearson framework and the Bayesian framework, both of which are widely used in the real world, but which are a lot more complicated to understand.

### Murder rates and gun laws under permutation



actual coin flips, which makes the virtual coin-flipping approach somewhat unhelpful as a general strategy.

It turns out, however, that in most situations, we can still harness the power of Monte Carlo simulation to understand what our data would look like if the null hypothesis were true. Rather than flipping virtual coins, we run something called a *permutation test*, which involves repeatedly permuting (or shuffling) the predictor variable and recalculating the statistic of interest.

To understand how this works, let's see an example. Figure 7.3 shows a map and dotplot very similar to those in Figure 7.2, with one crucial difference: in Figure 7.3, the identities of the states with notionally “passing” and “failing” gun laws have been randomly permuted. These grades bear no correspondence to reality. It's as though we took a deck of 51 cards, each card having some state's grade on it (treating D.C. as a state); shuffled the deck; and then dealt one card randomly to each state. The mathematical term for this is a *permutation* of the grades.

As expected, the median gun-murder rates of these two ran-

Figure 7.3: This map is almost identical to Figure 7.2, with one crucial difference: the identities of the states with passing and failing grades have been randomly permuted. There is still a small difference in the medians of the notionally passing and failing groups, due to random variation in the permutation process.

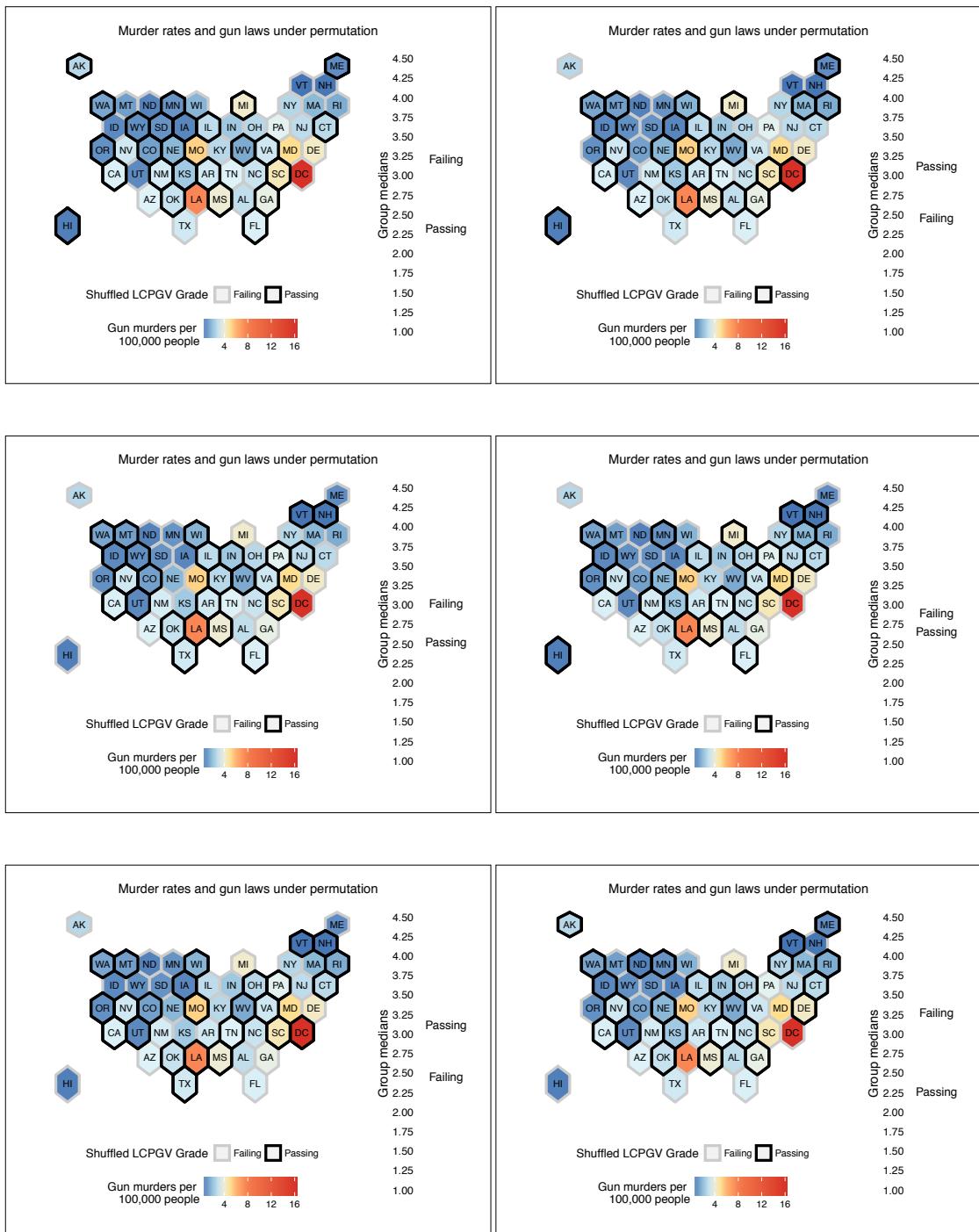


Figure 7.4: Six maps with permuted gun-law grades, with the medians for the passing and failing groups.

dom chosen “passing” and “failing” groups aren’t identical (right panel). The randomly chosen “failing” states have a median of 2.6, while the randomly chosen “passing” states have a slightly larger median of 2.8. Clearly we can get a difference in medians of at least 0.2 quite easily, just by random chance—that is, when the null hypothesis is true by design.

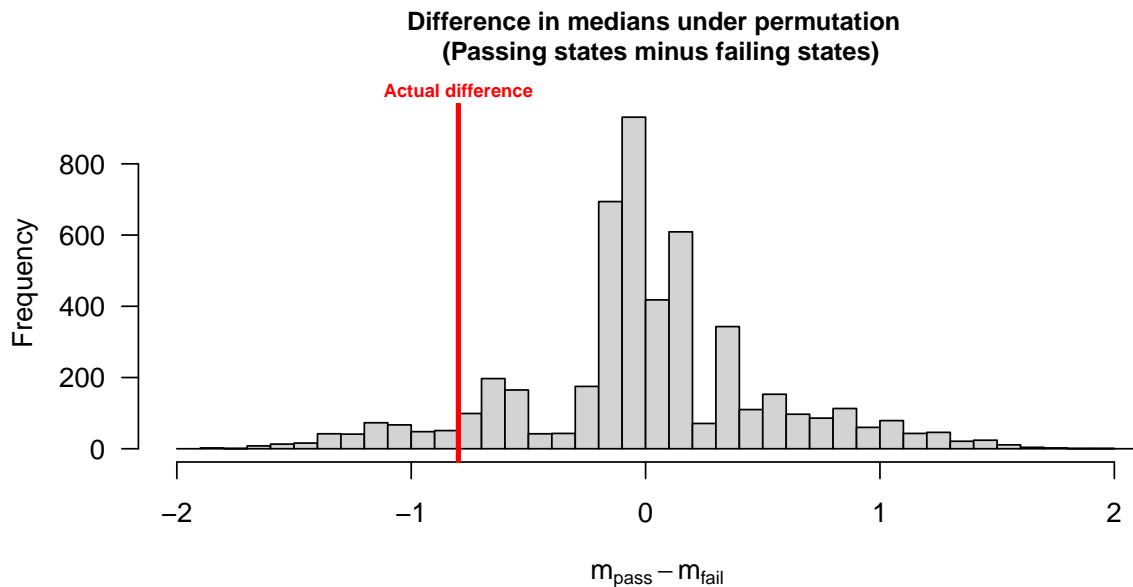
But Figure 7.3 shows the difference in medians for only a single permutation of the states’ gun-law grades. This permutation is random, and a different permutation would have given as a slightly different answer. Therefore, to assess whether could we get a difference in group medians as large as 0.8 just by random chance, we need to try several more permutations.

Figure 7.4 shows 6 more maps generated using the same permutation procedure. For each map, we shuffle the grade variables for all the states and recompute the median murder rates for the notionally “passing” and “failing” groups. Each map leads to its own difference in medians. In some maps, the difference is positive (“passing” states are higher), while in others it is negative (“failing” states are higher). In at least one of the 6 maps—the bottom right one—the median for the “failing” states exceeds the median for the “passing” states by more than 1 murder per 100,000 people, just by chance. This is a larger difference than we see for the real map, in Figure 7.2.

Six permutations give us some idea of how much a difference in the medians we could expect to see if the null hypothesis were true. But ideally we’d have many more than 6. Figure 7.5 addresses this need, showing the result of a much larger Monte Carlo simulation in which we generated 5,000 random maps, each one with its own random permutation of the states’ gun-law grades. For each of these 5,000 maps, we computed the difference in medians between the notionally passing and failing groups. These 5,000 differences in group medians across the 5,000 maps are shown as a histogram in Figure 7.5.

#### *Hypothesis testing: a four-step process*

Let’s review the vocabulary that describes what we’ve done here. First, we specified a null hypothesis: that the correlation between rates of gun violence and state-level gun policies could be explained by other unrelated sources of random variation. We decided to measure this correlation using a specific statistic: the difference in medians between the states with passing grades and



those with failing grades. (Remember that a statistic is just some numerical summary of a data set.) To give this statistic a name, let's call it  $\Delta$  (for difference in medians). It's intuitively clear that the larger  $\Delta$  is, the less plausible the null hypothesis seems.

Figure 7.5 quantifies this intuition by giving us an idea of how much variation we can expect in the sampling distribution of our  $\Delta$  statistic under the hypothesis that there is no systematic relationship between gun laws and rates of gun violence. As before, the sampling distribution is simply the probability distribution of the statistic under repeated sampling from the population—in this case, assuming that the null hypothesis is true.

There are two possibilities here, corresponding to the null and alternative hypotheses. First, suppose that we frequently get at least as extreme a value of  $\Delta$  for a random map, like those in Figure 7.4, as we do in the real map from Figure 7.2. Then there's no reason to be especially impressed by the actual value of  $\delta = -0.8$  we calculated from the real map.<sup>4</sup> It could have easily happened by chance. Hence we will be unable to reject the null hypothesis; it could have explained the data after all. (An important thing to remember is that *failing to reject* the null hypothesis is not the

Figure 7.5: The histogram shows the difference in group medians for 5,000 simulated maps generated by the same permutation procedure as the 6 maps in Figure 7.4. Negative values indicate that the “failing” states had higher rates of gun violence than the “passing” states. The actual difference in medians for the real map in Figure 7.2 is shown as a vertical red line. This difference seems to be consistent with (although does not prove) the null hypothesis that other sources of random variation, and not necessarily state-level gun policy, explains the observed difference in murder rates.

<sup>4</sup> We use the lower-case  $\delta$  to denote the value of the test statistic for your specific sample, to distinguish it from the  $\Delta$ 's simulated under permutation.

same thing as *accepting* the null hypothesis as truth. To use a relationship metaphor: failing to reject the null hypothesis is not like getting married. It's more like agreeing not to break up this time.)

On the other hand, suppose that we almost always get a smaller value of  $\Delta$  in a random map than we do in the real map. Then we will probably find it difficult to believe that the correlation in the real map arose due to chance. We will instead be forced to reject the null hypothesis and conclude that it provides a poor description of the observable data.

Which of these two possibilities seems to apply in Figure 7.5? Here, the actual difference of  $-0.8$  for the real map in Figure 7.2 is shown as a vertical red line. Its position on the histogram suggests possibility (1) here:  $\delta = -0.8$  is consistent with (although does not prove) the null hypothesis that other sources of random variation unrelated to state-level gun policy can explain the observed difference in murder rates between the passing-grade and the failing-grade states.

To summarize, the four steps we followed above were:

- (1) Choose a null hypothesis  $H_0$ , the hypothesis that there is no systematic relationship between the predictor and response variables.
- (2) Choose a test statistic  $\Delta$  that is sensitive to departures from the null hypothesis.
- (3) Approximate  $P(\Delta | H_0)$ , the sampling distribution of the test statistic  $T$  under the assumption that  $H_0$  is true.
- (4) Assess whether the observed test statistic for your data,  $\delta$ , is consistent with  $P(\Delta | H_0)$ .

For the gun-laws example, our test statistic in step (2) was the difference in medians between the “passing” states and the “failing” states. We then accomplished step (3) by randomly permuting the values of the predictor (gun laws) and recomputing the test statistic for the permuted data set. This shuffling procedure is called a permutation test when it’s done in the context of this broader four-step process. There are other ways of accomplishing step (3)—for example, by appealing to probability theory and doing some math. But the permutation test is nice because it works for any test statistic (like the difference of medians in the previous example), and it doesn’t require any strong assumptions.

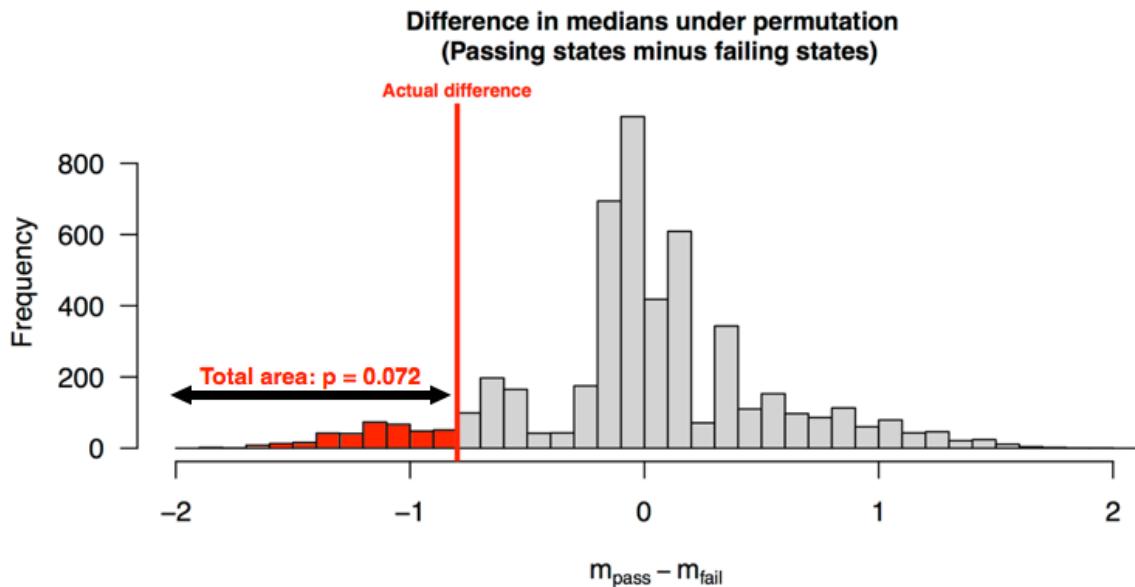


Figure 7.6: Assuming that the null hypothesis is true, the probability of observing a difference in medians at least as extreme as  $\delta = -0.8$  is  $p = 0.072$ . This tail area to the left of  $\delta = -0.8$  is the  $p$ -value of the test.

### Using and interpreting $p$ -values

There's one final question we haven't answered. How do we accomplish step (4) in the hypothesis test? That is, how can we measure whether the observed statistic for your data is consistent with the null hypothesis?

The typical approach here is to compute something called a *p-value*. Although we didn't call it by the name "*p-value*," this is exactly what we did for the Patriots' coin-flipping example at the beginning of the chapter.

Let's begin with a concise definition of a *p-value*, before we slowly unpack the definition (which is dense and non-intuitive).

*A p-value is the probability of observing a test statistic as extreme as, or more extreme than, the test statistic actually observed, given that the null hypothesis is true.* The way to compute the *p-value* is to calculate a *tail area* indicating what proportion of the sampling distribution,  $P(\Delta | H_0)$ , lies beyond the observed test statistic  $\delta$ .

This all sounds a bit abstract, but is much easier to understand by example. Let's go back to the gun-laws hypothesis test, where we observed a difference in the medians of  $\delta = -0.8$ . If the null hypothesis were true, the probability of getting  $\delta = -0.8$  (or

something more extreme in the negative direction) would be  $p = 0.072$ . We calculate this by taking the tail area under the sampling distribution that to the left of our observed  $\delta$  of  $-0.8$ . Figure 7.6 highlights this area in the left tail of the sampling distribution  $P(\Delta | H_0)$ . This is the  $p$ -value.

Using  $p$ -values has both advantages and disadvantages. The main advantage is that the  $p$ -value gives us a continuous measure of evidence against the null hypothesis. The smaller the  $p$ -value, the more unlikely it is that we would have seen our data under the null hypothesis, and therefore the greater the evidence the data provide that  $H_0$  is false.

The main disadvantage is that the  $p$ -value is hard to interpret correctly. Just look at the definition—it's pretty counterintuitive! To avoid having to think too hard about what a  $p$ -value actually means, people often take  $p \leq 0.05$  as a very important threshold that demarcates “significant” ( $p \leq 0.05$ ) from “insignificant” ( $p > 0.05$ ) results. While there are some legitimate reasons<sup>5</sup> for thinking in these terms, in practice, the  $p \leq 0.05$  criterion can feel pretty silly. After all, there isn't some magical threshold at which a result becomes important: in all practical terms,  $p = .049$  and  $p = .051$  are nearly identical in terms of the amount of evidence they provide against a null hypothesis.

Because of how counterintuitive  $p$ -values are, people make mistakes with them all the time, even (perhaps especially) people with Ph.D.'s quoting  $p$ -values in original research papers. Here is some advice about a few common misinterpretations:

- The  $p$ -value is *not* the probability that the null hypothesis is true, given that we have observed our statistic.
- The  $p$ -value is *not* the probability of having observed our statistic, given that the null hypothesis is true. Rather, it is the probability of having observed our statistic, *or any more extreme statistic*, given that the null hypothesis is true.
- The  $p$ -value is *not* the probability that your procedure will falsely reject the null hypothesis, given that the null hypothesis is true.<sup>6</sup>

The moral of the story is: always be careful when quoting or interpreting  $p$ -values. In many circumstances, a better question to ask than “what is the  $p$ -value?” is “what is a plausible range for the size of the effect?” This question can be answered with a confidence interval.<sup>7</sup>

<sup>5</sup> If you are interested in these reasons, you should read up on the Neyman–Pearson school of hypothesis testing.

<sup>6</sup> To get a guarantee of this sort, you have to set up a pre-specified rejection region for your  $p$ -value (like 0.05), in which case the size of that rejection region—and not the observed  $p$ -value itself—can be interpreted as the probability that your procedure will reject the null hypothesis, given that the null hypothesis is true. As above: if you're interested, read about the Neyman–Pearson approach to testing.

<sup>7</sup> In this case, you could get a confidence interval by bootstrapping the difference in medians between the two groups of states.

## Hypothesis testing in regression

To finish off this chapter, we will show how the permutation-testing framework can be used to answer questions about partial relationships in multiple regression modeling.

In a previous chapter, we asked the following question about houses in Saratoga, NY: what is the partial relationship between heating system type (gas, electric, or fuel oil) and sale price, once we adjust for the effect of living area, lot size, and the number of fireplaces? We fit a multiple regression model with these four predictors, which led to the following equation:

$$\begin{aligned} \text{Price} = & \$29868 + 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{Acres}) + 7546 \cdot \text{Fireplaces} \\ & - 14010 \cdot \mathbf{1}_{\{\text{fuel} = \text{electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel} = \text{oil}\}} + \text{Residual}. \end{aligned}$$

Remember that the baseline case here is gas heating, since it has no dummy variable. Our model estimated the premium associated with gas heating to be about \$14,000 over electric heating, and about \$16,000 over fuel-oil heating.

But are these differences due to heating-system type statistically significant, or could they be explained due to chance?

To answer this question, you could look at the confidence intervals for every coefficient associated with the heating-system variable, just as we learned to do in the chapter on multiple regression. The main difference is that before, we had one coefficient to look at, whereas now we have two: one dummy variable for fuel = electric, and one for fuel = oil. Two coefficients means two confidence intervals to look at.

Sometimes this strategy—that is, looking at the confidence intervals for all coefficients associated with a single variable—works just fine. For example, when the confidence intervals for all coefficients associated with a single variable are very far from zero, it's pretty obvious that the categorical variable in question is statistically significant.

But at other times, this strategy can lead to ambiguous results. In the context of the heating-system type variable, what if the 95% confidence interval for one dummy-variable coefficient contains zero, but the other doesn't? Or what if both confidence intervals contain zero, but just barely? Should we say that heating-system type is significant or not? This potential for ambiguous confidence intervals gets even worse when your categorical variable has more than just a few levels, because then there will be many more confi-

dence intervals to look at.

The core of the difficulty here is that we want to assess the significance of the heating-system variable itself, not the significance of any individual *level* of that variable. To assess the significance of the whole variable, with all of its levels, we'll use a permutation test. Specifically, we will compare two models:

- The *full model*, which contains variables for square footage, lot size, number of fireplaces, and heating system.
- The *reduced model*, which contains variables for square footage, lot size, and number of fireplaces, but not for heating system. We say that the reduced model is *nested* within the full model, since it contains a subset of the variables in the full model, but no additional variables.

As always, we must start by specifying  $H_0$ . Loosely speaking, our null hypothesis is that the reduced model provides an adequate description of house prices, and that the full model is needlessly complex. To be a bit more precise: the null hypothesis is that *there is no partial relationship* between heating system and house prices, once we adjust for square footage, lot size, and number of fireplaces. This implies that all of the *true* dummy variable coefficients for heating-system type are zero.

Next, we must pick a test statistic. A natural way to assess the evidence against the null hypothesis is to use improvement in  $R^2$  under the full model, compared to the reduced model. This is the same quantity we look at when assessing the importance of a variable in an ANOVA table. The idea is simple: if we see a big jump in  $R^2$  when moving from the reduced to the full model, then the variable we added (here, heating system) is important for predicting the outcome, and the null hypothesis of no partial relationship is probably wrong.

You might wonder here: why not use the coefficients on the dummy variables for heating-system type as test statistics? The reason is that there are two such coefficients (or in general,  $K - 1$  coefficients for a categorical variable with  $K$  levels). But we need a single number to use as our test statistic in a permutation test. Therefore we use  $R^2$ : it is a single number that summarizes the predictive improvement of the full model over the reduced model.

Of course, even if we were to add a useless predictor to the reduced model, we would expect  $R^2$  to go up, at least by a little bit, since the model would have more degrees of freedom (i.e. param-

Remember the four basic steps in a permutation test:

- (1) Choose a null hypothesis  $H_0$ .
- (2) Choose a test statistic  $\Delta$  that is sensitive to departures from the null hypothesis.
- (3) Repeatedly shuffle the predictor of interest and recalculate the test statistic after each shuffle, to approximate  $P(\Delta \mid H_0)$ , the sampling distribution of the test statistic  $T$  under the assumption that  $H_0$  is true.
- (4) Check whether the observed test statistic for your data,  $\delta$ , is consistent with  $P(\Delta \mid H_0)$ .

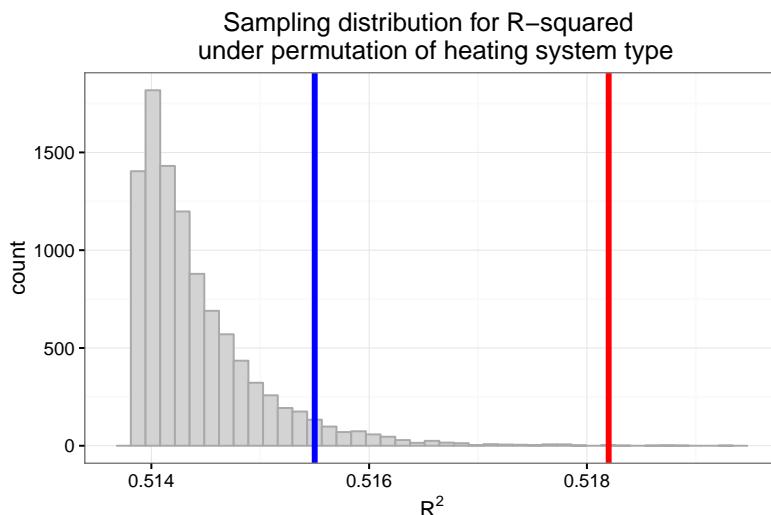


Figure 7.7: Sampling distribution of  $R^2$  under the null hypothesis that there is no partial relationship between heating system and price after adjusting for effects due to square footage, lot size, and number of fireplaces. The blue vertical line marks the 95th percentile of the sampling distribution (and so corresponds to a rejection region at the 5% level). The red line marks the actual value of  $R^2 = 0.518$  when we fit the full model by adding heating system to a model already containing the other three variables.

eters) that it can use to predict the observed outcome. Therefore, a more precise way of stating our null hypothesis is that, when we add heating system to a model already containing variables for square footage, lot size, and number of fireplaces, the improvement we see in  $R^2$  could plausibly be explained by chance, even if this variable had no partial relationship with price.

To carry out a hypothesis test, we need to approximate the sampling distribution of  $R^2$  under the null hypothesis. We will do so by repeatedly shuffling the heating system for every house (keeping all other variables the same), and re-fitting our model to each permuted data set. This breaks any partial relationship between heating system and price that may be present in our data. It tells us how big an improvement in  $R^2$  we'd expect to see when fitting the full model, even if the null hypothesis were true.

This sampling distribution is shown in Figure 7.7, which was generated by fitting the model to 10,000 data sets in which the heating-system variable had been randomly shuffled, but where the response and the variables in the reduced model have been left alone. As expected,  $R^2$  of the full model under permutation is always bigger than the value of  $R^2 = 0.513$  from the reduced model—but rarely by much. The blue line at  $R^2 = 0.5155$  shows the 95th percentile of the sampling distribution (i.e. the critical value for a rejection region at the 5% level). The red line shows the actual value of  $R^2 = 0.518$  from the full model fit the original

data set (i.e. with no shuffling). This test statistic falls far beyond the 5% rejection region. We therefore reject the null hypothesis and conclude that there is statistically significant evidence for an effect on price due to heating-system type.

One key point here is that we shuffled *only* heating-system type—or in general, whatever variable is being tested. We don’t shuffle the response or any of the other variables. That’s because we are interested in a partial relationship between heating-system type and price. Partial relationships are always defined with respect to a specific context of other control variables, and we have to leave these control variables as they are in order to provide the correct context for that partial relationship to be measured.

To summarize: we can compare any two nested models using a permutation test based on  $R^2$ , regardless of whether the variable in question is categorical or numerical. To do so, we repeatedly shuffle the extra variable in the full model—without shuffling either the response or the control variables (i.e. those that also appear in the reduced model). We fit the full model to each shuffled data set, and we track the sampling distribution of  $R^2$ . We then compare this distribution with the  $R^2$  we get when fitting the full model to the *actual* data set. If the actual  $R^2$  is a lot bigger than what we’d expect under the sampling distribution for  $R^2$  that we get under the permutation test, then we conclude that the extra variable in the full model is statistically significant.

*F tests and the normal linear regression model.* Most statistical software will produce an ANOVA table with an associated  $p$ -value for all variables. These  $p$ -values are approximations to the  $p$ -values that you’d get if you ran sequential permutation tests, adding and testing one variable at a time as you construct the ANOVA table. To be a bit more specific, they correspond to something called an  $F$  test under the normal linear regression model that we met awhile back:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + e_i, \quad e_i \sim N(0, \sigma^2).$$

You might want to revisit the discussion of the normal linear regression model starting on page 120. But the upshot is that an  $F$  test is conceptually similar to a permutation test based on  $R^2$ —and if you’re happy with the assumption of normally distributed residuals, you can treat the  $p$ -values from these two tests as virtually interchangeable.<sup>8</sup>

<sup>8</sup> If you’re not happy with this assumption, then you’re better off with the permutation test.

# *Building predictive models*

## **Building predictive models**

Suppose you have a house in Saratoga, NY that you’re about to put up for sale. It’s a 1900 square-foot house on a 0.7-acre lot.

It has 3 bedrooms, 2.5 bathrooms,<sup>1</sup> 1 fireplace, gas heating, and central air conditioning. The house was built 16 years ago. How much would you expect it to sell for?

Although we’ve been focusing on only a few variables of interest so far, our house-price data set actually has information on all these variables, and a few more besides. A great way to assess the value of the house is to use the available data to fit a multiple regression model for its price, given its features. We can then use this model to make a best guess for the price of a house with some particular combination of features—and, optionally, to form a prediction interval that quantifies the uncertainty of our guess.

We refer to this as the process of *building a predictive model*. Although we will still use multiple regression, the goal here is slightly different than in the previous examples. Here, we don’t care so much about isolating and interpreting one particular partial relationship (like that between fireplaces and price). Instead, we just want the most accurate predictions possible.

The key principle in building predictive models is *Occam’s razor*, which is the broader philosophical idea that models should be only as complex as they need to be in order to explain reality well. The principle is named after a medieval English theologian called William of Occam. Since he wrote in Latin, he put it like this: *Frustra fit per plura quod potest fieri per pauciora* (“It is futile to do with more things than which can be done with fewer.”) A more modern formulation of Occam’s razor might be the **KISS rule**: keep it simple, stupid.

In regression modeling, this principle is especially relevant for *variable selection*—that is, deciding which possible predictor variables to add to a model, and which to leave out. In this context,

<sup>1</sup> A half-bathroom has a toilet but no bath or shower.

Occam's razor is about finding the right set of variables to include so that we fit the data, without overfitting the data. Another way of saying this is that we want to find the patterns in the data, without memorizing the noise.

In this chapter, we'll consider two main questions:

- (1) How can we measure the predictive power of a model?
- (2) How can we find a model with good predictive power?

## Measuring generalization error

To understand how we measure the predictive power of a regression model, we first need a bit of notation. Specifically, let's say that we have estimated a multiple regression model with  $p$  predictors  $(x_1, x_2, \dots, x_p)$  to some data, giving us coefficients  $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ . Now we encounter a new case, not in our original data set. We'll let  $x^* = (x_1^*, x_2^*, \dots, x_p^*)$  be the predictor variables for this new case, and  $y^*$  denote the corresponding response. We will use the fitted regression model, together with  $x^*$ , to make a prediction for  $y^*$ :

$$\hat{y}^* = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j^*.$$

Our goal is to make the *generalization error*—that is, the difference between  $y^*$  and  $\hat{y}^*$ —as small as possible, on average.

A natural way to measure the generalization error of a regression model is using a quantity called the *mean-squared predictive error*, or MSPE. The mean-squared predictive error is a property of a fitted model, not an individual data point. It summarizes the magnitude of the errors we typically make when we use the model to make predictions  $\hat{y}^*$  on new data:

MSPE = Average value of  $(y^* - \hat{y}^*)^2$  when sampling new data points.

Here a “new” data point means one that hasn't been used to fit the model. You'll notice that, in calculating MSPE, we square the prediction error  $y^* - \hat{y}^*$  so that both positive and negative errors count equally.

Low mean-squared predictive error means that  $y^* - \hat{y}^*$  tends to be close to zero when we sample new data points. This gives us a simple principle for building a predictive model: find the model (i.e. the set of variables to include) with the lowest mean-squared predictive error.

### *Estimating the mean-squared predictive error*

Conceptually, the simplest way to estimate the mean-squared predictive error of a regression model is to actually collect new data and calculate the average predictive error made by our model. Specifically, suppose that, after having fit our model in the first place, we go out there and collect  $n^*$  brand new data points, with responses  $y_i^*$  and predictors  $(x_{i1}^*, \dots, x_{ip}^*)$ . We can then estimate the mean-squared predictive error of our model in two simple steps:

1. Form the prediction for each new data point:

$$\hat{y}_i^* = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}^*.$$

2. Calculate the average squared error of your predictions:

$$\widehat{\text{MSPE}}_{\text{out}} = \frac{1}{n^*} \sum_{i=1}^{n^*} (\hat{y}_i^* - y_i^*)^2.$$

Notice that we put a hat on MSPE, because the expression on the right-hand side is merely an *estimate* of the true mean-squared predictive error, calculated using a specific sample of new data points. (Calculating the *true* MSPE would require us, in principle, to average over all possible samples of new data points, which is obviously impractical.) We also use the subscript “out” to indicate that it is an *out-of-sample* measure—that is, calculated on new data, that falls outside of our original sample.

Conventionally, we report the square root of  $\widehat{\text{MSPE}}_{\text{out}}$  (which is called *root mean-squared predictive error*, or RMSPE), because this has the same units as the original  $y$  variable. You can think of the RMSPE as the standard deviation of future forecasting errors made by your model.

Assuming your new sample size  $n^*$  isn’t too small, these two steps are a nearly foolproof way to estimate the mean-squared predictive error of your model. The drawback, however, is obvious: you need a brand new data set, above and beyond the original data set that you used to fit the model in the first place. This new data set might be expensive or impractical to collect.

Thus we’re usually left in the position of needing to estimate the mean-squared predictive error of a model, without having access to a “new” data set. For this reason, the usual practice is

make a *train/test split* of your data: that is, to randomly split your original data set into two subsets, called the *training* and *testing* sets.

- The training set is used only to fit (“train”) the model—that is, to estimate the coefficients  $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ .
- The testing set is used only to estimate the mean-squared predictive error of the model. It is not used at all to fit the model. For this reason, the testing set is sometimes referred to as the “hold-out set,” since it is held out of the model-fitting process.

From this description, it should be clear that the training set plays the role of the “old” data, while the testing set plays the role of the “new” data.

This gives us a simple three-step procedure for choosing between several candidate models (i.e. different possible sets of variables to include).

- (1) Split your data into training and testing sets.
- (2) For each candidate model:
  - A. Fit the model using the training set.
  - B. Calculate  $\widehat{MSPE}_{out}$  for that model using the testing set.
- (3) Choose the model with the lowest value of  $\widehat{MSPE}_{out}$ .

*Choosing the training and testing sets.* A key principle here is that you must *randomly* split your data into a training set and testing set. Splitting your data nonrandomly—for example, taking the first 800 rows of your data as a training set, and the last 200 rows as a testing set—may mean that your training and testing sets are systematically different from one another. If this happens, your estimate of the mean-squared prediction error can be way off.

How much of the data should you reserve for the testing set? There are no hard-and-fast rules here. A common rule of thumb is to use about 75% of the data to train the model, and 25% to test it. Thus, for example, if you had 100 data points, you would randomly sample 75 of them to use for model training, and the remaining 25 to estimate the mean-squared predictive error. But other ratios (like 50% training, or 90% training) are common, too.

My general guideline is that the more data I have, the larger the fraction of that data I will use for training the predictive model.

Thus with only 100 data points, I might use a 75/25 split between training and testing; but with 10,000 data points, I might use more like a 90/10 split between training and testing. That's because estimating the model itself is generally harder than estimating the mean-squared predictive error.<sup>2</sup> Therefore, as more data accumulates, I like to preferentially allocate more of that data towards the intrinsically harder task of model estimation, rather than MSPE estimation.

<sup>2</sup> By "harder" here, I mean "subject to more sources of statistical error," as opposed to computationally more difficult.

*Averaging over different test sets.* It's a good idea to average your estimate of the mean-squared predictive error over several different train/test splits of the data set. This reduces the dependence of  $\widehat{\text{MSPE}}_{\text{out}}$  on the particular random split into training and testing sets that you happened to choose. One simple way to do this is average your estimate of MSPE over many different random splits of the data set into training and testing sets. Somewhere between 5 and 100 splits is typical, depending on the computational resources available (more is better, to reduce Monte Carlo variability).

Another classic way to estimate MSPE it is to divide your data set into  $K$  non-overlapping chunks, called *folds*. You then average your estimate of MPSE over  $K$  different testing sets, one corresponding to each fold of the data. This technique is called *cross validation*. A typical choice of  $K$  is five, which gives us five-fold cross validation. So when testing on the first fold, you use folds 2-5 to train the model; when testing on fold 2, you use folds 1 and 3-5 to train the model; and so forth.

*Can we use the original data to estimate the MSPE?*

A reasonable question is: why do even we need a new data set to estimate the mean-squared prediction error? After all, our fitted model has residuals,  $e_i = y_i - \hat{y}_i$ , which tell us how much our model has "missed" each data point in our sample. Why can't we just use the residual variance,  $s_e^2$ , to estimate the MSPE? This approach sounds great on the surface, in that we'd expect the past errors to provide a good guide to the likely magnitude of future errors. Thus you might be tempted to use the *in-sample* estimate of MSPE, denoted

$$\widehat{\text{MSPE}}_{\text{in}} = s_e^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where we recall that  $p$  is the number of parameters in the model.

Using  $\widehat{\text{MSPE}}_{\text{in}}$  certainly removes the need to collect a new data set. This turns out, however, to be a false economy:  $\widehat{\text{MSPE}}_{\text{in}}$  is usually too optimistic as an estimate of a model's generalization error. Practically speaking, this means the following. When we use  $\widehat{\text{MSPE}}_{\text{in}}$  to quantify the *in-sample* error of a model, and then we actually go out and take new data to calculate the *out-of-sample* generalization error  $\widehat{\text{MSPE}}_{\text{out}}$ , we tend to discover that the out-of-sample error is larger—sometimes much larger! This is called overfitting, and it is especially likely to happen when the size of the data set is small, or when the model we're fitting is very complex (i.e. has lots of parameters).

### *An example*

Let's see these ideas in practice, by comparing three predictive models for house prices in Saratoga, New York. Our models will draw from the following set of variables:

- lot size, in acres
- age of house, in years
- living area of house, in square feet
- percentage of residents in neighborhood with college degree
- number of bedrooms
- number of bathrooms
- number of total rooms
- number of fireplaces
- heating system type (hot air, hot water, electric)
- fuel system type (gas, fuel oil, electric)
- central air conditioning (yes or no)

We'll consider three possible models for price constructed from these 11 predictors.

*Small model:* price versus lot size, bedrooms, and bathrooms (4 total parameters, including the intercept).

*Medium model:* price versus all variables above, main effects only (14 total parameters, including the dummy variables).

*Big model:* price versus all variables listed above, together with all pairwise interactions between these variables (90 total parameters, include dummy variables and interactions).

Table 8.1 shows both  $\widehat{\text{MSPE}}_{\text{in}}$  and  $\widehat{\text{MSPE}}_{\text{out}}$  for these three models. To calculate  $\widehat{\text{MSPE}}_{\text{out}}$ , we used 80% of the data as a training

	In-sample RMSPE	Out-of-sample RMSPE	Difference
Small model: underfit	\$76,144	\$76,229	\$85
Medium model: good fit	\$65,315	\$65,719	\$403
Big model: overfit	\$61,817	\$71,426	\$9,609

set, and the remaining 20% as a test set, and we averaged over 100 different random train/test splits of the data. The final column, labeled “difference,” shows the difference between the in-sample and out-of-sample estimates of prediction error.

There are a few observations to take away from Table 8.1. The first is that the big model (with all the main effects and interactions) has the lowest in-sample error. With a residual standard deviation of \$61,817, it seems nearly \$3,500 more accurate than the medium model, which is next best. This is a special case of a very general phenomenon: a more complex model will always fit the data better, because it has more degrees of freedom to play with.

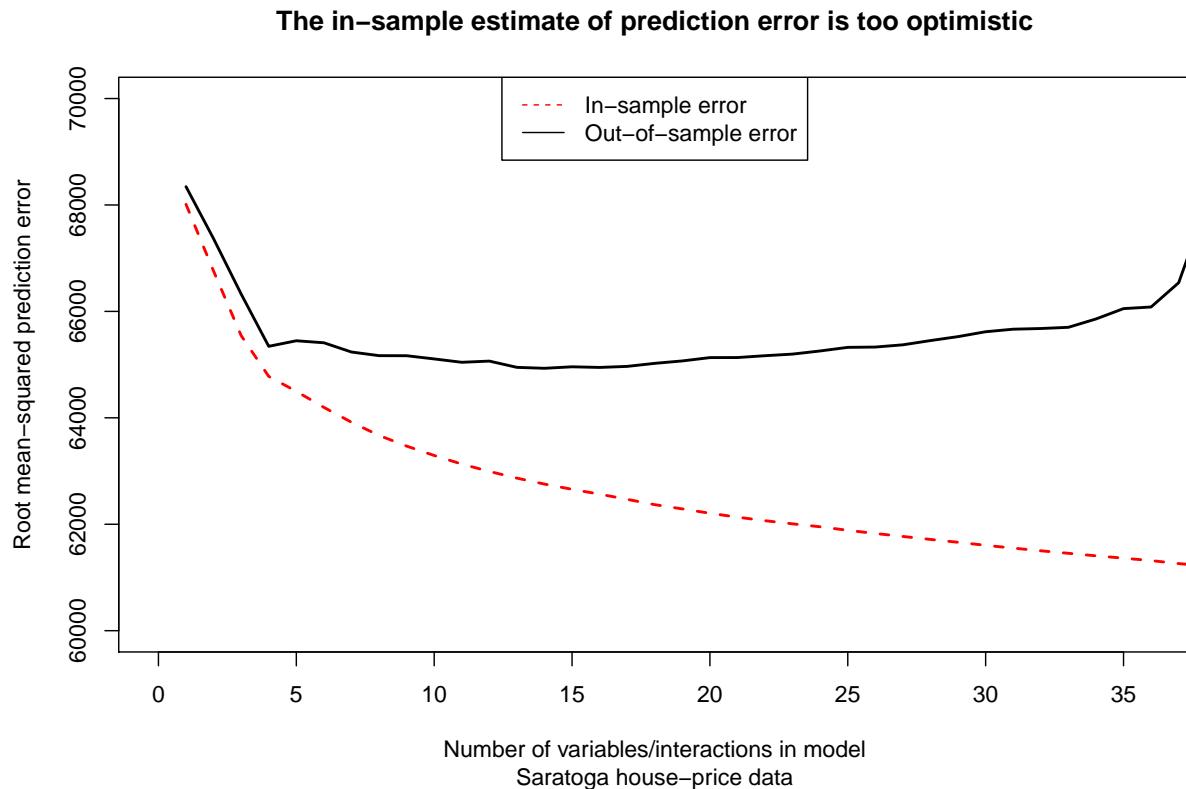
However, the *out-of-sample* measure of predictive error tells a different story. Here, the medium-sized model is clearly the winner. Its predictions on new data are off by about \$65,719, on average, which is nearly \$6,000 better than the big model.

Finally, notice how severely degraded the predictions of the big model become when moving from old (in-sample) data to new (out-of-sample) data: about \$9,600 worse, on average. This kind of degradation is a telltale sign of overfitting. The medium model suffers only a mild degradation in performance on new data, while the small model suffers hardly any degradation at all—although it’s still not competitive on the out-of-sample measure, because it wasn’t that good to begin with. This is also a special case of a more general phenomenon: *some* degradation in predictive performance on out-of-sample versus in-sample data is inevitable, but simpler models tend to degrade a lot less.

Figure 8.1 demonstrates this point visually. Starting from a very simple model of price (using only lot size as a predictor), we’ve added one variable or interaction at a time<sup>3</sup> from the list on page 168. For each new variable or interaction, we recalculated both the in-sample ( $\widehat{MSPE}_{in}$ ) and out-of-sample ( $\widehat{MSPE}_{out}$ ) estimates of the generalization error. As we add variables, the out-of-sample error initially gets smaller, reflecting a better fitting model that still generalizes well to new data. But after 15 or 20 variables,

Table 8.1: In-sample versus out-of-sample estimates of the root mean-squared predictive error for three models of house prices in Saratoga, NY. The “difference” column shows the difference between the in-sample and out-of-sample estimates. The big model has a very large difference (over \$9,000), indicating that the in-sample estimate is way too optimistic, and that the model is probably overfit to the data.

<sup>3</sup> To be specific here, at each stage we added the single variable or interaction that most improved the fit of the model. See the next section on stepwise selection.



eventually the out-of-sample error starts creeping back up, due to overfitting. The in-sample estimate of error, however, keeps going down, falling even further out of line with the real out-of-sample error as we add more variables to the model.

In summary, you should remember the basic mantra of predictive model building: out-of-sample error is larger than in-sample error, especially for bigger models. If you care about minimizing out-of-sample error, you should always use an out-of-sample estimate of a model's MSPE, to make sure that you're not overfitting the original data. Our goal here should be obvious: to find the “turning point” in Figure 8.1, and to stop adding variables before we start overfitting.

Figure 8.1: Starting from a small pricing model with just lot size as a predictor, we've added one variable or interaction at a time from the list on page 168. The red line shows the in-sample estimate of error, while the black line shows the out-of-sample estimate. After we add about 15 variables and interactions, the out-of-sample error starts to creep back up. Clearly the in-sample estimate is too optimistic, especially as the model gets more complex.

## Iterative model building via stepwise selection

Now that we know how to measure generalization error of a model, we're ready to introduce the overall steps in the process of building and using a predictive model from a set of candidate variables  $x_1, x_2$ , etc. We sometimes use the term *scope* to refer to this set of candidate variables.

The seemingly obvious approach is to fit all possible models under consideration to a training set, and to measure the generalization error of each one on a testing set. If you have only a few variables, this will work fine. For example, with only 2 variables, there are only  $2^2 = 4$  possible models to consider: the first variable in, the second variable in, both variables in, or both variables out. You can fit and test those four models in no time. This is called *exhaustive enumeration*.

However, if there are lots of variables, exhaustive enumeration of all the models becomes a lot harder to do, for the simple fact that it's too exhausting—there are too many models to consider. For example, suppose we have 10 possible variables, each of which we could put in or leave out of the model. Then there are  $2^{10} = 1024$  possible models to consider, since each variable could be in or out in any combination. That's painful enough. But if there are 100 possible variables, there are  $2^{100}$  possible models to consider. That's 1 *nonillion* models—about  $10^{30}$ , or a thousand billion billion billion. This number is larger than the number of atoms in a human body.

You will quite obviously never be able to fit all these countless billions of models, much less compare their generalization errors on a testing set, even with the most powerful computer on earth. Moreover, that's for just 100 candidate variables *with main effects only*. Ideally, we'd like the capacity to build a model using many more candidate variables than that, or to include the possibility of interactions among the variables.

Thus a more practical approach to model-building is *iterative*: that is, to start somewhere reasonable, and to make small changes to the model, one variable at a time. Model-building in this iterative way is really a three-step process:

- (1) Choose a baseline model, consisting of initial set of predictor variables to include in the model, including appropriate transformations, polynomial terms and interactions. Exploratory

data analysis (i.e. plotting your data) will generally help you get started here, in that it will reveal obvious relationships in the data. Then fit the model for  $y$  versus these initial predictors.

- (2) Check the model. If necessary, change what variables are included, what transformations are used, etc.:
  - (a) Are the assumptions of the model met? This is generally addressed using residual plots, of the kind shown in Figures 6.7 and 6.8. This allows you to assess whether the response varies linearly with the predictors, whether there are any drastic outliers, etc.
  - (b) Are we missing any important variables or interactions? This is generally addressed by *adding* candidate variables or interactions to the model from step (1), to see how much each one improves the generalization error (MSPE).
  - (c) Are there signs that the model might be overfitting the data? This is generally addressed by *deleting* variables or interactions that are already in the model, to see if doing so actually improves the model's generalization error.

You may need to iterate these three questions a few times, going through many rounds of adding or deleting variables, before you're satisfied with your final model. Remember that the best way to measure generalization error is using an out-of-sample measure, like  $\widehat{MSPE}_{out}$  derived from a train/test split of the data.

Once you're happy with the model itself, then you can. . . .

- (3) Use your fitted model to form predictions (and optionally, prediction intervals) for your new data points.

*Can this process be automated?*

In this three-step process, step 1 (start somewhere reasonable) and step 3 (use the final model) are usually pretty easy. The part where you'll spend the vast majority of your time and effort is step 2, when you consider many different possible variables to add or delete to the current model, and check how much they improve or degrade the generalization error of that model.

This is a lot easier than considering all possible combinations of variables in or out. But with lots of candidate variables, even this

iterative process can get super tedious. A natural question is, can it be automated?

The answer is: sort of. We can easily write a computer program that will automatically check for iterative improvements to some baseline (“working”) model, using an algorithm called *stepwise selection*:

- (1) From among a candidate set of variables (the scope), check all possible one-variable additions or deletions from the working model;
- (2) Choose the single addition or deletion that yields the best improvement to the model’s generalization error. This becomes the new “working model.”
- (3) Iteratively repeat steps (1) and (2) until no further improvement to the model is possible.

The algorithm terminates when it cannot find any one-variable additions or deletions that will improve the generalization error of the working model.

*Some caveats.* Stepwise selection tends to work tolerably well in practice. But it’s far from perfect, and there are some important caveats. Here are three; the first one is minor, but the second two are pretty major.

First, if you run stepwise selection from two different baseline models, you will probably end up with two different final models. This tends not to be a huge deal in practice, however, because the two final models usually have similar mean-squared predictive errors. Remember, when we’re using stepwise selection, we don’t care too much about *which* combinations of variables we pick, as long as we get good generalization error. Especially if the predictors are correlated with each other, one set of variables might be just as good as another set of similar (correlated) variables.

Second, stepwise selection usually involves some approximation. Specifically, at each step of stepwise selection, we have to compare the generalization errors of many possible models. Most statistical software will perform this comparison *not* by actually calculating  $\widehat{\text{MSPE}}_{\text{out}}$  on some test data, but rather using one of several possible heuristic approximations for MSPE. The most common one is called the AIC approximation:<sup>4</sup>

$$\widehat{\text{MSPE}}_{\text{AIC}} = \widehat{\text{MSPE}}_{\text{in}} \left( 1 + \frac{p}{n} \right) = s_e^2 \left( 1 + \frac{p}{n} \right),$$

<sup>4</sup> In case you’re curious, AIC stands for “Akaike information criterion.” If you find yourself reading about AIC on Wikipedia or somewhere similar, it will look absolutely nothing like the equation I’ve written here. The connection is via a related idea called “Mallows’  $C_p$  statistic,” which you can [read about here](#).

where  $n$  is the sample size and  $p$  is the number of parameters in the model.

The AIC estimate of mean-squared predictive error is not a true out-of-sample estimate, like  $\widehat{\text{MSPE}}_{\text{out}}$ . Rather, it is like an “inflated” or “penalized” version of the in-sample estimate,  $\widehat{\text{MSPE}}_{\text{in}} = s_e^2$ , which we know is too optimistic. The inflation factor of  $(1 + p/n)$  is always larger than 1, and so  $\widehat{\text{MSPE}}_{\text{AIC}}$  is always larger than  $\widehat{\text{MSPE}}_{\text{in}}$ . But the more parameters  $p$  you have relative to data points  $n$ , the larger the inflation factor gets. It’s important to emphasize that  $\widehat{\text{MSPE}}_{\text{AIC}}$  is just an approximation to  $\widehat{\text{MSPE}}_{\text{out}}$ . It’s a better approximation than  $\widehat{\text{MSPE}}_{\text{in}}$ , but it still relies upon some pretty specific mathematical assumptions that can easily be wrong in practice.

The third and most important caveat is that, when using any kind of automatic variable-selection procedure like stepwise selection, we lose the ability to use our eyes and our brains each step of the way. We can’t plot the residuals to check for outliers or violations of the model assumptions, and we can’t ensure that the combination of variables visited by the algorithm make any sense, substantively speaking. It’s worth keeping in mind that your eyes, your brain, and your computer are your three most powerful tools for statistical reasoning. In stepwise selection, you’re taking two of these tools out of the process, for the sake of doing a lot of brute-force calculations very quickly.

None of these caveats are meant to imply that you *shouldn’t* use stepwise selection—merely that you shouldn’t view the algorithm as having God-like powers for discerning the single best model, or treat it as an excuse to be careless. You should instead proceed cautiously. Always verify that the stepwise-selected model makes sense and doesn’t violate any crucial assumptions. It’s also a good idea to perform a quick train/test split of your data and compute  $\widehat{\text{MSPE}}_{\text{out}}$  for your final model, just as a sanity check, to make sure that you’re actually improving the generalization error versus your baseline model.

# 9

## *Understanding cause and effect*

### **Statistical questions versus causal questions**

WHY have some nations become rich while others have remained poor? Do small class sizes improve student achievement? Does following a Mediterranean diet rich in vegetables and olive oil reduce your risk of a heart attack? Does a “green” certification (like LEED, for [Leadership in Energy and Environmental Design](#)) improve the value of a commercial property?

Questions of cause and effect like these are, fundamentally, questions about *counterfactual statements*. A counterfactual is an if–then statement about something that has not actually occurred. For example: “If Colt McCoy had not been injured early in the [2010 National Championship football game](#), then the Texas Longhorns would have beaten Alabama.” If you judge this counterfactual statement to be true—and who but the most hopelessly blinkered Crimson Tide fan doesn’t?—then you might say that Colt McCoy’s injury caused the Longhorns’ defeat.

Statistical questions, on the other hand, are about correlations. This makes them fundamentally different from causal questions.

- Causal: “If we invested more money in our school system, how much faster would our economy grow?” Statistical: “In looking at data on a lot of countries, how are education spending and economic growth related?”
- Causal: “If I ate more vegetables than I do now, how much longer would I live?” Statistical: “Do people who eat a lot of vegetables live longer, on average, than people who don’t?”
- Causal: “If we hire extra teachers at our school and reduce our class sizes, will our students’ test scores improve?” Statistical: “Do students in smaller classes tend to have higher test scores?”

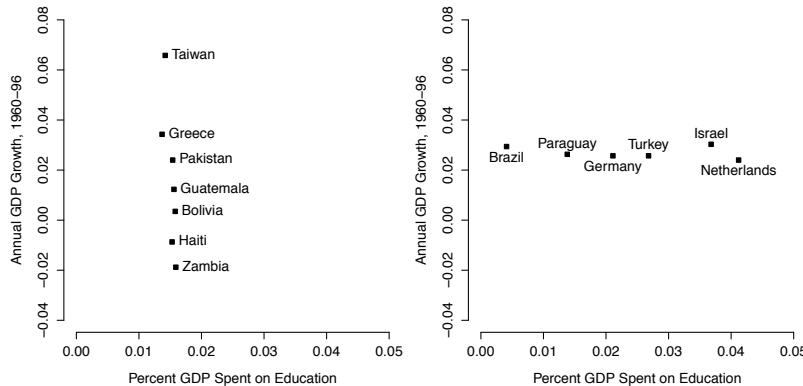


Figure 9.1: Two egregious examples of selective reporting.

Causal questions all invoke some kind of hypothetical intervention, where one thing is changed and everything else is held equal. In such a hypothetical intervention, there is no competing explanation for what might be causing the change we expect to see—in our economy, our lifespan, our students’ test scores, a football game, or whatever outcome we’re interested in.

Statistical questions, on the other hand, are about the patterns we observe in the real world. And the real world is rarely so simple as the hypothetical interventions we imagine. For example, people who eat more vegetables live longer—that’s a clear pattern. But those same people also tend to exercise more, live in better housing, and have higher-status jobs. These other factors are *confounders*. A confounder is a competing explanation—some other factor correlated with both the “treatment” assignment (whether someone eats vegetables) and the response (lifespan). So in light of these confounders, how do we know it’s the vegetables, rather than all that other stuff, that’s making veggie-eaters live longer?

This is just a specific version of the general question we’ll address in this chapter: under what circumstances can causal questions be answered using statistics?

### *Good evidence . . . and bad*

Most of the cause-and-effect reasoning that you’ll see out there in the real world is of depressingly poor quality. A common flaw is *cherry picking*: that is, pointing to data that seems to confirm some argument, while ignoring contradictory data.

Here’s an example. In the left panel in Figure 9.1 we see a

group of seven countries that all spend around 1.5% of their GDP on education, but with very different rates of economic growth for the 37 years spanning 1960 to 1996. In the right panel, we see another group of six countries with very different levels of spending on education, but similar growth rates of 2–3%.

Both highly selective samples make it seem as though education and economic growth are barely related. If presented with the left panel alone, you'd be apt to conclude that the differences in growth rates must have been caused by something other than differences in education spending (of which there are none). Likewise, if presented with the right panel alone, you'd be apt to conclude that the large observed differences in education spending don't seem to have produced any difference in growth rates. The problem here isn't with the data—it's with the biased, highly selective *use* of that data.

This point seems almost obvious. Yet how tempting it is just to cherry pick and ignore the messy reality. Perhaps without even realizing it, we're all accustomed to seeing news stories that marshal highly selective evidence—usually even worse than that of Figure 9.1—on behalf of some plausible because-I-said-so story:

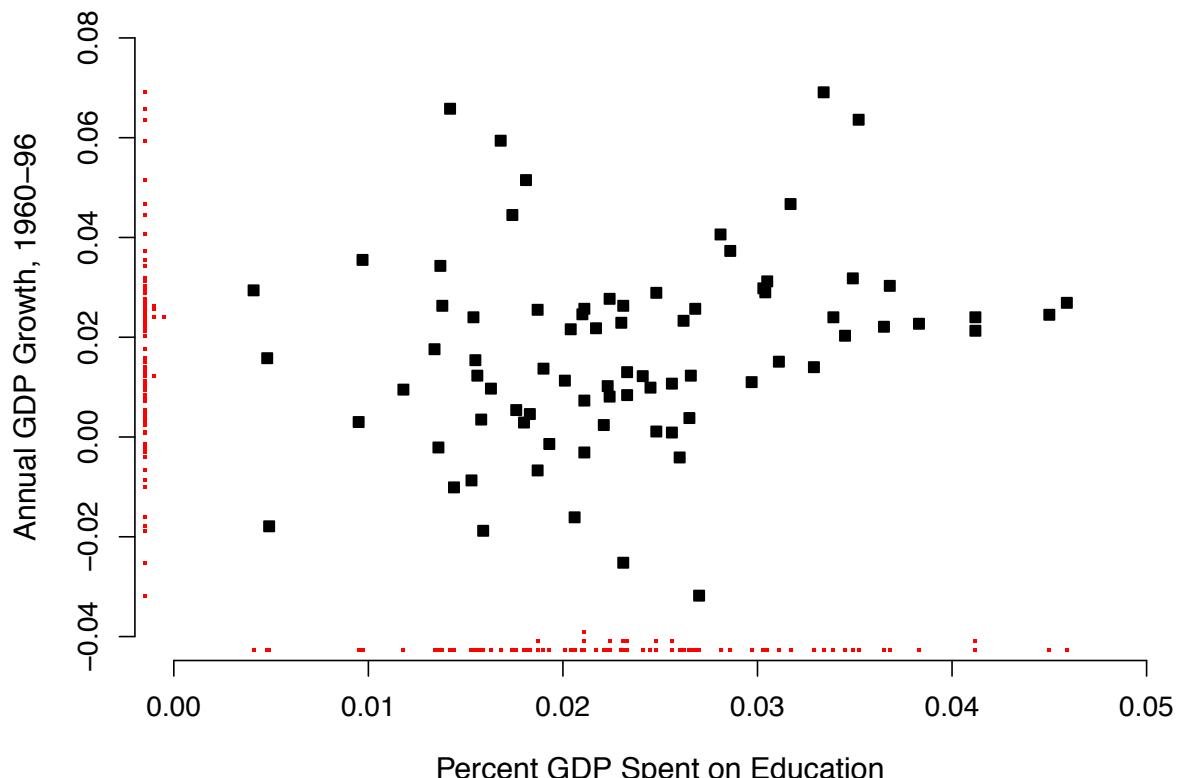
[H]igher levels of education are critical to economic growth. . . . Boston, where there is a high proportion of college graduates, is the perfect example. Well-educated people can react more quickly to technological changes and learn new skills more readily. Even without the climate advantages of a city like San Jose, California, Boston evolved into what we now think of as an “information city.” By comparison, Detroit, with lower levels of education, languished.<sup>1</sup>

And this from a reporter who presumably has no hidden agenda. Notice how the selective reporting of evidence—one causal hypothesis, two data points—lends an air of such graceful inevitability to what is a startlingly superficial analysis of the diverging economic fates of Boston and Detroit over the last half century.

Of course, most bad arguments are harder to detect than this howler from the New York Times. After all, using data to understand cause-and-effect relationships is hard. For example, consider the following summary of a recent neuroscience study:

A study presented at the Society for Neuroscience meeting, in San Diego last week, shows people who start using marijuana at a young age have more cognitive shortfalls. Also, the more marijuana a person used in adolescence, the more trouble they had with focus and attention. “Early onset smokers

<sup>1</sup> “Economic Scene.” *New York Times* (Business section); August 5, 2004



have a different pattern of brain activity, plus got far fewer correct answers in a row and made way more errors on certain cognitive tests," says study author Staci Gruber.<sup>2</sup>

Did the marijuana smokers get less smart, or were the less-smart kids more likely to pick up a marijuana habit in the first place? It's an important question to consider in making drug policy, especially for states and countries where marijuana is legal. But can we know the answer on the basis of a study like this?

For another example, consider the bigger sample of countries in Figure 9.2, which provides a much more representative body of evidence on the GDP-versus-education story. This evidence takes the form of a scatter plot of GDP growth versus education spending for a sample of 79 countries worldwide. Notice the following two facts:

- (1) Of the 29 countries that spent less than 2% of GDP on education, 18 fall below the median growth rate (1.58%).

Figure 9.2: A scatter plot of GDP growth versus education spending for 79 countries. The tiny red dots clustered near the  $x$  and  $y$  axes are called *rug plots*. They are miniature histograms aligned with the axes of the predictor and the response.

<sup>2</sup> [www.usatoday.com/yourlife/health/medical/pediatrics/2010-11-20-teendrugs22\\_ST\\_N.htm](http://www.usatoday.com/yourlife/health/medical/pediatrics/2010-11-20-teendrugs22_ST_N.htm)

- (2) Of the 18 countries that spent more than 3% of GDP on education, 16 fall above the median growth rate.

These two facts, together with the upward trend in the scatter plot, suggest that economic growth and education spending are correlated. But this does not settle the causal question. For example, it might be that countries spend a lot on education because they are rich, rather the other way around.

The generic difficulty is that there are many different ways that two variables  $X$  and  $Y$  can appear correlated.

(1) *One-way causality*: the first domino falls, then the second; the rain falls, and the grass gets wet. ( $X$  causes  $Y$  directly.)

(2) *Two-way causality*: flowers and honey bees prosper together.  
(Both  $X$  and  $Y$  play a role in causing each other.)

(3) *Common cause*: People who go to college tend to get higher-paying jobs than those who don't. Does education directly lead to better economic outcomes? Or are a good education and a good job both just markers of a person's underlying qualities? (The role of  $X$  in causing  $Y$  is hard to distinguish from the role of  $C$ , which we may not have observed.)

(4) *Common effect*: either musical talent ( $X$ ) or athletic talent ( $Y$ ) will help you get into Harvard ( $Z$ ). Among a population of Harvard freshmen, musical and athletic talent will thus appear negatively correlated, even if they are independent in the wider population. ( $X$  and  $Y$  both contribute to some common outcome  $C$ , inducing a correlation among a subset of the population defined by  $Z$ . This is often called Berkson's paradox; it is subtle, and we'll encounter it again.)

(5) *Luck*: the observed correlation is a coincidence.

This is the point where most books remind you that “correlation does not imply causation.” Obviously. But if not to illuminate causes, what is the point of looking for correlations? Of course correlation does not imply causality, or else playing professional basketball would make you tall. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observed correlations. The important question is: what distinguishes the good evidence-based arguments from the bad?

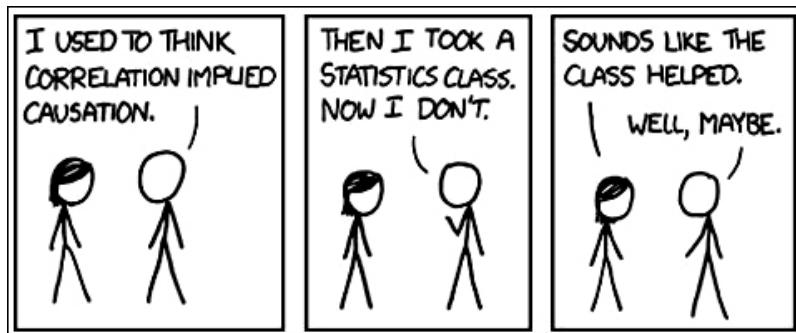


Figure 9.3: Originally published online at xkcd: <http://xkcd.com/552/>

#### *Four common identification strategies*

The key principle in using evidence to draw causal conclusions is that of a *balanced comparison*. To make things simple, we'll imagine that our predictor  $X$  is binary (i.e. has two groups), and we'll borrow the lingo of a clinical trial by referring to the two groups as the "treatment" and "control." To reach the conclusion that  $X$  causes  $Y$ , you must do two things: (1) *compare cases* in the treatment and control groups, to see how their  $Y$  values differ; and (2) *ensure balance*, by removing all other systematic differences between the cases in the treatment and control groups. Balance is crucial; it's what allows us to conclude that the differences in  $X$  (and not something else) cause the differences we observe in  $Y$ .

In general, there are four common ways to make a balanced comparison. These are often called *identification strategies*, in the sense that they are strategies for identifying a causal effect.

- (1) *Run a real experiment*, randomizing subjects to the treatment and control groups. The randomization will ensure that, on average, there are no systematic differences between the two groups, other than the treatment.
- (2) *Find a natural experiment*: that is, find a situation where the way that cases fall naturally into the treatment and control groups plausibly resembles a random assignment.
- (3) *Matching*: artificially construct a balanced data set from an unbalanced data set, by explicitly matching treated cases with similar control cases, and discarding the cases without a good match. This will correct for lack of balance between control and treatment groups.

(4) *Modeling*: use multiple regression modeling to adjust for confounders and isolate a partial relationship between the response and the treatment of interest.

We'll take each of these four ideas in turn.

## The power of experiment

THE idea of an experiment is simple. If you want to know what would happen if you intervened in some system, then you should intervene, and measure what happens. There is simply no better way to establish that one thing causes another.

Indeed, one kind of experiment—the randomized, controlled clinical trial—is one of the most important medical innovations in history. Suppose we want to establish whether a brand new cholesterol drug—we'll call it Zapaclot—works better than the old drug. Also suppose that we've successfully recruited a large cohort of patients with high cholesterol. We know that diet and genes play a role here, but that drugs can help, too. We express this as

$$\text{Cholesterol} \sim \text{Diet} + \text{Genes} + \text{Drugs}.$$

Interpret the plus sign as the word “and,” not like formal addition: we're assuming that cholesterol depends upon diet, genes, and drugs, although we haven't said how. Of course, it's that third predictor in the model we care about; the first two, in addition to some others that we haven't listed, are potential confounders.

First, what not to do: don't proceed by giving Zapaclot to all the men and the old drug to all the women, or Zapaclot to all the marathon runners and the old drug to the couch potatoes. These highly non-random assignments would obviously bias any judgment about the relative effect of the new drug compared to the old one. We refer to this sort of thing as *selection bias*: that is, any bias in the selection of cases that receive the treatment. Moreover, you shouldn't just give the new drug to whomever wants it, or can afford it. The people with more engagement, more knowledge, more money, or more trust in the medical system would probably sign up in greater numbers—and if those people have systematic differences in diet or genes from the people who don't sign up, then you've just created a hidden selection bias.

Instead, you should two simple steps.

*Randomize:* randomly split the cohort into two groups, denoted the treatment group and the control group.

*Intervene:* allocate everyone in the treatment group to take the treatment (e.g. Zapaclot, the new drug), and everyone in the control group to take something else (e.g. the old drug or a placebo).<sup>3</sup>

Randomize and intervene: a simple prescription, but the surest way to establish causality. The intervention allows you to pick up a difference between the new and old drug, if there's one to be found. The randomization ensures that other factors—even unknown factors, in addition to known ones like diet and lifestyle—do not lead us astray in our causal reasoning. The Latin phrase *ceteris paribus*, which translates roughly as “everything else being equal,” is often used to describe such a situation. By randomizing and intervening, we have ensured that the only *systematic* difference between the groups is the treatment itself. The randomization gives us a balanced comparison.

This last point is crucial. It's not that diet, genes, and other lifestyle factors somehow stop affecting a patient's cholesterol level when we randomize and intervene. It's just that diet, genes, and lifestyle factors aren't correlated with the treatment assignment, and so they're balanced between the two groups, on average.

The need to avoid selection bias sounds obvious. But if selection bias in medical trials were not rigorously policed, then it would be easy for doctors to cherry pick healthy patients for newly proposed treatments. After all, a doctor who invents a new, seemingly effective form of treatment will almost surely become both rich and famous. As one physician reminisces:

One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, “Do you have any controls?” Well, the great surgeon drew himself up to his full height, hit the desk, and said, “Do you mean did I not operate on half of the patients?” The hall grew very quiet then. The voice at the back of the room very hesitantly replied, “Yes, that's what I had in mind.” Then the visitor's fist really came down as he thundered, “Of course not. That would have doomed half of them to their death.” God, it was quiet then, and one could scarcely hear the small voice ask, “Which half?”<sup>4</sup>

<sup>3</sup> Everyone in the control group should be taking the *same* something else, whether it's the old drug or a placebo.

<sup>4</sup> Dr. E. Peacock, University of Arizona. Originally quoted in *Medical World News* (September 1, 1972). Reprinted pg. 144 of *Beautiful Evidence*, Edward Tufte (Graphics Press, 2006).

These last two words—"Which half?"—should echo in your mind whenever you are asked to judge the quality of evidence offered in support of a causal hypothesis. There is simply no substitute for a controlled experiment: not a booming authoritative voice, not even fancy statistics.

In fact, government regulators are so fastidious in their attention to possible selection biases that, in most real clinical trials, neither the doctors nor the patients are allowed to know which drug each person receives. Such a "double-blind" experiment avoids the possibility that patients might simply imagine that the latest miracle drug has made them feel better, in a feat of unconscious self-deception called the placebo effect.

A placebo, from the Latin *placere* ("to please"), is a fake treatment designed to simulate the real one.

### *Some history*

The notion of a controlled experiment was certainly around in pre-Christian times. The first chapter of the book of Daniel relates the tale of one such experiment. Daniel and his three friends Hananiah, Mishael, and Azariah arrive in the court of Nebuchadnezzar, the King of Babylon. They enroll in a Babylonian school, and are offered a traditional Babylonian diet. But Daniel wishes not to "defile himself with the portion of the king's meat, nor with the wine which he drank." He goes to Melzar, the prince of the eunuchs, who is in charge of the school. Daniel asks not to be made to eat the meat or drink the wine. But Melzar responds that he fears for Daniel's health if he were to let them follow some crank new-age diet. More to the point, Melzar observes, if the new students were to fall ill, "then shall ye make me endanger my head to the king."

So Daniel proposes a trial straight out of a statistics textbook:

Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink.

Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants.<sup>5</sup>

<sup>5</sup> King James Bible, Daniel 1:12–13.

The King agreed. When Daniel and his friends were inspected ten days later, "their countenances appeared fairer and fatter in flesh" than all those who had eaten meat and drank wine. Suitably impressed, Nebuchadnezzar brings Daniel and his friends in for an audience, and he finds that "in all matters of wisdom and understanding," they were "ten times better than all the magicians and astrologers that were in all his realm."

As for a placebo-controlled trial, in which some of the patients are intentionally given a useless treatment (the “placebo”): that came much later.<sup>6</sup> The first such trial seems to have taken place in 1784. It was directed by none other than Benjamin Franklin, the American ambassador to the court of King Louis XVI of France. A German doctor by the name of Franz Mesmer had gained some degree of notoriety in Europe for his claim to have discovered a new force of nature that he called “magnétisme animal,” and which was said to have magical healing powers. The demand for Dr. Mesmer’s services soon took off among the ladies of Parisian high society, whom he would “Mesmerize” using a wild contraption involving ropes and magnetized iron rods.

Much to the king’s dismay, his own wife, Marie Antoinette, was one of Mesmer’s keenest followers. The king found the whole Mesmerizing thing frankly a bit dubious, and presumably wished for his wife to have nothing to do with the Herr Doctor’s magnétisme animal. So he convened several members of the French Academy of Sciences to investigate whether Dr. Mesmer had indeed discovered a new force of nature. The panel included Antoine Lavoisier, the father of modern chemistry, along with Joseph Guillotin, whose own wild contraption was soon to put the King’s difficulties with Mesmer into perspective. Under Ben Franklin’s supervision, the scientists set up an experiment to replicate some of Dr. Mesmer’s prescribed treatments, substituting non-magnetic materials—history’s first placebo—for half of the patients. In many cases, even the patients in the control group would flail about and start talking in tongues anyway. The panel concluded that the doctor’s method produced no effect other than in the patients’ own minds. Mesmer was denounced as a charlatan, although he continues to exact his revenge via the dictionary.

A more recent and especially striking example of a placebo comes from Thomas Freeman, director of the neural reconstruction unit at Tampa General Hospital in Florida. Dr. Freeman performs placebo brain surgery. (You read that correctly.) According to the British Medical Journal,

In the placebo surgery that he performs, Dr Freeman bores into a patient’s skull, but does not implant any of the fetal nerve cells being studied as a treatment for Parkinson’s disease. The theory is that such cells can regenerate brain cells in patients with the disease. Some colleagues decry the experimental method, however, saying that it is too risky and unethical, even though patients are told before the operation

<sup>6</sup> See “The Power of Nothing” in the December 12, 2011 edition of *The New Yorker* (pp. 30–6).

that they may or may not receive the actual treatment.<sup>7</sup>

<sup>7</sup> BMJ. 1999 October 9; 319(7215): 942

"There has been a virtual taboo of putting a patient through an imitation surgery," Dr. Freeman said. (Imagine that.) "This is the way to start the discussion." Freeman has performed 106 real and placebo cell transplant operations since 1992. Dr. Freeman argues that the medical history is littered with examples of unsafe and ineffective surgical procedures—think of that small voice at the back of the room, asking "which half?"—that were not tested against a placebo and resulted in needless deaths, year after year, before doctors abandoned them.

### *Experimental evidence is the best kind of evidence*

Let's practice here, by comparing two causal hypotheses arising from two different data sets. The first comes from a clinical trial in the 1980's on a then-new form of adjuvant chemotherapy for treating colorectal cancer, a dreadful disease that, as of 2015, has a five-year survival rate of only 60-70% in the developed world.

The trial followed a simple protocol. After surgical removal of their tumors, patients were randomly assigned to different treatment regimes. Some patients were treated with fluorouracil (the chemotherapy drug, also called 5-FU), while others received no follow-up therapy. The researchers followed the patients for many years afterwards and tracked which ones suffered from a recurrence of colorectal cancer.

The outcome of the trial are in Table 9.1, below. Among the patients who received chemotherapy, 39% (119/304) had relapsed by the end of the study period, compared with 57% of patients (177/315) in the group who received no therapy:

Chemotherapy?	Yes	No
Recurrence?	Yes	119 177
	No	185 138

The evidence strongly suggests that the chemotherapy reduced the risk of recurrence by a substantial amount: the relative risk of a relapse under the treatment group is 0.7, with a 95% confidence interval of (0.59, 0.83).

We can be confident that this evidence reflects causality, and not merely correlation, because patients were randomly assigned

Table 9.1: Data from: J. A. Laurie et. al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. J. Clinical Oncology, 7:1447–56, 1989. There was also a third treatment arm of the study in which patient received a drug called levamisole, which isn't discussed here. Survival statistics on colorectal cancer from Cunningham et. al (2010). "Colorectal cancer." Lancet 375 (9719): 1030–47.

to the treatment and control groups. Randomization ensures *balance*: that is, it ensures that there are no systematic differences between the two groups with respect to any confounding factors that might be correlated with the patients' survival chances. This would obviously not be true if we had non-randomly assigned all the healthiest patients to the treatment group, and all the sickest patients to the control group.

It's worth emphasizing a key fact here. Randomization ensures balance both for the possible confounders that we can measure (like a patient's age or baseline health status), as well as for the ones we might *not* be able to measure (like a patient's will to live). This is what makes randomization so powerful, and randomized experiments so compelling. We don't even have to know what the possible confounding variables are in order for the experiment to give us reliable information about the causal effect of the treatment. *Randomization balances everything*, at least on average.

Next, let's examine data from a study from the 1990's conducted in sub-Saharan Africa about HIV, another dreadful disease which, at the time, was spreading across the continent with alarming speed. Several studies in Kenya had found that men who were uncircumcised seemed to contract HIV in greater numbers. This set off a debate among medical experts about the extent to which this apparent association had a plausible biological explanation.

Circumcised?	Yes	No
HIV positive?	Yes	105
	No	527
		85
		93

Table 9.2, above, shows some data from one of these studies, which found that among those recruited for the survey, 48% of uncircumcised men were HIV-positive, versus only 17% of circumcised men. The evidence seems to suggest that circumcision reduced a Kenyan man's chance of contracting HIV by a factor of 3.

Table 9.2: Data from Tyndall et. al. Increased risk of infection with human immunodeficiency virus type 1 among uncircumcised men presenting with genital ulcer disease in Kenya. Clin. Infect. Dis. 1996 Sep; 23(3):449–53.

*Evaluating the evidence.* If you suffer from colon cancer, should you get chemotherapy? Almost certainly: the researchers in the first study randomized and intervened, giving chemotherapy only to a random subset of patients. Unless you believe that the chemotherapy patients in this trial just happened to be much luckier than their peers, this result establishes that the reduction in recurrence must have been caused by the treatment.

But should all Kenyan men head straight to a surgeon? In this case we can't really be sure. The researchers in the second study neither randomized nor performed any snipping themselves. They merely asked whether each man was circumcised. It is therefore possible that they've been fooled by a confounder. To give one plausible example, a man's religious affiliation might affect both the likelihood that he is circumcised and the chances that he contracts HIV from unprotected sex. If that were true, the observed correlation between circumcisions and HIV rates might be simply a byproduct of an imbalanced, unfair comparison, rather than a causal relationship.<sup>8</sup>

## Natural experiments

A randomized, controlled experiment is the gold standard of evidence for a causal hypothesis. Yet many times an experiment is impossible, impractical, unethical, or too expensive in time or money. In these situations, it often pays to look for something called a *natural experiment*, also called a *quasi-experiment*. A natural experiment is not something that you, as the investigator, design. Rather, it is an "experiment" where nature seems to have done the randomization and intervention for you, thereby giving you the same type of balance between treatment and control groups that you'd expect to get out of a real experiment.

This idea is best understood by example. Suppose you want to study the effect of class size on student achievement. You reason that, in smaller classes, students can get more individual attention from the instructor, and that instructors will feel a greater sense of personal connection to their students. All else being equal, you believe that smaller class sizes will help students learn better.

A cheap, naïve way to study this question would be to compare the test scores of students in small classes to those of students in larger classes. Any of these confounders, however, might render such a comparison highly unbalanced, and therefore dubious: (1) students in need of remediation are sometimes put in very small classes; (2) highly gifted students are also sometimes put in very small classes; (3) richer school districts can afford both smaller classes and many other potential sources of instructional advantage; or (4) better teachers successfully convince their bosses to let them teach the smaller classes themselves.

An expensive, intelligent way to study this question would

<sup>8</sup> The authors of the study were obviously aware of these possible confounders. They used a technique called logistic regression to attempt to account for some them and isolate the putative effect of circumcision on HIV infection. This is like our fourth method for making balanced comparisons: use a model to adjust for confounders statistically. See the original paper for details.

Question	Problem	Natural experiment	Lingering issues
Does being rich make people happy?	Even if richer people are happier on average, maybe happiness and success are the common effect of a third factor. Or maybe the rich grade on a different curve than the rest of us.	Compare a group of lottery winners with a similar group of people who played the lottery but didn't win.	Lottery winners may play the lottery far more often than people who played the lottery but didn't win, which might correlate with other important differences.
Does smoking increase a person's risk for Type-II diabetes?	People who smoke may also engage in other unhealthy behaviors at systematically different rates than non-smokers.	Compare before-and-after rates of diabetes in cities that recently enacted bans on smoking in public places.	Maybe the incidence of diabetes would have changed anyway.
Do bans on mobile phone use by drivers in school zones reduce the rate of traffic collisions?	Groups of citizens that enact such bans may differ systematically in their attitudes toward risk and behavior on the road.	Go to Texarkana, split by State Line Avenue. Observe what happens when Texas passes a ban and Arkansas doesn't.	There may still be systematic differences between the two halves of the city.

Table 9.3: Three hypothetical examples of natural experiments.

be to design an experiment, in conjunction with a scientifically inclined school district, that randomly assigned both teachers and students to classes of varying size. In fact, a few school systems have done exactly this. A notable experiment is Project STAR in Tennessee—an expensive, lengthy experiment that studied the effect of primary-school class sizes on high-school achievement, and showed that reduced class sizes have a long-term positive impact both on test scores and drop-out rates.<sup>9</sup>

But suppose you are neither naïve nor rich, and yet still want to study the question of whether small class sizes improve test scores. If you're in search of a third way—one that's better than merely looking at correlations, yet cheaper than a full-fledged experiment—you might be interested to know the following fact about the Israeli school system.

[I]n Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."<sup>10</sup>

This is a lovely example of a natural experiment—something you didn't design yourself, but that is almost as good as if you

<sup>9</sup> The original study is described in Finn and Achilles (1990). "Answers and Questions about Class Size: a Statewide Experiment." *American Educational Research Journal* 28, pp. 557–77

<sup>10</sup> Angrist and Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press, p. 21

had. The researchers in this study compared the students in a group of 40 (“control group,” in one large class) versus the students in a group of 41 (“treatment group,” split into two smaller classes). This is a plausibly random assignment: the “randomization mechanism” is whether a student fell into a peer group of 40 versus a peer group of 41, and we would not expect this difference to be confounded by anything else that might predict test scores. Therefore, if we see a big difference in performance between the two groups, the most likely explanation is that class size caused the difference.

Some natural experiments, of course, are better than others. Consider the examples in Table 9.3, on page 188. For each one, ask yourself two questions. (1) What are the “treatment” and “control” groups? (2) How balanced are these two groups? (Said another way: how good is the quasi-randomization of cases to these groups?) Think carefully about each one, and you may begin to see “experiment” versus “non-experiment” as the black and white ends of a spectrum, with many shades of grey in between.

## Matching

To estimate a causal effect by matching, we artificially construct a balanced data set out of an unbalanced one, by explicitly matching treated cases with similar control cases. We then compare the outcomes in treatment versus control groups, using only the balanced data set. This is most easily seen by example.

### *An example: the value of going green*

For many years now, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious (“green”) buildings. There are both ethical and economic forces at work here. To quote a recent report by Mercer, an investment-consulting firm, entitled “Energy efficiency and real estate: Opportunities for investors”:

Investing in energy efficiency has two intertwined virtues that make it particularly attractive in a world with a changing climate and a destabilized economy: It cuts global-warming greenhouse gas emissions and saves money by reducing energy consumption. Given that the built environment accounts for 39 percent of total energy use in the US and 38 percent of total indirect CO<sub>2</sub> emissions, real estate investment represents

one of the most effective avenues for implementing energy efficiency.

This only scratches the surface. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. Every new project involves negotiating a trade-off between costs incurred and benefits realized over the lifetime of the building. In this context, the decision to invest in an eco-friendly building could pay off in at least four ways.

- (1) Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.
- (2) Green buildings are often associated with indoor environments that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify *ex ante*; you cannot simply ask an engineer in the same way that you could ask a question such as, “How much are these solar panels likely to save on the power bill?”
- (3) Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.
- (4) Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is “life-cycle analysis,” which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for

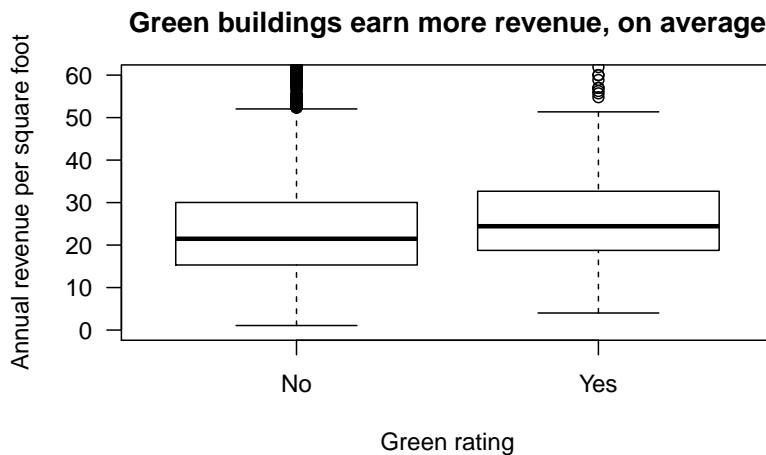


Figure 9.4: Green buildings seem to earn more revenue per square foot, on average, than non-green buildings.

rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market and comparing “green” versus “non-green” buildings. By “green,” we mean that a commercial property has received some official certification, because its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks, as certified by outside engineers.<sup>11</sup>

Let’s look at some data on 678 green-certified buildings in the United States, together with 6,298 non-green buildings in similar geographic areas. The boxplot above shows that, when we measure revenue by a building’s rental rate per square foot per year, green buildings tend to earn noticeably higher revenue (mean = 26.97) than non-green buildings (mean = 24.51). That’s a difference of \$2.46 per square foot, or nearly a 10% market premium.

<sup>11</sup> The two most common certifications are LEED and EnergyStar; you can easily find out more about these rating systems on the web, e.g. at [www.usgbc.org](http://www.usgbc.org).

Original data		
	Non-green buildings	Green buildings
Sample size	6928	678
Mean revenue/sq ft.	24.51	26.97
Mean age	49.2	23.9
Class A	37%	80%
Class B	48%	19%
Class C	15%	1%

Table 9.4: Covariate balance for the original data. Class A, B, and C are relative classifications within a specific real-estate market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

However, there's a problem with this comparison. As Table 9.4 shows, the green buildings tend to be newer than the non-green buildings, and are more likely to be "Class A" buildings.

So the important question is: do green buildings command a market premium *because* they are green, or simply because they are newer, better buildings in the first place? We can't tell by simply computing the average revenue in each group, because the green ("treatment") and non-green ("control") groups are highly unbalanced with respect to some important confounders.

This is where matching comes in. Matching means constructing a balanced data set from an unbalanced one. It involves three steps:

- (1) For each case in the treatment group, find the case in the control group that is the closest match in terms of confounding variables, and pair them up. Put these matched pairs into a new matched data set, and discard the cases in the original data set for which there are no close matches.
- (2) Verify covariate balance for the matched data set, by checking that the confounders are well balanced between the treatment and control groups.
- (3) Assuming that the confounders are approximately balanced, then compare the treatment-group outcomes with the control-group outcomes, using *only* the matched pairs.

Matching relies on a simple principle: compare like with like. In this example, that means if we have a 25-year-old, Class A building with a green rating, we try to find another 25-year old, Class A building without a green rating to compare it to.

In this particular example, once we've constructed the data set of matched pairs, the confounder variables are much more closely

		Matched data	
	Non-green buildings	Green buildings	
Sample size	678	678	
Mean revenue/sq ft.	25.94	26.97	
Mean age	23.9	23.9	
Class A	80%	80%	
Class B	19%	19%	
Class C	1%	1%	

Table 9.5: Covariate balance for the matched data.

balanced between the treatment and control groups (see Table 9.5). A comparison of revenue rates for this matched data set makes the premium for green buildings look a lot smaller: \$26.97 versus \$25.94, or about a 4% premium. Compare that with the 10% green premium we estimated from the original, unmatched data.

*How do we actually find matches?* The nitty-gritty algorithmic details of actually finding good matched pairs of cases are best left to the experts who write the software for these things. The two most common types of matching are called *nearest-neighbor search* and *propensity-score matching*; follow the links if you'd like to know more. In R, the package MatchIt uses propensity-score matching as a default; this is a very commonly used algorithm in real-world data analysis. In addition, [the paper linked here<sup>12</sup>](#) has a much more detailed overview of different matching methods.

#### *Matching isn't a silver bullet: a bigger example*

If you've ever been admitted to the intensive-care unit at a hospital, you may have undergone a diagnostic procedure called *right heart catheterization*, or RHC. RHC is used to see how well a patient's heart is pumping, and to measure the pressures in that patient's heart and lungs. RHC is widely believed to be helpful, since it allows the doctor to directly measure what's going on inside a patient's heart. But it is an invasive procedure, since it involves inserting a small tube (the catheter) into the right side of your heart, and then passing that tube through into your pulmonary artery. It therefore poses some risks—for example, excessive bleeding, partial collapse of a lung, or infection.

A natural question is: do the diagnostic benefits of RHC outweigh the possible risks? But this turns out to be tricky to answer. The reason is that doctors would not consider it ethical to run a randomized, controlled trial to see if RHC improves patient outcomes. As the authors of one famous study from the 1990s pointed out:<sup>13</sup>

Many cardiologists and critical care physicians believe that the direct measurement of cardiac function provided by right heart catheterization (RHC) . . . is necessary to guide therapy for certain critically ill patients, and that such management leads to better patient outcomes. While the benefit of RHC has not been demonstrated in a randomized controlled trial (RCT), the popularity of this procedure, and the widespread

<sup>12</sup> "Matching Methods for Causal Inference: A Review and a Look Forward." Elizabeth A. Stuart, *Statistical Science*, 2010.

<sup>13</sup> "The effectiveness of right heart catheterization in the initial care of critically ill patients." Connors et. al. *Journal of the American Medical Association*. 1996 Sep 18; 276(11):889-97.

	Original data		Matched data	
	No RHC	RHC	No RHC	RHC
Sample size	3551	2184	2184	2184
180-day survival rate	0.370	0.320	0.354	0.320
mean APACHE score	50.934	60.739	57.643	60.739
Trauma	0.005	0.016	0.008	0.016
Heart attack	0.030	0.043	0.036	0.043
Congestive heart failure	0.168	0.195	0.209	0.195
Sepsis	0.148	0.321	0.24	0.321

belief that it is beneficial, make the performance of an RCT difficult. Physicians cannot ethically participate in such a trial or encourage a patient to participate if convinced the procedure is truly beneficial.

We're therefore left with only observational data on the effectiveness of RHC—which, on the surface, doesn't look good! Here's the data from the study quoted above, showing that critically ill patients undergoing RHC actually have a *worse* 180-day survival rate (698/2184, or 32%) than patients not undergoing RHC (1315/3551, or 37%):

	No RHC	RHC
Survived 180 days	1315	698
Died within 180 days	2236	1486

What's going on here? Should we conclude that right heart catheterization is actually killing people, and that the doctors are all just plain wrong about its putative benefits?

Not so fast. The problem with this conclusion is that the treatment (RHC) and control (no RHC) groups are heavily unbalanced with respect to baseline measures of health. Put simply, the patients who received RHC were a lot sicker to begin with, so it's no surprise that they have a lower 6-month survival rate. To cite a few examples: the RHC patients were three times more likely to have suffered acute trauma, 50% more likely to have had a heart attack, and 16% more likely to be suffering from congestive heart failure. The RHC patients also had an average **APACHE score** that was 10 points higher than the non-RHC patients.<sup>14</sup> The left half of Table

Table 9.6: A before-and-after table of summary statistics showing covariate balance for the observational study on right-heart catheterization. The entries for trauma, heart attack, etc. show rates of these complications in the two groups. The left half of the table shows the original data set, while the right half shows the matched data set.

<sup>14</sup> The APACHE score is a composite severity-of-disease score used by hospital ICUs to estimate which patients have a higher risk of death. Patients with higher numbers have a higher risk of death.

[9.6](#) shows these rates of various complications for the two groups in the original data set. They're quite different, implying that the survival rates of these two groups cannot be fairly compared.

And what about after matching? Unfortunately, Table [9.6](#) shows that, even after matching treatment cases with controls having similar complications, the RHC group still seems to have a lower survival rate. The gap looks smaller than it did before, on the unmatched data—a 32% survival rate for RHC patients, versus a 35.4% survival rate for non-RHC patients—but it's still there.

Again we find ourselves asking: what's going on? Is the RHC procedure actually killing patients? Well, it might be, at least indirectly! The authors of the study speculate that one possible explanation for this finding is “that RHC is a marker for an aggressive or invasive style of care that may be responsible for a higher mortality rate.” Given the prevalence of [overtreatment](#) within the American health-care system, this is certainly plausible.

But we can't immediately jump to that conclusion on the basis of the matched data. In fact, this example points to a couple of basic difficulties with using matching to estimate a causal effect.

The first (and most important) difficulty is that *we can't match on what we haven't measured*. If there is some confounder that we don't know about, then we'll never be able to make sure that it's balanced between the treatment and control groups within the matched data. This is why experiments are so much more persuasive: because they also ensure balance for unmeasured confounders. The authors of the study acknowledge as much, writing:

A possible explanation is that RHC is actually beneficial and that we missed this relationship because we did not adequately adjust for some confounding variable that increased both the likelihood of RHC and the likelihood of death. As we found in this study, RHC is more likely to be used in sicker patients who are also more likely to die.

Another possible explanation is that we simply haven't been able to match treatment cases with control cases very effectively. The right half of Table [9.6](#) shows that covariate balance for the matched data is noticeably better than for the unmatched data, but it's not perfect. We still see some small differences in complication rates and APACHE scores between the treatment and control group. There are two main reasons for this.

- (1) First, and most importantly, although finding a match on one or two variables is relatively easy, finding a match on several

variables is pretty hard. Think of this in terms of your own life experience—for example, in seeking a spouse or partner. It probably isn't too hard to find someone who's a good match for you in terms of your interests and your sense of humor. But if you require that this person *also* match you in terms of age, career, education, home town, height, weight, looks, and favorite sport, then you're a lot less likely to find a match. *Picky people are less likely to find a satisfying match in life.* For this same reason, it's unlikely that we'll be able to find an exact match for each treatment case in a matching problem, especially with lots of possible confounders.

- (2) Second, finding matches for cases with rare confounders is especially hard—by definition, since the confounder is rare!

These two points underline a basic difficulty with matching: perfect matches usually don't exist, and we have no choice but to accept approximate matches. In practice, therefore, we give up on the requirement that every single pair of matched observations is similar in terms of all possible confounders, and settle for having matched groups that are similar in their confounders, *on average*. That's why it's so important to check the covariate balance after finding matched pairs, to make sure that there's nothing radically different between the two groups.

## Model-based statistical adjustment

A fourth identification strategy for estimating a causal effect is to build a regression model. If some important (and quite strong) assumptions are met, then such a model is capable of isolating a causal relationship between predictor and response, by adjusting for the effects of confounders *statistically*, rather than experimentally. You may have heard this process described as “statistical control” or “statistical correction,” both in the popular media and in scientific publications:

- “Schatz’s numbers are unique in that they evaluate each play against the league average for plays of its type, adjust for the strength of the opponents’ defense, and even try to divide credit for a given play among teammates.”<sup>15</sup>
- “The committee concluded that a statistical adjustment of the 1990 census leads to an improvement of the counts.”<sup>16</sup>

<sup>15</sup> “Pigskin Pythagoras: A guy from Framingham tries to remake the muddy field of football statistics.” *Boston Globe*, February 1, 2004

<sup>16</sup> “Judge must decide on census adjustment.” *Chicago Tribune*, 6/8/1992

- “Further adjustment for weight change and leukocyte count attenuated these risks substantially.”<sup>17</sup>

Estimating a causal effect using a regression model is, in principle, no different than estimating a partial relationship, which we’ve already learned how to do:

- (1) Build a multiple regression model for the outcome ( $y$ ) versus the predictor of interest ( $x$ ) and other possible confounders;
- (2) Interpret the coefficient on the  $x$  variable of interest as the partial linear relationship between  $y$  and  $x$ , holding confounders constant.

The key question is: under what circumstances can we interpret the partial relationship in a multiple regression model as the *causal* effect of  $x$  on  $y$ ? By *causal effect*, you should think in terms of the counterfactuals we entertained at the beginning of the chapter: *if* I were to intervene and change  $x$  by one unit, holding all other variables constant, *then* how much would  $y$  change on average?

There are three important assumptions that must be met in order to give a causal interpretation to a regression coefficient. First, your model must include all confounding variables (that is, variables that have a causal effect on both the treatment assignment and the outcome). Second, the model must be correct. In this context, “correct” means that you have included the right interactions among confounding variables, and that you have specified the right functional form of the model (linear, polynomial, power law, etc.). Finally, you must *not* include any post-treatment effects as covariates in the model. A post-treatment effect is something causally “downstream” from the treatment variable, and that becomes known only as a result of receiving or not receiving the treatment. This is a subtle point, and we won’t discuss it in detail. But the important thing is: include those confounders, and *only* those confounders, that affect the allocation of cases to the treatment and control groups.

If, and only if, these three assumptions about your model are true, then the regression coefficient of  $y$  on  $x$  has a causal interpretation. If, on the other hand, there are any unmeasured confounders affecting your  $x$  and  $y$  variable, then the coefficient of  $y$  on  $x$  measures association, not causation. This is called *omitted-variable bias*.<sup>18</sup>

Another way of saying this is that *if* the possible confounders are all observed, then accurately estimating the causal effect of

<sup>17</sup> “Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus: A Cohort Study.” *Annals of Internal Medicine*, January 4, 2010

<sup>18</sup> Or *lurking-variable bias*.

$x$  on  $y$  really just boils down to modeling the data well, and not using that model to extrapolate beyond the range of available data. However, the assumption that we've observed all relevant confounders, and can therefore adjust for them appropriately, is very strong. It's also unverifiable using the data; as with matching, you have to believe this assumption, and convince people of it, on extrinsic grounds.

Using regression analysis to estimate causal effects is a big, serious topic. Here are two full books about it:

- *Causality*, by Judea Pearl
- *Observational Studies*, by Paul Rosenbaum

For some additional, more easily digestible advice on choosing which covariates to include in a causal model, see [Chapter 17](#) of Daniel Kaplan's book on statistical modeling.<sup>19</sup>

#### *Matching versus regression, or matching and regression?*

We've seen that it's easiest to infer causality if the cases in the treatment group are comparable to those in the control group. One way to do this is via matching: explicitly constructing a balanced data set from an unbalanced one. Another way to do this is via regression: adjust for confounders using a statistical model, so that we can evaluate the partial relationship between treatment and response, holding confounders constant.

This makes it sound as though regression and matching are competing identification strategies for causal inference. Sociologically speaking, there is certainly some truth to this, in that some people tend to use matching more often, and others tend to use regression more often. So which one should *you* use?

In the real world, if you're going to use only one strategy or the other, my advice is to use matching, mainly for three reasons:

- (1) Matching is a lot easier for non-experts to understand, since you can point to the matched treatment and control groups and show that they are visibly balanced with respect to observed confounders. In other words, the nature of the "balanced comparison" being made via matching is much more transparent than the idea of a partial slope in a regression model. This will make it easier for you to convince others of your conclusions.

<sup>19</sup> Kaplan also has a good explanation for why it's not a good idea to include post-treatment effects (i.e. variables causally downstream of the treatment) as covariates in a regression model.

- (2) Matching is a bit more robust than regression, at least in their “off the shelf” versions. The regression-based approach to causal inference relies on a whole bunch of hard-to-verify assumptions: linearity, all necessary interactions included, and so forth. By comparison, it’s a lot easier to verify covariate balance using before-and-after tables of summary statistics. (Of course, neither method is robust to unmeasured confounders—only an experiment can fix that problem.)
- (3) Unwarranted extrapolations are more apparent when matching than when using regression. Suppose that the treatment and control groups have highly nonoverlapping distributions of confounders—for example, that most the men are in the treatment group and most of the women in the control group. In such cases, the data are inherently limited in what they can tell us about the treatment-response relationship in this region of nonoverlap (i.e. how the treatment will work for women). This lack of overlap will be obvious if you use matching, because you’ll still have drastic post-match covariate imbalances that will stick out like a sore thumb. But the lack of overlap will be less obvious if you throw all the confounders into a multiple regression model without plotting your data.

In summary, it’s easier to convince others with matching, and easier to fool yourself with regression. These aren’t intrinsic *statistical* advantages to matching; they are merely *practical* advantages worth keeping in mind.

It turns out, however, that there’s no need to choose between matching and regression. Better still is to use both matching *and* regression, to get better estimates of causal effects than either technique is capable of getting on its own. In other words: first run matching to get an approximately balanced data set. Then run a regression model for the response versus the treatment variable and the confounders, to correct for minor imbalances in the matched data set. Under this approach, the primary role of matching is to correct for major covariate imbalances between the groups, while the primary role of regression is to model the treatment-response relationship in a way that adjusts for any minor confounding that remains in the matched data set.

There’s one other major advantage of using matching and regression together. By fitting a regression model to a matched data set, you are able to search for interactions *between* the treatment

variable and possible confounders. For example, what if the treatment effect is different for men than for women? You can discover this kind of modulating effect much more easily using a regression model than you can with matching alone.

In summary, matching and regression make for an excellent pair. There's rarely a good reason to use just one or other!