2
Exploring Multivariate Data

Key terms and concepts: contingency table, mosaic plot, between-group variability versus within-group variability, boxplot, dot plot, group means and grand mean, regression equation, linear predictor, parameter, model/fitted value, residual, statistical adjustment, brushing, influence plot

Supply and demand, chocolate and peanut butter, education and income . . . some things just go hand in hand. Empirical relationships like these are the backbone of causal reasoning, regardless of whether the theory—the why, the how—comes first or second.

Supply and demand: Economic theory and common sense predict that people will demand less of a thing as its price rises. How are we to explain, then, why demand for certain goods—a Rolls Royce, a Gucci handbag, a bottle of Dom Perignon—seems to increase with price, in apparent defiance of gravity? The empirical price–demand relationship for these *Veblen goods* requires a causal story, the canonical one being that people buy luxury items as status symbols.

Veblen goods are named after economist Thorstein Veblein. These are not to be confused with *Giffen goods*, where the usual price–demand relationship is reversed because of something called the income effect.

Chocolate and peanut butter: Sometimes empirical relationships can suggest new theories. For example, one notable culinary fact is that people, regardless of culture, tend to like sweet and salty things together: melon and ham; coconut and curry; and yes, Reese's pieces. How does this fact influence the research agenda of scientists who study the physiology of human taste buds? Might sodium play a role in the perception of every basic taste, and not just the taste of salty things?

In the "you'll just have to trust us" department: if you like coffee but don't like the bitter taste, try adding a pinch of coarse salt for every 2 tablespoons of coffee grounds, all mixed together when dry, next time you brew a pot. . . .

Education and income: "Human capital" sounds like a phrase of recent vintage. But it dates at least to Adam Smith:

> The acquisition of such talents, by the maintenance of the acquirer during his education, study, or apprenticeship, always costs a real expense, which is a capital fixed and realized, as

> it were, in his person. Those talents, as they make a part of his fortune, so do they likewise that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labor, and which, though it costs a certain expense, repays that expense with a profit.[1]

The natural empirical question is: if education is a capital investment, how good is the return?

[1] *An Inquiry into the Nature And Causes of the Wealth of Nations*, Book 2.

In each case, a particular causal story—a particular idea about how things work—turns upon the interpretation of an observed relationship between two variables. To do this correctly requires both care and judgment. But it also requires a richer mathematical language for summarizing joint variation. The goal of this chapter is to provide you with just such a language.

Variation across categories

MUCH OF the data you'll meet will involve unordered categories: man or woman; chocolate or vanilla; butcher or baker or candlestick maker. A simple, effective way to summarize these *categorical variables* 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:

The R script `titanic.R` has all the R code used to generate the tables and plots for the Titanic data set on this page. Useful commands here include `xtabs`, `table`, `factor`, and `mosaicplot` (in the `graphics` package).

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

Table 2.1: 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*.

But as this three-way table reveals, richer passengers, of either sex, fared better than others:

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

Table 2.2: 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.

The categories go along the rows and columns of the table; the cell counts show how many cases fall into each class; this is often

called *cross-tabulation*. Given the constraints of a two-dimensional page, multiway tables are usually displayed as a series of two-way tables, just as you see here.

Some categories have a natural ordering, like measures of severity for a hurricane, or responses to a survey about consumer satisfaction. These are called *ordinal variables*. Likewise, sometimes a categorical variable has only two options, one of which is considered a success—heads or tails, survived or died. Sometimes we code these outcomes as 1's and 0's, respectively, in which case they are called indicator or dummy variables.

Tables are almost always the best way to display small data sets with few classifying variables—they convey much information in little space. A few guidelines will help keep your tables clean and easily readable: (1) Do not use vertical lines. (2) Put units in column labels. (3) Align columns of numbers at the decimal point. Thus:

Good

Item	Price ($)
Laptop	1149.00
Mouse	39.00
Ink pen	2.50
Paper clip	0.02

Bad

Item	Price
Laptop	$1149.00
Mouse	$39.00
Ink pen	$2.50
Paper clip	$0.02

Categorical data can also be displayed in a *mosaic plot*. The area of each box tells you what fraction of cases fall into the corresponding cell of the contingency table. At a glance, this gives you an idea of how category membership predicts some of the variation in the outcome variable. Compare Figure 2.1 with the previous table of Titanic survivors stratified by sex.

Mosaic plots are at their best when used for data sets with more than two cross-classifying variables. In this context they are more useful for quick visual comparisons than are tables. For example, Figure 2.2 depicts joint variation in hair color, eye color, and sex for a sample of 592 students at the University of Delaware. Try answering the following questions using the plot.

(1) Among students with blond hair and blue eyes, who is overrepresented: men or women?

(2) For which hair color would green eyes be the most surprising?

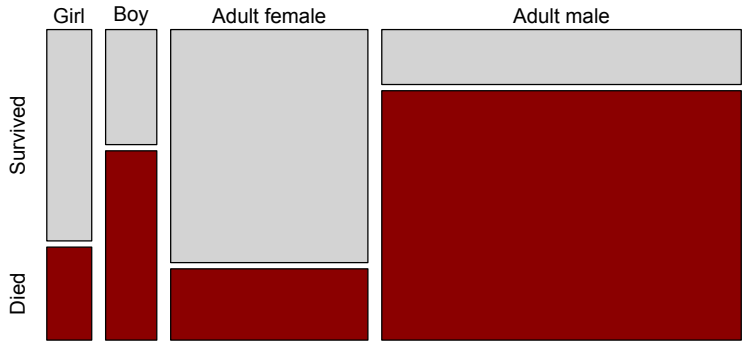


Figure 2.1: Survival of passengers on the Titanic, stratified by the categorical predictor "girl/boy/man/woman."

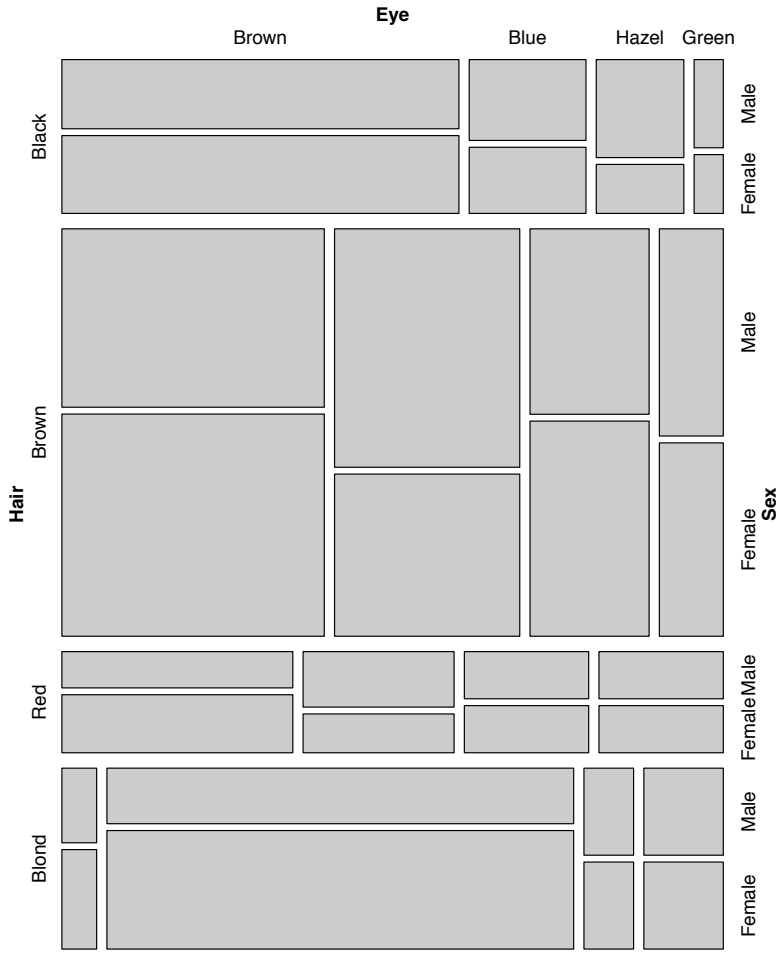


Figure 2.2: Hair color, eye color, and sex for 592 students at the University of Delaware. Data from Snee (1974), "Graphical display of two-way contingency tables." *The American Statistician*, 28, 9–12. It appears as though, among people with blond hair and blue eyes, women are over-represented. (Compare the sizes of the male and female boxes for the blond-haired, blue-eyed crowd in the lower left.) It also appears as though green eyes are rarest among people with black hair, in a relative sense. (Notice how the "Green eyes" boxes are narrowest left to right in the row for black hair, compared to the row for any other hair color.)

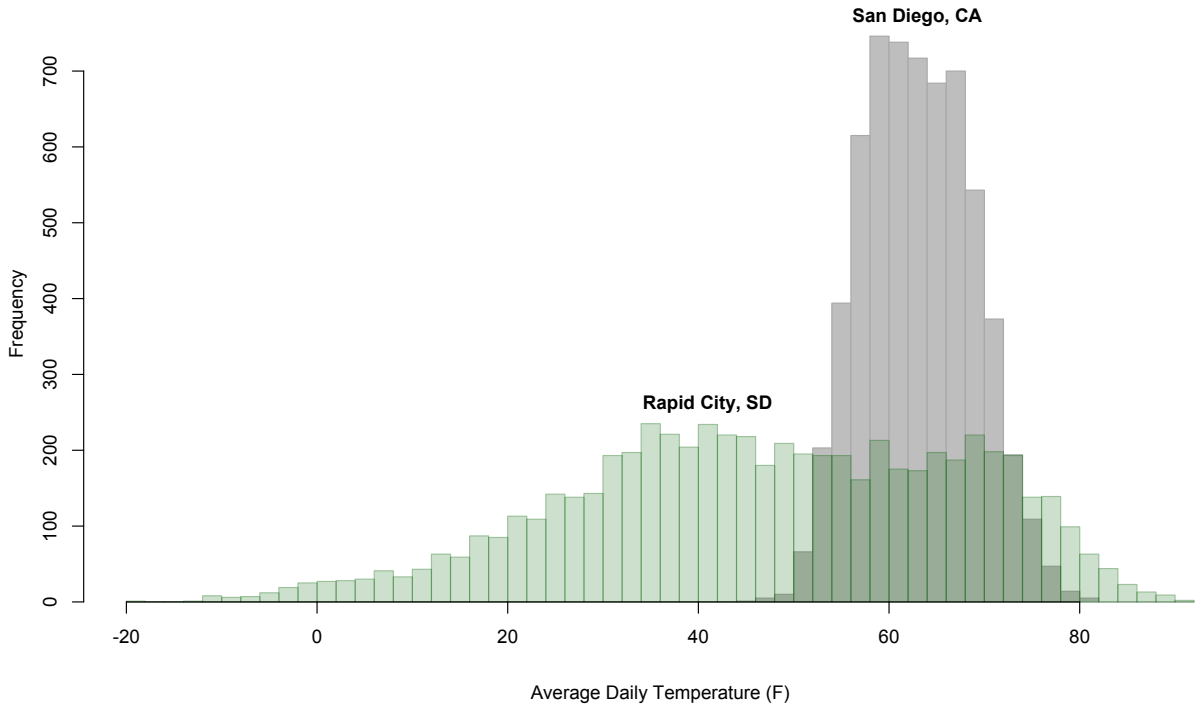


Figure 2.3: Daily average temperatures for San Diego and Rapid City, 1995–2011. Temperature is an example of a *quantitative variable*, or something for which numerical comparisons are meaningful (twice as far, six times as fast, $17 cheaper, and so forth). Quantitative variables can be discrete or continuous. Marbles are discrete; we count them on our fingers and toes. Speed and distance are continuous; we measure them in arbitrarily small increments.

Variation of a typical case

FIGURE 2.3 depicts the average daily temperatures in two American cities—San Diego, CA, and Rapid City, SD—for every day from January 1995 to November 2011. These pictures suggest two obvious, meaningful questions we can ask about a *quantitative variable* like temperature: where is the middle of the sample, and how much does a typical case vary from the middle?

You're already aware of more than more way to answer the question, "Where is the middle?"

- There's the sample mean, written as $\bar{y}$.
- There's the median, or the halfway point in a sample.
- There's also the mode, or the most common value.

If we have n samples $\{y_1, \ldots, y_n\}$, then

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

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.

For example, consider the data set $\{1, 2, 3, 3, 4, 4, 5\}$.

It's much the same with the question, "How much does a typical case vary from the middle?" There is more than one way of answering it, 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. Suppose we choose to measure the middle of a sample $y_1, \ldots, y_n$ 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

The subscript i's run from case 1 to case n, where n is the number of data points in the sample. In most data sets the actual ordering of cases won't matter, and will just reflect the arbitrary ordering of the rows in your data frame. A notable exception is in the analysis of time-series data, where the ordering of observations in time may be highly meaningful.

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

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.

I said this quantity "makes sense," and at this point you may be objecting to that claim! Two natural questions are:

(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 highly overrated.) Mainly we use $n-1$ to follow convention.

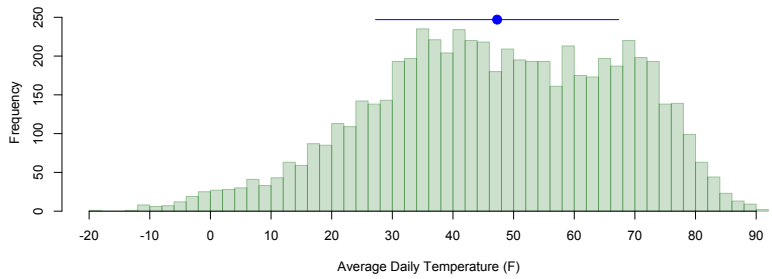
As for the second question: because sums of squares are special! In all seriousness, there are deep mathematical reasons why we choose to measure dispersion using sums of squared deviations, rather than the seemingly more natural sums of absolute deviations. You'll learn why in the next chapter—but it 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 Y is measured in. This is not very intuitive; imagining telling someone that the mean daily 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 by non-statisticians, and thus cold comfort.

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

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

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.



Two other 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^{th} and 25^{th} percentiles. This is fairly robust to extreme values, since it involves only the middle 50% of the sample.

Variation between, and within, groups

MAGAZINES, blogs, and college brochures (among others) love to publish tables that stratify some numerical variable by a categorical predictor. For example, Table 2.3 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.

	Average SAT	
College	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 2.3: 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.

This is superficially similar to the contingency tables we just saw. After all, some of the variability in student SAT scores can be predicted by a student's college, and the table gives you some idea of this between-group variability:

$$\text{SAT score} \sim \text{College}\,.$$

Math skills, for example, are probably more important for engineering majors than English majors.

But Table 2.3 is different from a contingency table in one crucial respect: it throws away information. 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.

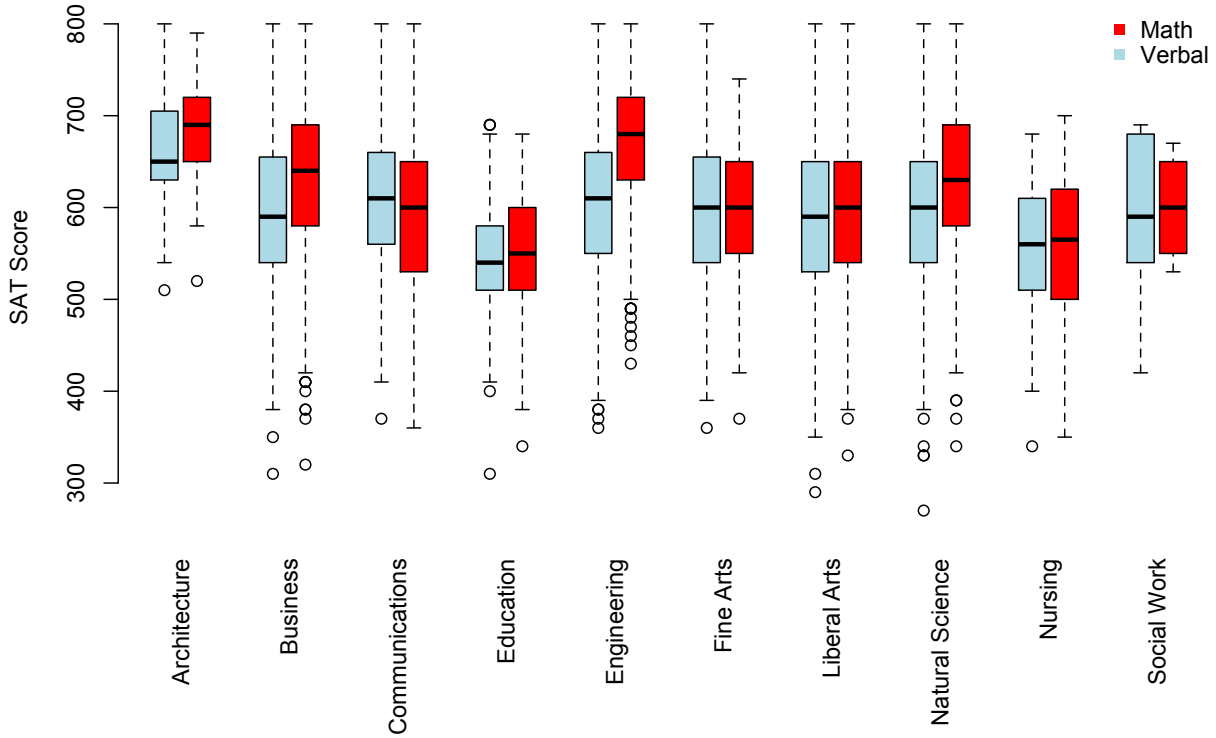


Figure 2.4: Boxplots of the full data set used to form the means in Table 2.3.

Boxplots

This is where boxplots are useful: they allow you to assess variability both between and within the groups. Many times it is the within-group variability that matters most. For example, as Figure 2.4 shows, SAT scores vary at least as much within a college as they do between colleges. For example, the 52-point difference in average SAT math scores between Architecture students and Natural Science students looks a little smaller compared to the roughly 100-point interquartile range of math scores within Natural Science! Although between-group variability cannot be ignored, it is nonetheless dwarfed by within-group variability.

The situation is quite different Figure 2.5. 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 survive, yet cannot synthesize their own. The amount they are given predicts much about their health, measured in this case by the length of their teeth. We therefore see compara-

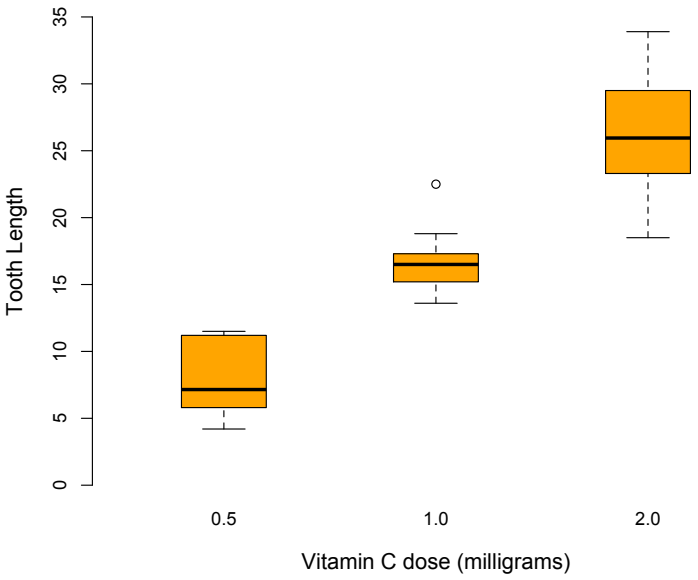


Figure 2.5: For comparison, the table of within-group means is below. Notice how the within-group variability evident in the boxplots at left simply disappears when presented in table form.

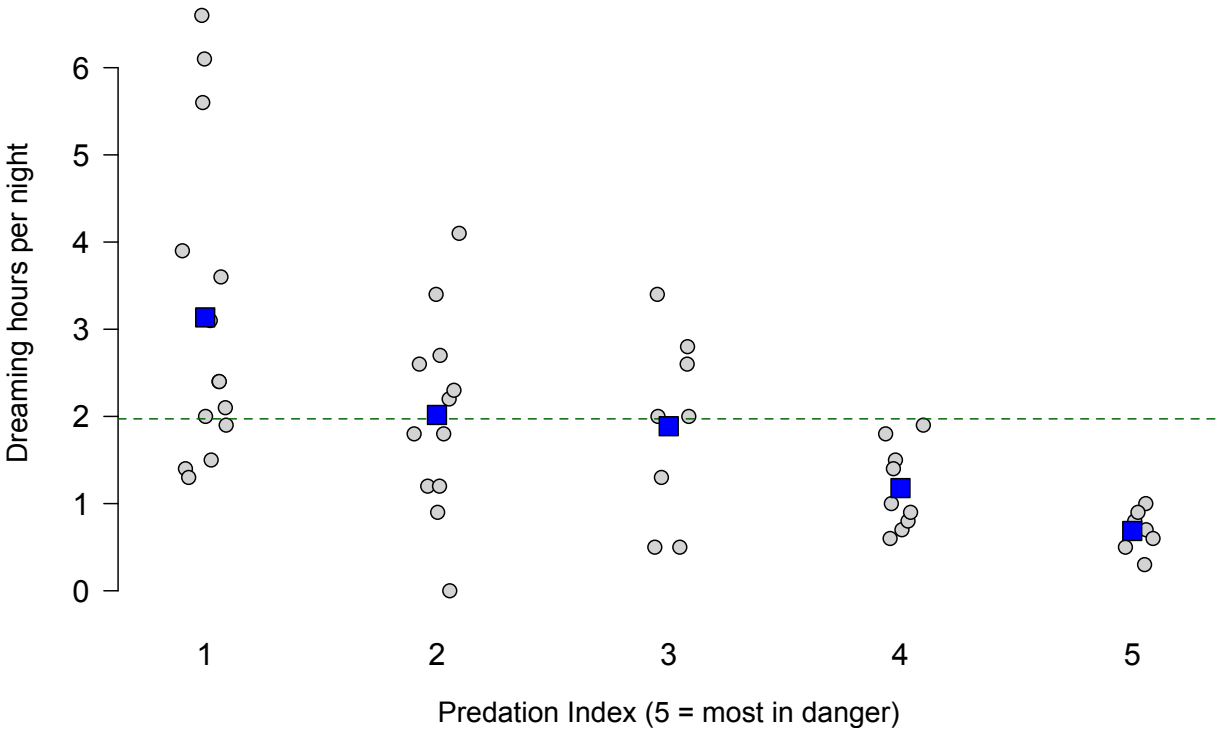
Dose (mg)	Tooth len.
0.5	7.98
1.0	16.77
2.0	26.14

tively more variability between the groups, whose boxplots almost don't overlap.

This idea of within-group variability versus between-group variability will come up again and again. Remember, statistical models partition variability into predictable and unpredictable components. We'll soon make this mathematically rigorous, but these examples convey the essence of the idea:

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

A table of group-wise means does not depict "data" as such, but an abstraction of some typical group member masquerading as data. This abstraction may be useful for some purposes. But sometimes within-group variability is also important, or even 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.

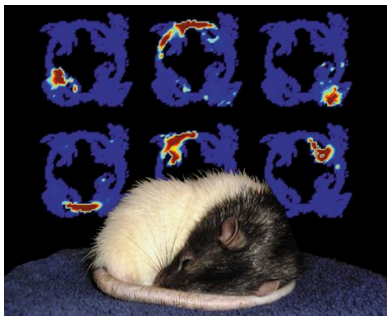


Dot plots

The *dot plot*, or strip chart, is a close cousin of the boxplot. For example, the plot above depicts a relationship between the length of a mammal's dreams and the severity of the danger it faces from predators. Each dot is a single species of mammal—like, for example, the critter at right. The predation index (X) 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 bit of meaningless horizontal jitter has been added to the dots, which allows them to be distinguished from one another more easily.

Figure 2.6: 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).



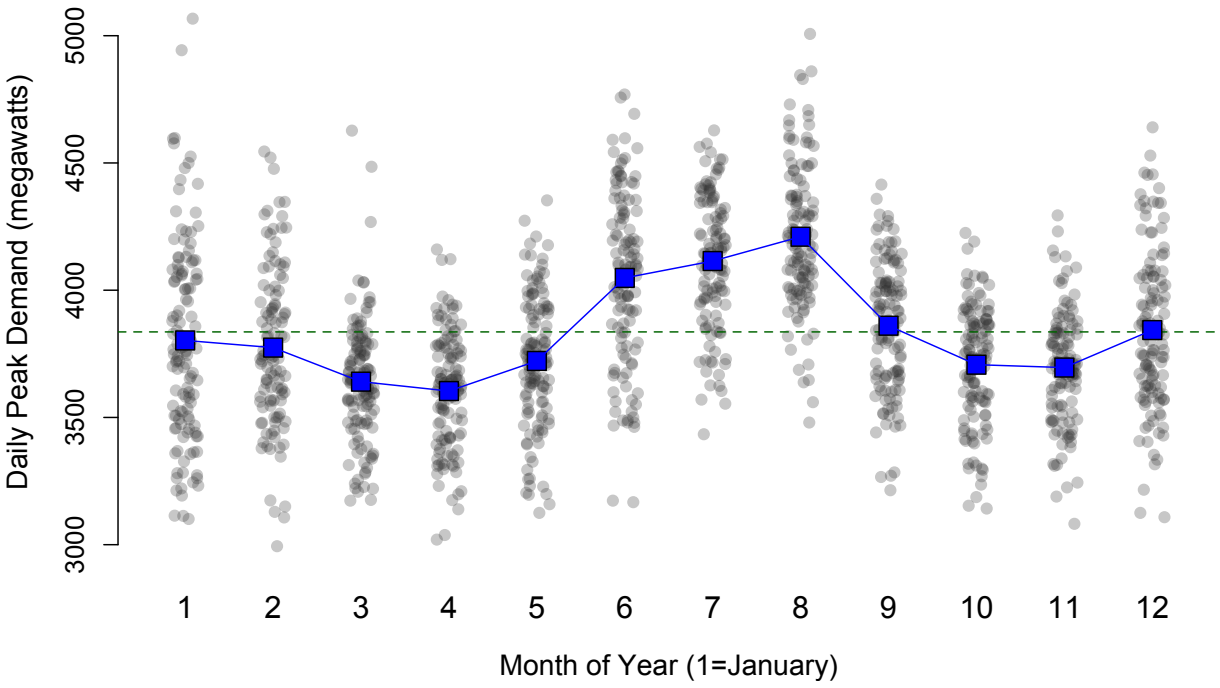


Figure 2.7: Daily peak electricity demand (stratified by month) in Raleigh, NC from 2006–09. The dashed line is the average peak demand for the whole data set, and the blue dots are the month-by-month means.

Dot plots can also be effective for larger data sets. Here 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 beautifully.

Group means and grand means

If you looked carefully, you may have noticed two extra features of the dot plots in Figures 2.6 and 2.7. 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. Notice that, in plotting these means along with the data, we have implicitly partitioned the variability:

$$\begin{aligned} \text{Individual case} &= \text{Group mean} + \text{Deviation of that case} \\ &= \text{Grand mean} + \text{Deviation of group} + \text{Deviation of case}\,. \end{aligned} \tag{2.1}$$

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

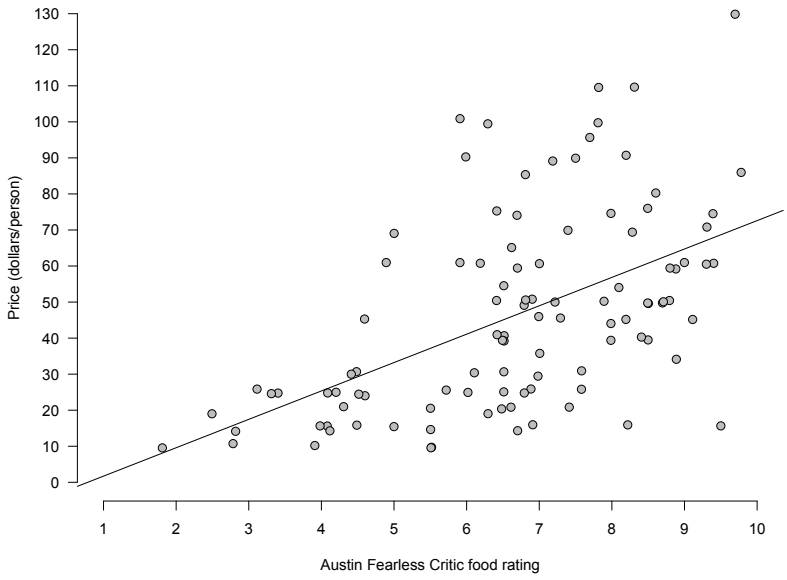


Figure 2.8: Price versus reviewer food rating for a sample of 104 restaurants near downtown Austin, Texas. The data are from a larger sample of 317 restaurants from across greater Austin, but downtown-area restaurants were chosen to hold location relatively constant. Data from Austin Fearless Critic, www.fearlesscritic.com/austin. Because of ties in the data, a small vertical jitter was added for plotting purposes only. The equation of the line drawn here is $y = -6.2 + 7.9x$.

Two quantitities varying together

Fitting straight lines

Group by group isn't the only way that numerical quantities can vary predictably. The scatter plot above 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. On average, it appears that people pay more for tastier food!

This is our first of many data sets where the predictor (price, Y) and response (food score, X) can be described by a linear model. We write the model in two parts as "$Y = \beta_0 + \beta_1 X +$ 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 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:

$$y_i = \beta_0 + \beta_1 x_i + e_i \,. \tag{2.2}$$

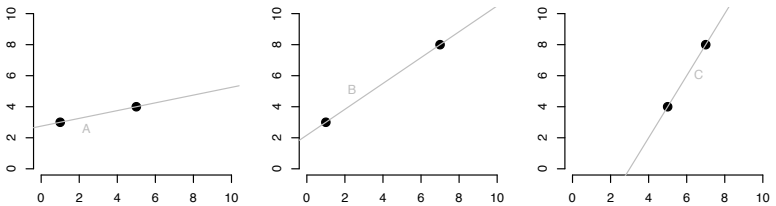
An equation like (2.2) is often called a *regression equation*. Together the *intercept* β_0 and the *slope* β_1 are called the *parameters* of the regression equation. Meanwhile, e_i is called the *residual* for the ith case—residual, because it is what's left over in the Y variable after accounting for the contribution of the X variable.

For every two points. . . .

An obvious question is: how do we fit the parameters β_0 and β_1 to the observed data? (This process is called *linear regression*). Historically, the standard approach, still in widespread use today, is by the method of least squares. We do this by choosing β_0 and β_1 so that the sum of squared residuals (the e_i's) will be as small as mathematically possible.

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.

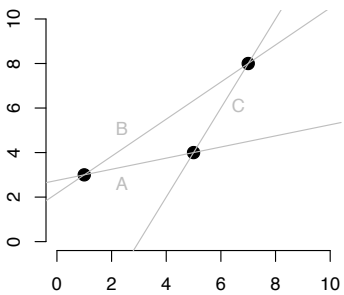


For every two points, a line. If life were always this simple, there would be no need for statistics!

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. No solution—and therefore no perfectly fitting linear trend—exists. Seen graphically, at right, it is clear that no line can pass through all three points.



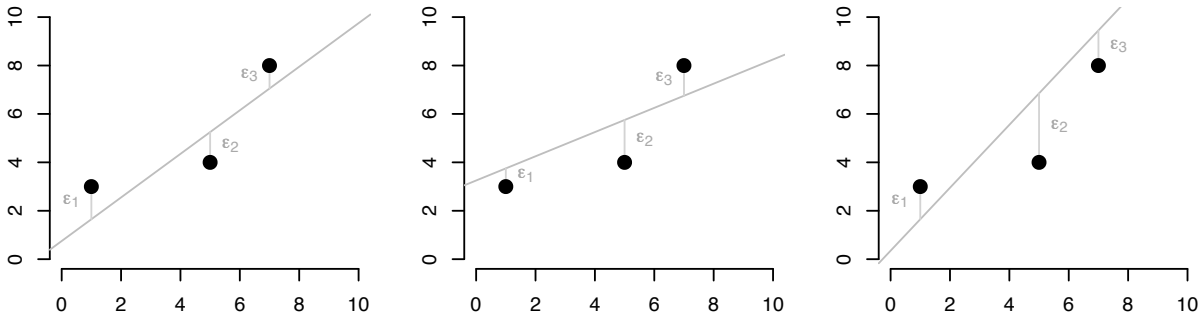


Figure 2.9: Three possible straight-line fits, each involving an attempt to distribute the "errors" among the observations.

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.

It is clear that, 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. Let's express these small misses mathematically:

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

The three little e's are, of course, the residuals.

But now we've created a different kind of predicament. Before we added the e_i's to give us some wiggle room, there was no solution to our system of linear equations. Now we have three equations and five unknowns: an intercept, a slope, and three residuals. Out of the frying pan and into the fire—this system has infinitely many solutions! How are we to choose, for example, among the three lines in Figure 2.9? 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 very 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 usings 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, essentially, fitting a straight line to observations of the moon's orbit. The same problem of fitting astronomical orbits arose in a wide variety of situations. Many proposals for doing so 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.[2]

[2] *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, more than half a century after scientists first entertained the idea of combining observations. What changed things was the *method of least squares*, independently invented by at least 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, in fact, a direct translation of Legendre's original 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. Legendre put it like this:

> 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.[3]

[3] Adrien-Marie Legendre (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*. Translation p. 13, Stigler's *A History of Statistics*.

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 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 $(\widehat{\beta}_0, \widehat{\beta}_1)$:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.3}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}\,, \tag{2.4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the X and Y variables, respectively. The line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ 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 nothing more difficult than taking a few derivatives and solving a simple matrix-algebra problem—something that scientists of the nineteenth century could do easily.

With modern computers, the computation of least-squares estimates is now entirely automatic. Their use and interpretation, however, isn't automatic at all. Here are four kinds of stories one can tell with a regression equation. Each is useful for a different purpose.

Least squares then and now: an 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 in 3 weeks—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 it, moreover, by hand! This is a task that would 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 in 2009 takes less than 5 seconds to solve the same problem.

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

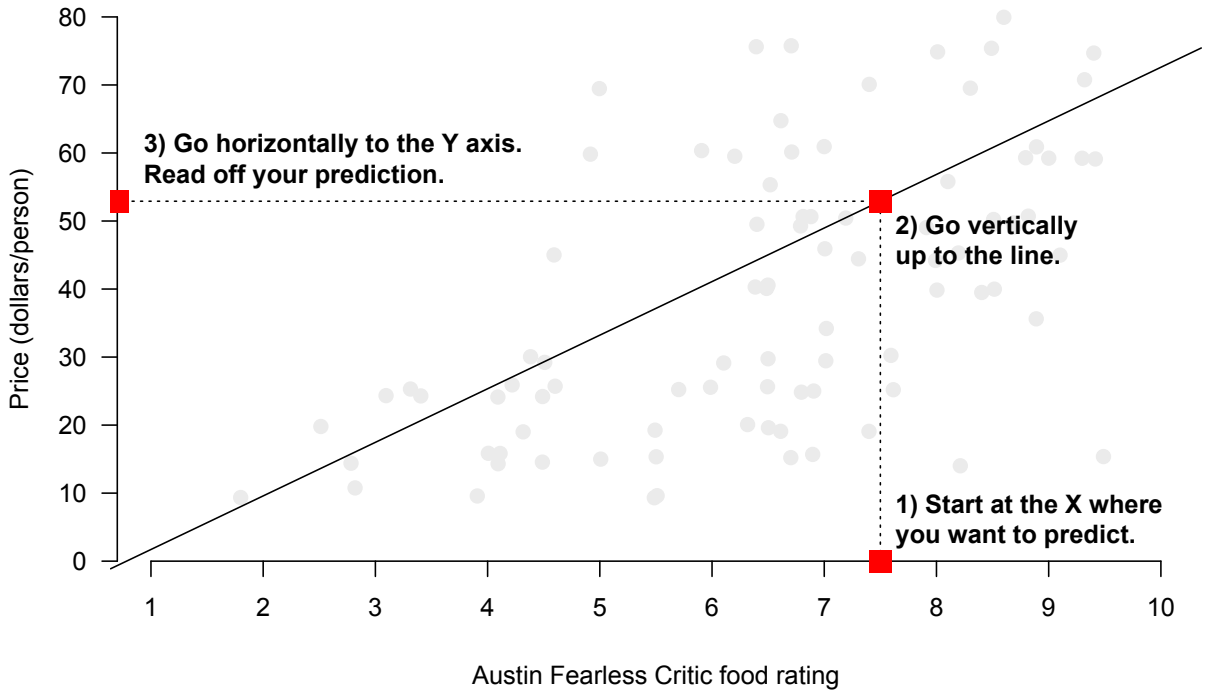
In other words, it is a function $\hat{y} = f(x)$ that maps inputs to outputs. When we plug in the original X values in to the least-squares linear predictor, we get back the so-called *model values*, or *fitted values*, denoted $\hat{y}_i$:

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

$$\hat{y}_i = \widehat{\beta}_0 + x_i \widehat{\beta}_1 \,. \tag{2.5}$$

We can also do this for observations not in the original data set. This is useful for forecasting the response for a known value of the predictor. If we see a new observation $x^\star$ and want to predict where the corresponding $y^\star$ will be, we can simply plug in $x^\star$ and read off our guess for $y^\star$ directly from the line.

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: $\hat{y}^\star = -6.2 + 7.9 \cdot 7.5$, or \$53.05 per person. 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.



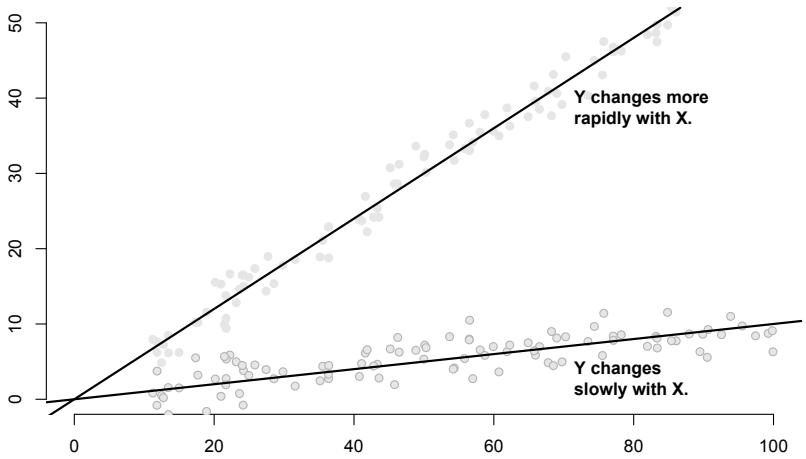
Story 2: A regression equation summarizes the trend in the data.

The linear predictor tells you how Y changes, on average, as a function of X. In particular, the slope β_1 tells you how the response 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 2.8, the slope is $\beta_1 = 7.9$. On average, then, one extra Fearless Critic food rating point (ΔX) is associated with an average increase of \$7.90 (ΔY) 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.

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.



The intercept β_0 is what we'd expect from the response Y if the predictor X took on a value of exactly 0. Try plugging in $x_i = 0$ into the regression equation and notice what you get for the linear predictor: $\beta_0 + \beta_1 \cdot 0 = \beta_0$.

Sometimes the intercept is easily interpretable, and sometimes it isn't. Take the trend line in Figure 2.8, 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 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 our first example of the dangers of extrapolation—that is, forecasting outside the bounds of past experience.

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

Notice that the regression equation splits up every observation in the sample into two pieces, a fitted value ($\beta_0 + \beta_1 x_i$) and a residual (e_i):

$$\text{Observed } y \text{ value} = (\text{Fitted value}) + (\text{Residual}), \tag{2.6}$$

or equivalently,

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

The residuals from a regression equation 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 this interpretation of a residual as an error can often be misleading. The regression equation can still leave a nonzero residual, even if there is no mistake in the measurement of the Y variable. Remember how a statistical model ($Y \sim X$) partitions variation? It's far more illuminating to think of the residual as the part of the Y variable that it is left unpredicted by the X variable.

In Figure 2.8, 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 Lance Armstrong or 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.

Said a different way, the regression equation takes the X-ness out of Y, leaving what remains in 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 2.10. Notice that, in the right panel, there is no evident correlation between food rating and the residuals. The X-ness has been taken out of Y.

Don't look now, but you've just seen your first example of statistical adjustment. Notice the red dot sitting in the lower right

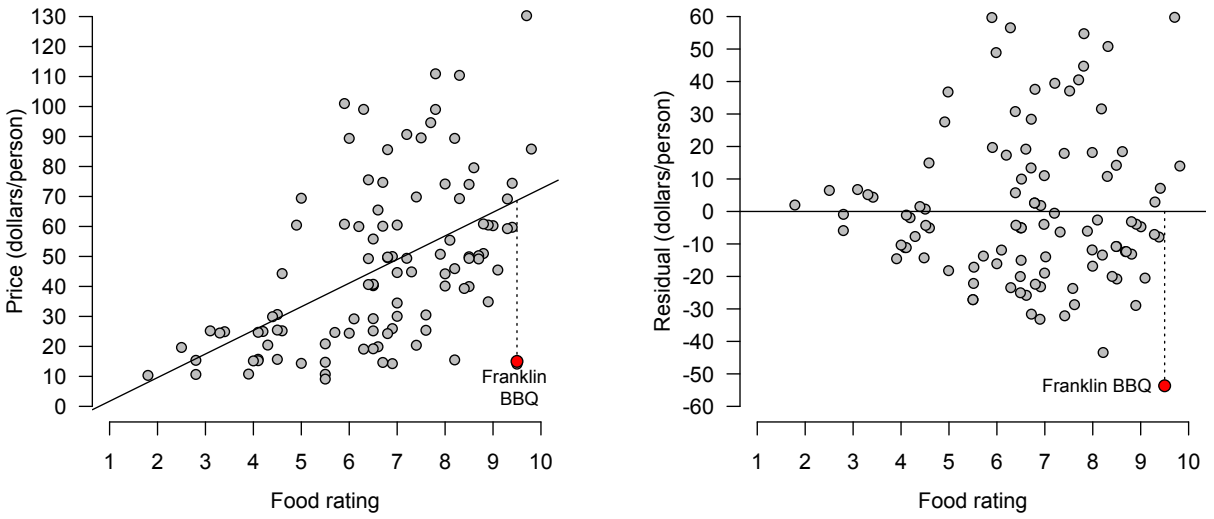


Figure 2.10: Left: the original data on price versus food rating. Right: the residuals from the least squares fit on the left. The residual for Franklin BBQ is the length of the dotted vertical line: $e_i = -\$53.85$.

of Figure 2.10, 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 $\hat{y}^\star = -6.2 + 7.9 \cdot 9.5$, 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.

This restaurant, incidentally, is Franklin Barbecue, declared the "Best Barbecue in America" in 2011 by *Bon Appétit* magazine:

> 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.[4]

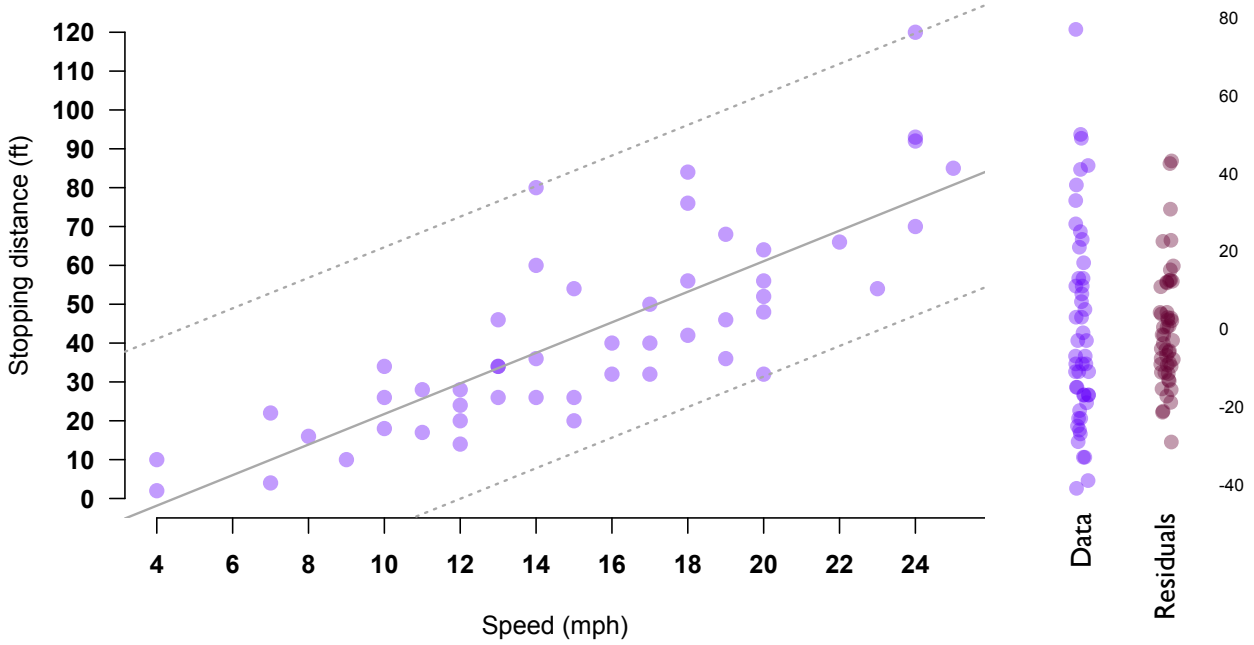
And undoubtedly the most delicious residual in the city.

[4] "A Day in the Life of a BBQ Genius." Andrew Knowlton, *Bon Appétit*, July 2011.

Story 4: A regression equation reduces uncertainty.

How long does it take for your car to stop once you slam on the brakes? Obviously, the answer depends upon many factors: the model of the car, the condition of its brakes and tires, how much weight it's carrying, the slickness of the road, and so forth.

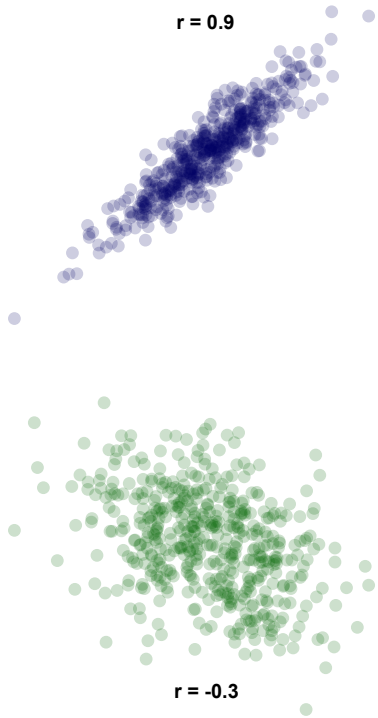
But surely one of the most important factors is the speed you were traveling in the first place. The evidence bears this out:



To the right of the scatter plot, we see two dot plots, both on the same scale: (1) the original deviations $y_i - \bar{y}$, aligned vertically with the actual data points in the scatter plot; and (2) the residuals from the regresson equation. In moving from blue (data) to grey (residuals), some of the variation has clearly been soaked up by the least squares line. This means less uncertainty in forecasting Y, compared to your uncertainty before you knew X.

We say the variation got soaked up—where did it go? The short answer is "into the fitted values." But the more detailed answer to this question turns out to be surprisingly beautiful, and cuts to the heart of an earlier question left unanswered: why measure variation using sums of squares? We'll consider both questions at length in the next chapter.

Sample correlation



The *sample correlation coefficient* is an oft-quoted measure of the strength of linear dependence between two observed quantities:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{2.7}$$

where s_x and s_y are the sample standard deviations of the X and Y variables. At right you see examples of strong positive (top) and weak negative (bottom) correlation. Sample correlation is always between -1 and 1, and is closely related to linear least squares:

(1) A sample correlation of 0 ("uncorrelated") means that the slope of the least-squares line of Y on X is exactly 0.

(2) Sample correlation is between 1 and -1, which are the extremes of perfect positive and perfect negative correlation. The correlation between x_i and the fitted values $\hat{y}_i$ from a least-squares regression of Y on X is either 1 or -1.

The key word here is *linear*. The correlation coefficient is useful to know—but still only a single number, and only able to convey so much information. Therefore, always plot your data, as in Figure 2.11. Four different data sets, four different stories about what's going on. Yet all have the same correlation: $r = 0.816$.

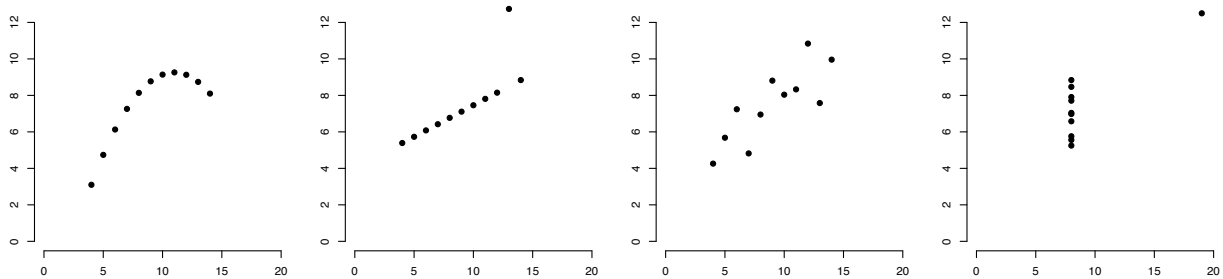


Figure 2.11: Data taken from F.J. Anscombe, "Graphs in Statistical Analysis." *American Statistician*, 27 (1973), pp. 17–21

In fact, the correlation coefficient is so intimately tied up with linear least squares that it breaks down entirely when used to quantify the strength of nonlinear relationships. In each of the six plots in Figure 2.12, for example, there is an obvious nonlinear relationship between X and Y. Yet the sample correlation coefficient for each of them is exactly zero. *Caveat correlator*.

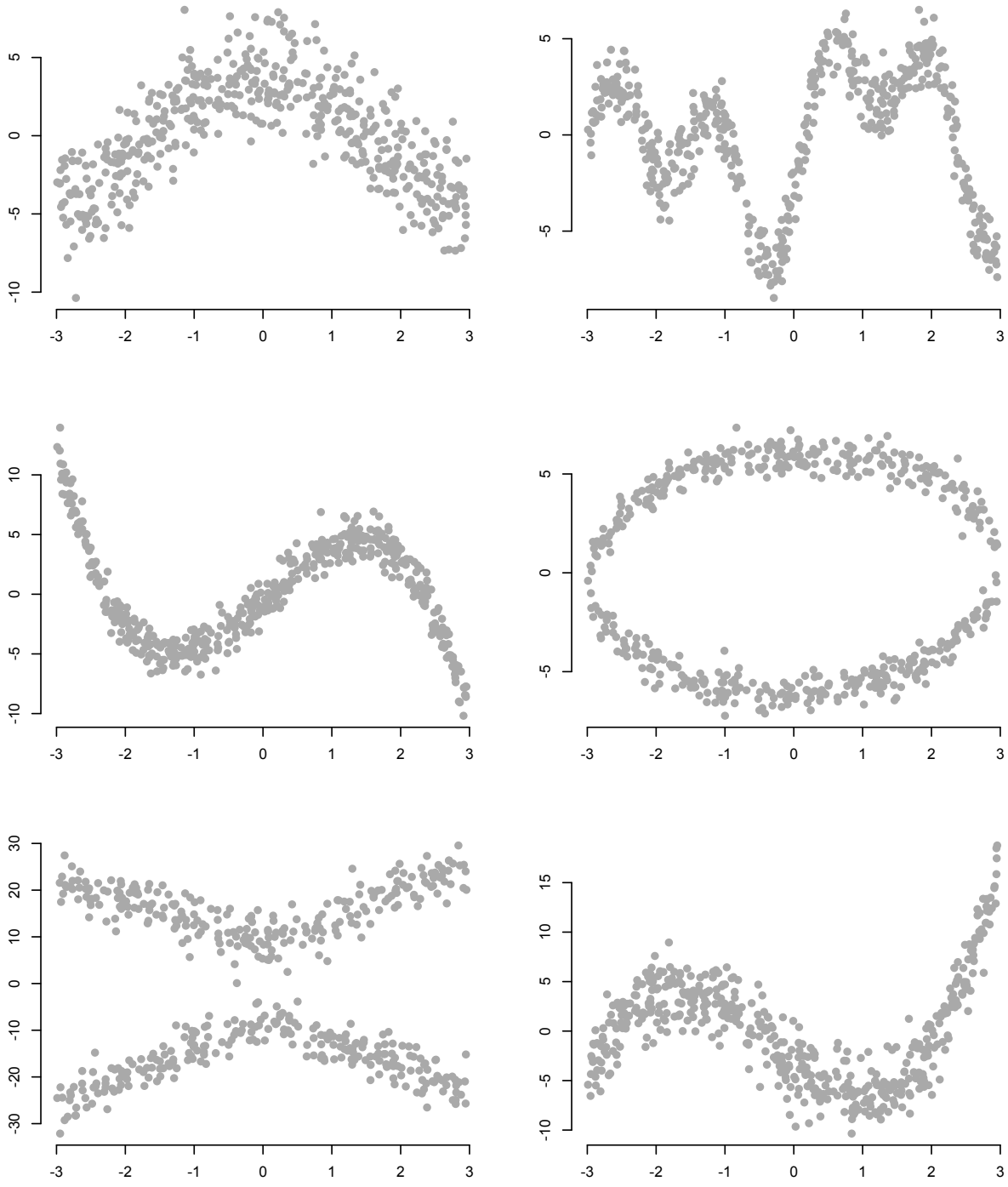


Figure 2.12: Obvious dependence, but zero sample correlation!

Sub-Saharan Africa	North America	South America	Asia–Pacific	Europe	Mid-East & N. Africa
Benin	Canada	Argentina	Australia	Austria	Algeria
Burundi	Dominican Rep.	Bolivia	Hong Kong	Denmark	Egypt
Cameroon	El Salvador	Brazil	India	France	Israel
Cent'l Afr. Rep.	Guatemala	Chile	Indonesia	Germany (West)	Jordan
Congo	Haiti	Colombia	Japan	Greece	Morocco
Ethiopia	Honduras	Ecuador	Korea	Ireland	Syria
Gabon	Jamaica	Paraguay	Malaysia	Italy	Turkey
Gambia	Mexico	Peru	Nepal	Netherlands	
Ghana	Trinidad & Tobago	Uruguay	New Zealand	Norway	
Lesotho	United States	Venezuela	Pakistan	Spain	
Liberia			Papua New Guinea	Sweden	
Madagascar			Philippines	United Kingdom	
Malawi			Singapore		
Mauritania			Taiwan		
Niger			Thailand		
Nigeria					
Rwanda					
Senegal					
South Africa					
Tanzania					
Togo					
Uganda					
Zaire					
Zambia					
Zimbabwe					

Table 2.4: Countries included in the sample for studying the determinants of long-term economic growth.

Case study: schools or guns?

Growth, education, and geography

LET'S NOW work our way through a single example, where we will use some of these basic techniques for assessing relationships among variables. In the process, we'll also see what can go awry in trying to assess causal claims about complex, real-world systems.

Our example comes from the field of political economy: why have some nations become rich while others have remained poor? The long-term fate of nations is a compelling, almost existential topic. Nobody knows for sure why some places enjoy peace and prosperity, while others suffer conflict and poverty. Many candidate explanations are to be found in the works of Smith, Marx, Ricardo, Keynes, Galbraith, Sen, and others still. But let's put aside grand theories for now and look at two specific models:

(1) Economic growth $\sim$ Spending on education; and

(2) Economic growth $\sim$ Spending on military defense.

Good data for studying this question comes a recent academic paper from the *American Economic Review*.[5] For each of the 79 countries listed in Table 2.4, we know its GDP growth rate from 1960–1996, along with a host of potential socio-economic causes of that growth.

[5] Sala-i-Martin, Doppelhofer, and Miller. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." *American Economic Review*, pp. 813–35, Sep. 2004

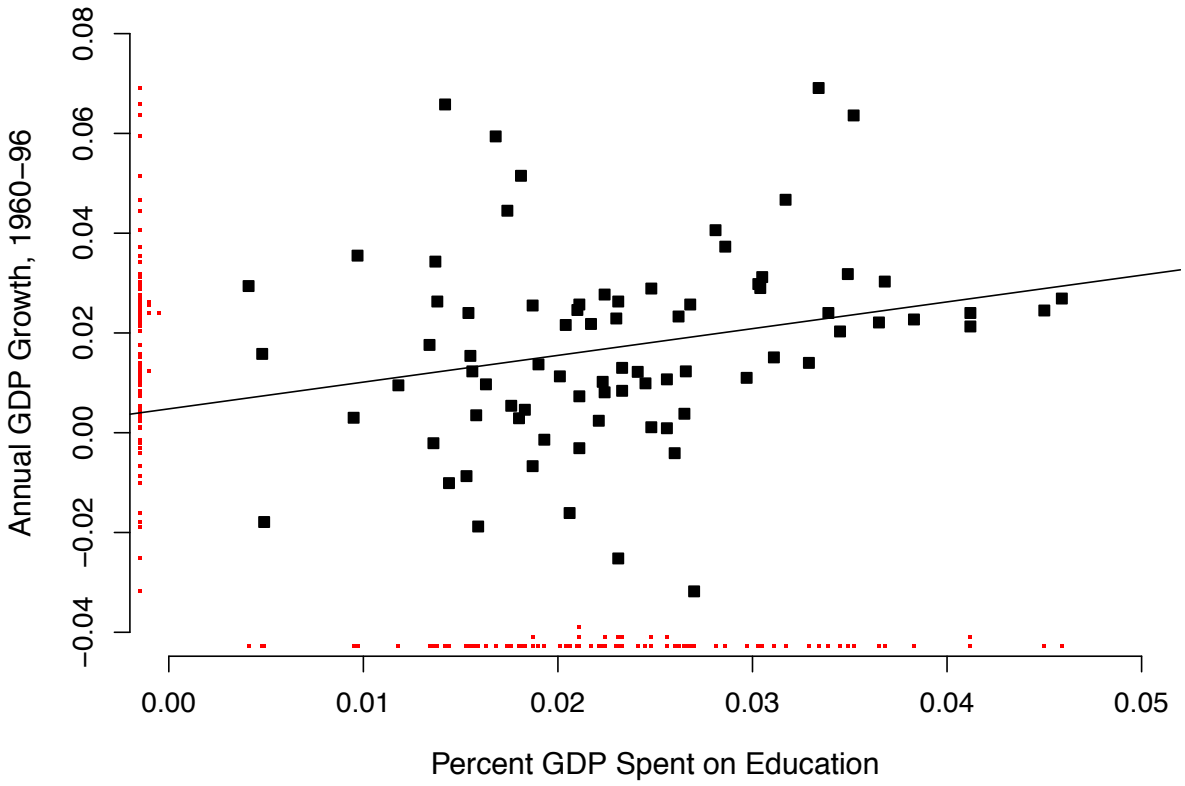


Figure 2.13: GDP growth since 1960 versus percent of GDP spent on education for the 79 countries shown in Table 2.4.

In Figure 2.13, for example, we see a scatter plot of GDP growth versus education spending, together with the upward-sloping least-squares line. On average, then, countries that spend more on education tend to grow at faster rates. But be wary of the causal interpretation, for the potential sources of confounding here are almost endless—for example, geography!

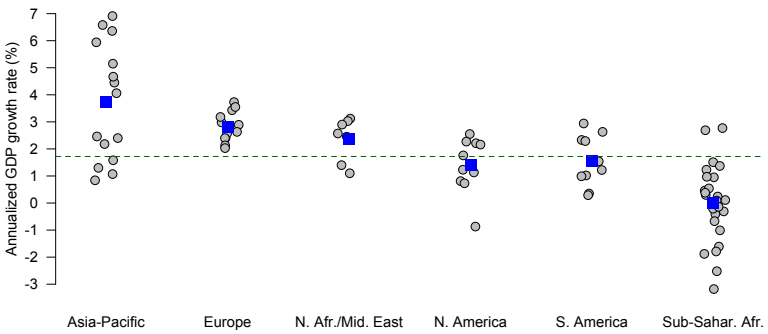


Figure 2.14: GDP growth rates stratified by region.

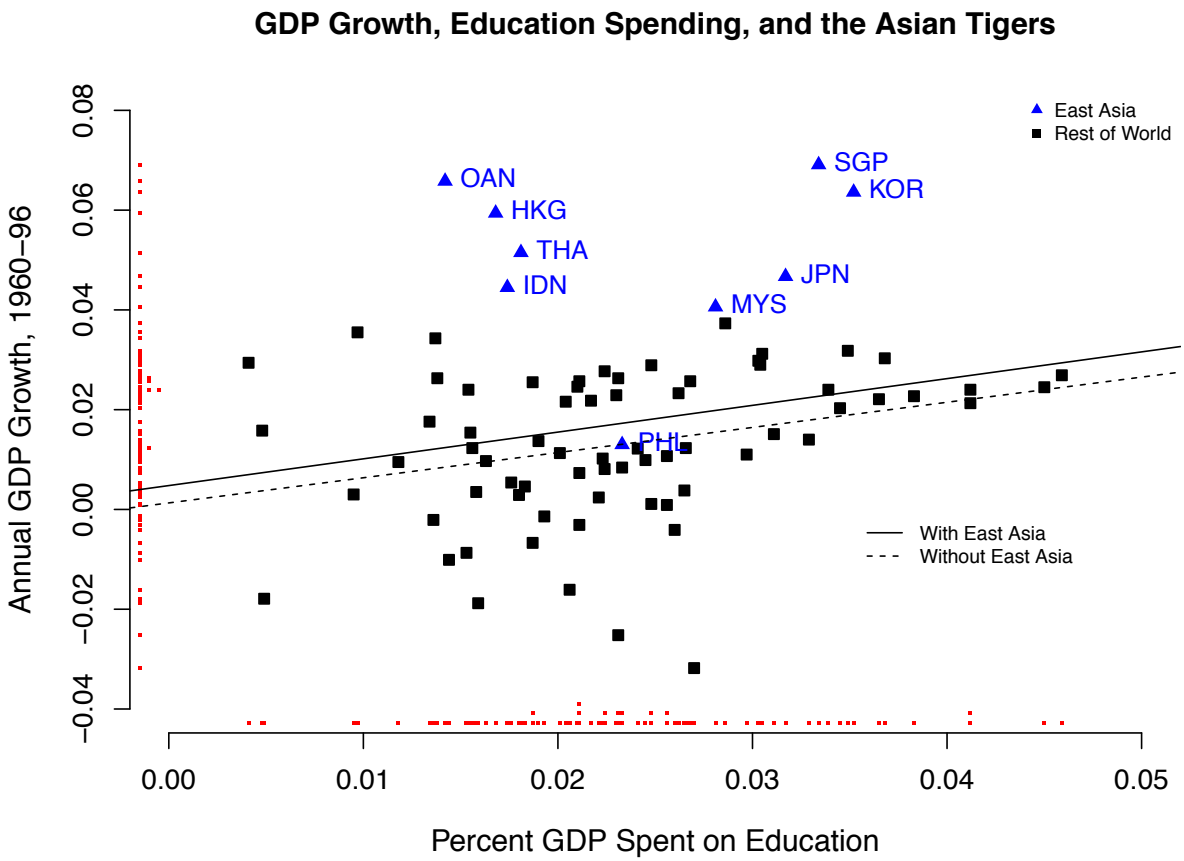


Figure 2.15: The nine east Asian countries in the sample have been brushed, or highlighted.

To see whether geography confounds the education/growth relationship, we'll use a technique called *brushing*—that is, highlighting a subset of points in a scatter plot to illustrate some third factor. Above we see nine countries brushed and labeled: Taiwan, Hong Kong, Indonesia, Thailand, the Philippines (its label lost in a sea of mediocrity), Malaysia, Japan, Singapore, and South Korea. This small group of nine east-Asian nations accounts for the eight fastest-growing countries in our sample.

These high-growth countries, moreover, are near each other geographically but all over the map in terms of education spending. Might there just be something special about these economies from 1960–1996 that has nothing to do with education? Such a possibility is especially worrying when we consider how these countries might affect our understanding of the underlying relationship be-

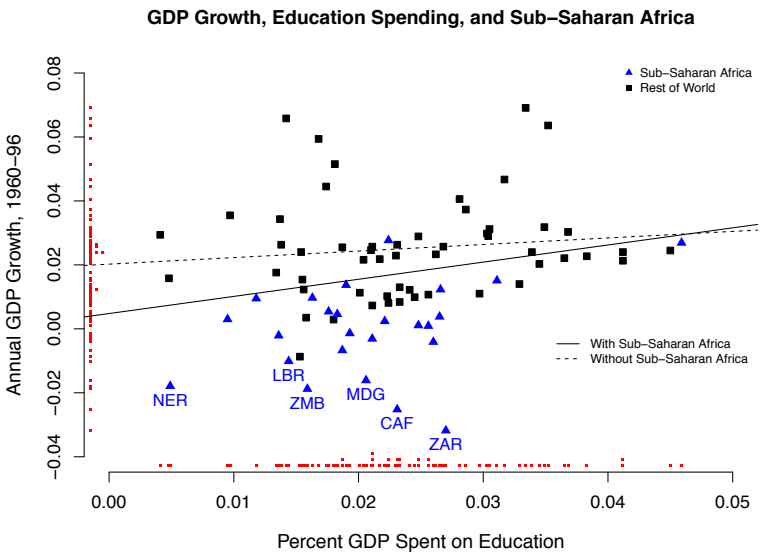


Figure 2.16: The 25 countries from sub-Saharan African have been highlighted, and those with the worst six GDP growth rates have been labeled.

tween education and growth. The two lines are the least-squares fits both with and without the east-Asian countries included in the sample. Note how the presence of the "Asian tiger" economies shifts our estimate of the best-fitting line systematically upward.

In Figure 2.16, we see another potential geographic confounder. Our sample includes 25 nations from sub-Saharan Africa, all brushed as blue triangles. This group accounts for the worst six growth rates—Niger, Liberia, Zambia, Madagascar, Central African Republic, and Zaire—along with 14 of the worst 15.

This does fit our causal story, since most of these countries are in the bottom half of education spending. Might these paltry investments in schools be the cause of Africa's poor economic growth? To be sure, if we exclude these nations from the sample, our estimate of the regression line changes noticeably. In particular, without sub-Saharan Africa in the sample, education looks a lot less important in predicting GDP growth. Notice how the dashed least-squares line is almost flat compared to the solid one.

Yet just as the economies of east Asia might have special reasons for their fast growth that have nothing to do with education, so too might the economies of sub-Saharan Africa have special reasons for their relative stagnation. Whatever these reasons may be—if they exist—*ceteris paribus* is a distant dream.

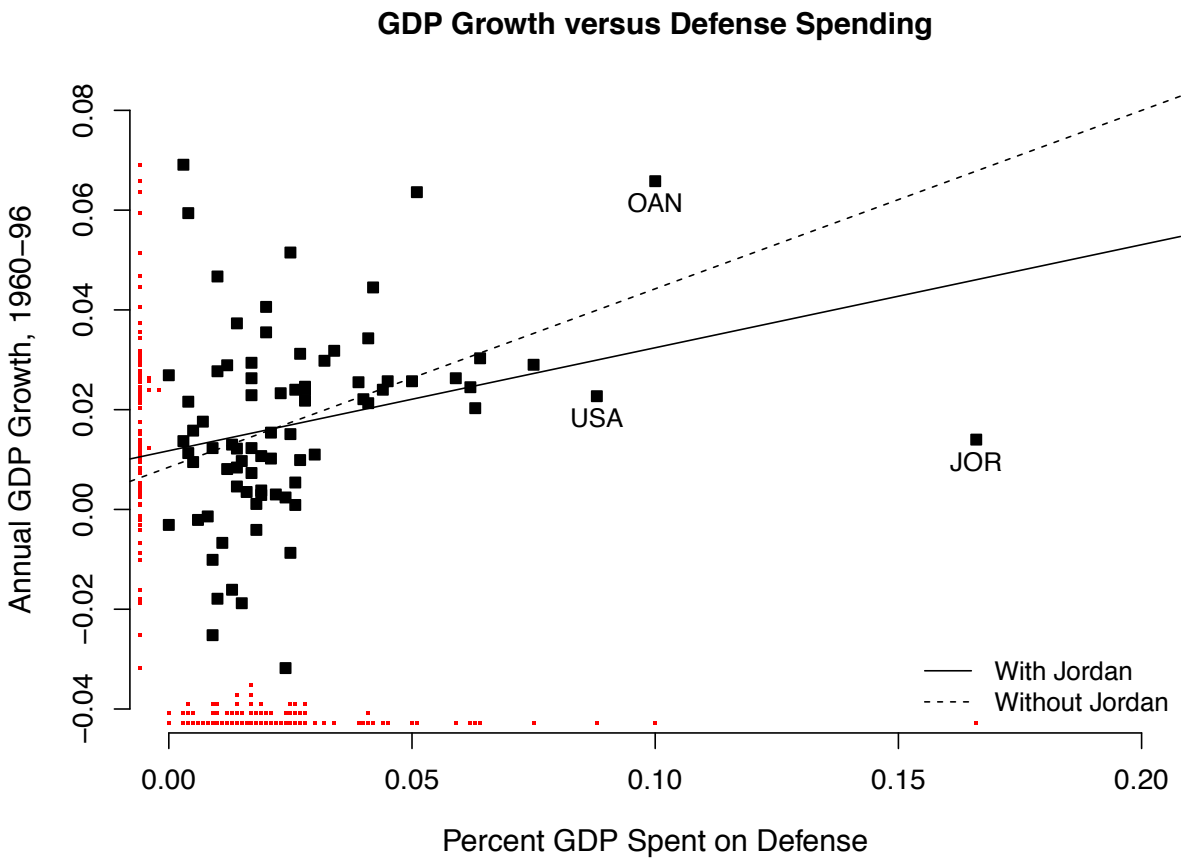


Figure 2.17: GDP growth since 1960 versus percent of GDP spent on national defense.

The guns question

Now let's turn to the second question, the one regarding the relationship between GDP growth and spending on military defense. In Figure 2.17 we see a scatter plot of these two variables. Keen observers may notice the following:

(1) Of the 40 countries that spent less than 2% of GDP on military defense, 27 fall below the median growth rate (1.58%).

(2) All 19 countries that spent more than 3% of GDP on military defense fall above the median growth rate.

The positive association between defense spending and growth seems, if anything, slightly stronger than the one where we use

education spending as the predictor. Of course, you should be wary by now of saying that defense spending is therefore a more direct cause of economic strength!

Also note the obvious outlier: the single nation all the way at the far right of the plot, at 16.6% of GDP spent on military defense, and firmly in the middle of the pack at 1.4% annualized GDP growth.

This outlier is not the United States, which comes in at a mere 8.8% of GDP spent on military defense. (Keep in mind this was back in the middle of the Cold War!) This is good for third place in the sample, just behind Taiwan (at 10% of GDP), and far behind Jordan.

You can probably think of some obvious geopolitical factors that might contribute to these two nations' elevated levels of military spending compared to the rest of the world. For now, though, simply compare the solid and dashed lines, and observe the large effect that Jordan exerts on the least-squares fit. It looks very much like a heavy weight, pulling the line downward about a fulcrum located near the middle of the sample, far away in the 2–3% range.

Appropriately enough, such a point is said to have high *leverage*. We can quantify this with an *influence plot*, below.

There is no formal statistical definition of an outlier, which is simply an observation far from the rest of the sample. Like Justice Potter Stewart said about pornography: you'll know an outlier when you see one.

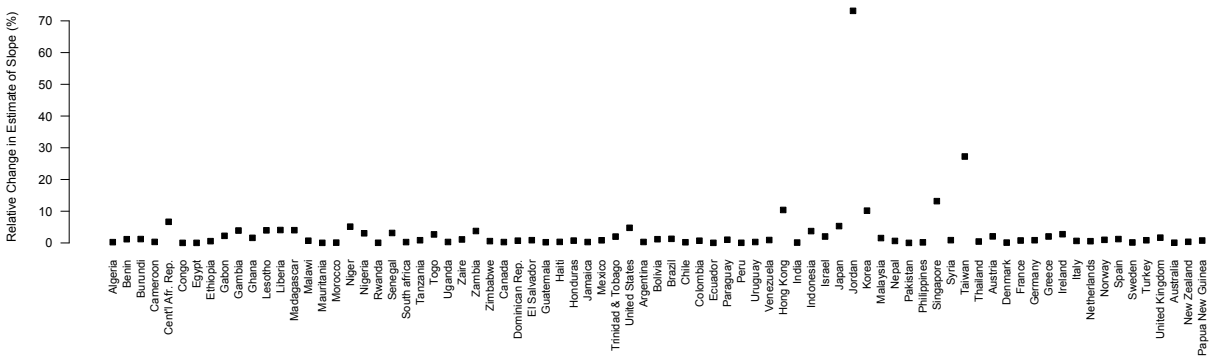


Figure 2.18: To construct an influence plot, we run 79 different least-squares fits, each time leaving one of the countries out. For each fit, we calculate how much the estimate of β_1, the defense-spending coefficient, changes compared to the estimate for the full data set. We express this as a percentage change—for example, when we leave out Jordan, the change in the regression coefficient is $|0.35 - 0.20|/0.20 = 75\%$.

As the plot shows, Jordan changes the estimated slope ($\widehat{\beta}_1$) by a relative margin of 75%, from 0.20 to 0.35. No other country even comes close to exerting such a strong influence over the line.

At least the qualitative conclusion—that defense spending and GDP growth are positively associated—remains intact, regardless of whether we include Jordan. But regression lines can be brittle things. In other cases the causal interpretation itself may turn on one or two decisive observations.

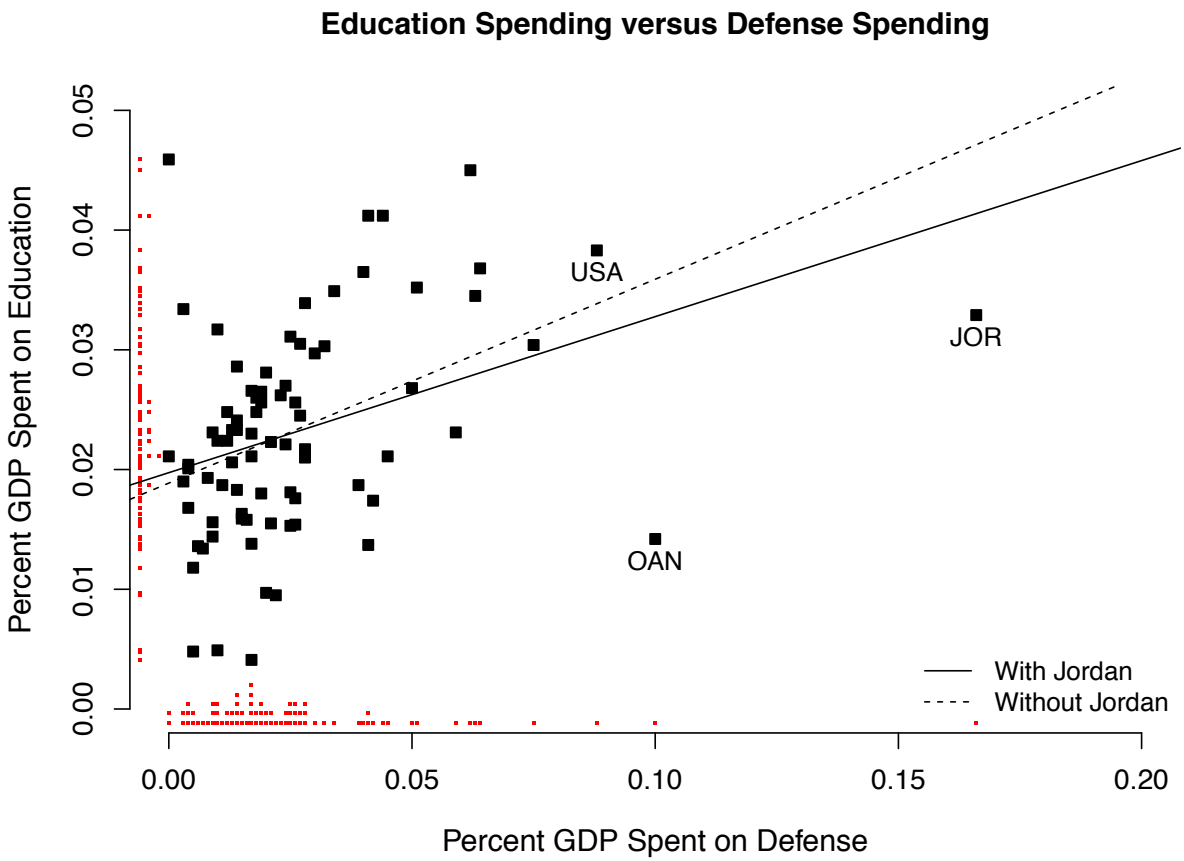


Figure 2.19: The relationship between defense spending and education spending.

The search for underlying causes

As it turns out, defense and education spending are positively associated both with GDP growth and with one another (as the plot above shows). In light of this, might there be some underlying cause that affects GDP growth, education spending, and defense spending all at once? The next two pages show some evidence for two such causal hypotheses. First, there's "rich get richer" hypothesis; it says that both future growth and current spending on luxuries are simply a reflection of current wealth. Second, there's the "healthy get wealthy" hypothesis; it attributes both spending and growth to the physical health of a nation's people.

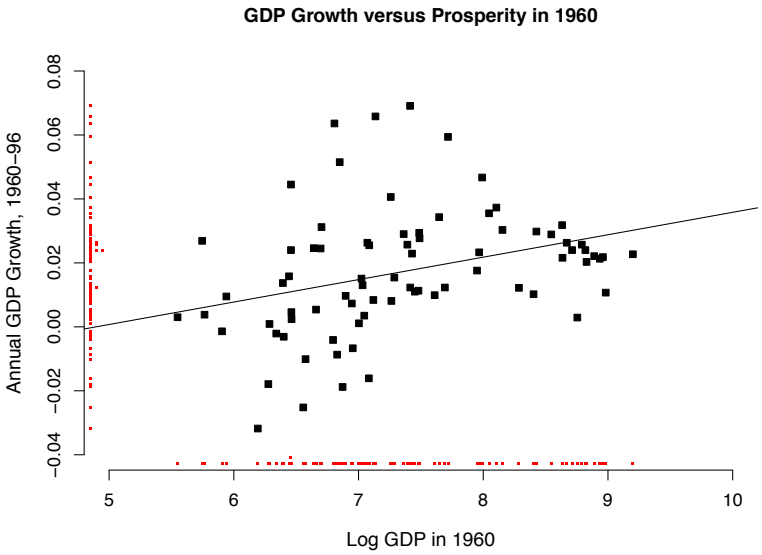


Figure 2.20: GDP growth versus beginning GDP (log scale) in 1960.

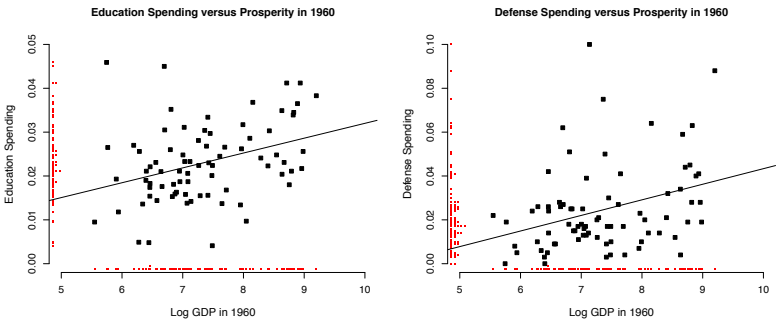


Figure 2.21: Defense spending and education spending versus beginning GDP (log scale) in 1960.

The pictures above are certainly consistent with the "rich get richer" hypothesis. Here we see that a nation's starting GDP in 1960 (as measured on a log scale) is positively associated both with subsequent GDP growth (top) and with elevated levels of spending on education and defense (bottom, left and right). They also cast doubt on the importance of education and defense spending as decisive factors in economic growth, since these expenditures can be reinterpreted as effects, not causes.

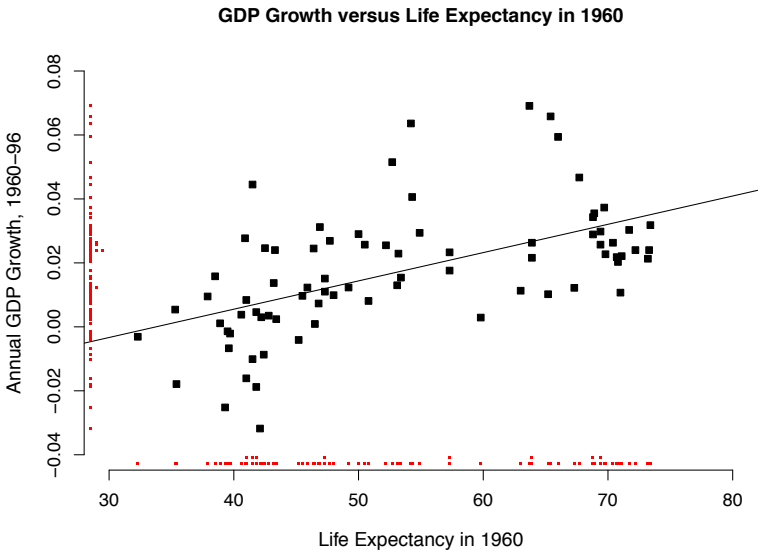


Figure 2.22: GDP growth versus life expectancy in 1960.

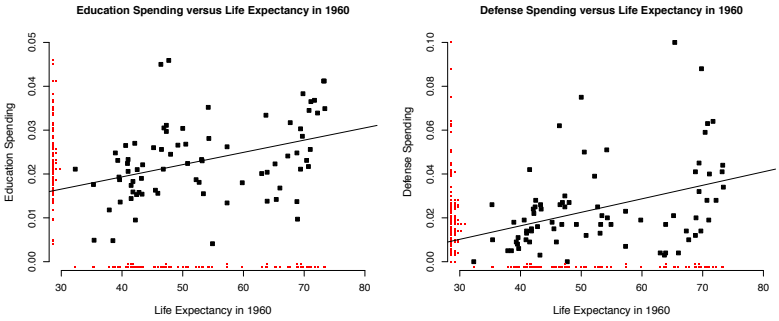


Figure 2.23: Defense spending and education spending versus life expectancy in 1960.

Is health the mystery factor? Only a country where most people died in old age, the argument might go, would believe it wise to spend money on schools or a military, in lieu of more basic, life-prolonging necessities. Above we see life expectancy in 1960 used as a predictor for GDP growth (top), education spending (bottom left), and defense spending (bottom right). In all three cases, the least-squares fit shows a positive association. The evidence at least does not rule out the "healthy get wealthy" hypothesis.

Lessons learned

We found several interesting relationships in the GDP-growth data.

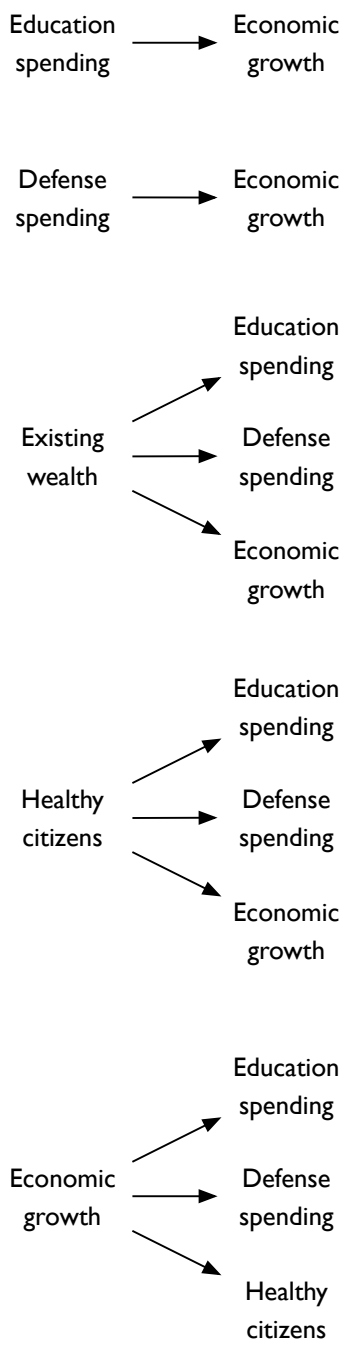
Countries with low growth tend to:	Countries with high growth tend to:
Spend less on education.	Spend more on education.
Spend less on defense.	Spend more on defense.
Be from sub-Saharan Africa.	Be from east Asia.
Have lower life expectancies.	Have higher life expectancies.
Have been poorer in 1960.	Have been richer in 1960.

We've entertained at least four different causal hypotheses that could explain these facts. (1) Investments in education drive economic growth. (2) Spending on military defense drives economic growth. (3) Existing wealth explains both subsequent economic growth and spending on education and defense. (4) A healthy, long-lived population explains both strong growth and elevated spending on education and defense.

At right we see the graphs for all four of these possibilities, along with a fifth: that growth itself is the cause, enabling people to spend their windfall profits on education, a strong military, fresh food, doctors, and all the other markers of a healthy, prosperous nation.

Although we could go on to consider many other possible confounders, let's stop here. The essential point has been made in the messy realities we've been forced to encounter: the confounding effects of high-growth and low-growth geographical clusters, the highly influential role played by just one or two outliers, and the fact that our two main predictors of interest are correlated with one another. On real problems, there is rarely any way of avoiding difficulties like these. Causal reasoning is a difficult business, and we simply must have data if we are to make any progress. But if those data do not arise from a controlled experiment, then skepticism must be the default position.

Figure 2.24: Some causal hypotheses.



Of course, that is why we learn statistics! We'll soon go beyond the rough comparisons we've made here to provide rigorous answers to questions like the following:

- How do we decide when one fit is better than another?
- How confident are we in the numerical description given by the least-squares regression line?

- How can we adjust for one predictor in evaluating the marginal effect of another predictor on the response?
- How can we decide whether a potential predictor is even relevant for describing changes in the response? (Remember Occam's razor.)

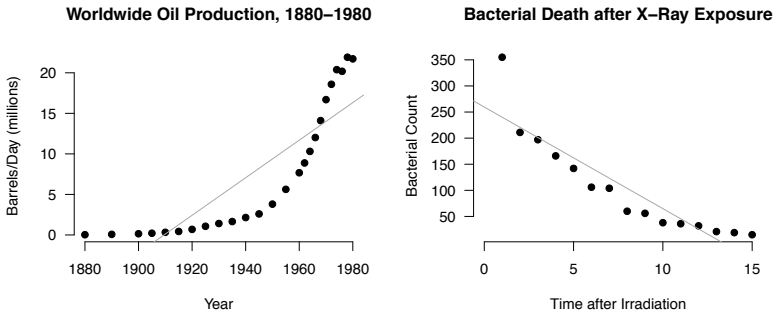
Answering these questions will involve computing a lot specific technical summaries, some of which you may have heard of before: standard errors, p-values, R^2, F-statistics, and so forth. But even as we get more technical in our descriptions, don't ever lose sight of two facts:

(1) These numbers are just quantitative versions of the intuitive notions you've already learned in the first two chapters, much like a mean is the quantitative version of "middle."

(2) Sometimes a picture is worth a thousand numbers.

Beyond straight lines

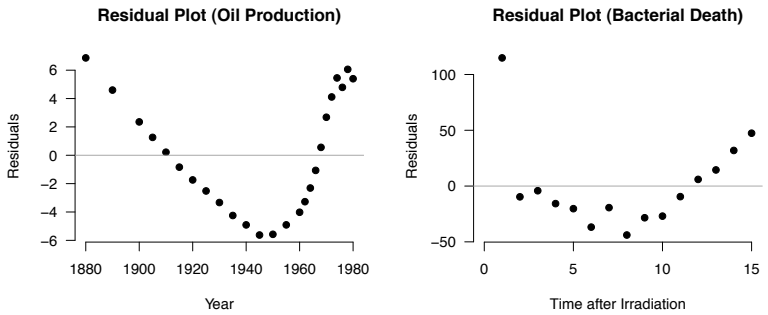
UP TO this point, we've talked about fitting straight lines to data. For many data sets, however, the "off-the-shelf" least-squares method just won't work:



The least-squares method does produce an answer, but the straight lines are clearly inadequate. The problem is equally obvious in the residuals:

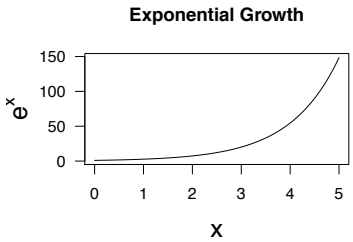
No random cloud in sight. Observe the systematic snake-like path traced out by each set of residuals: first above the line, then back below, and then above again. This should raise a red flag.

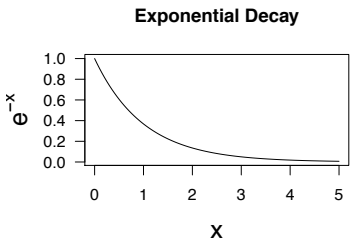
Now that we know a straight line isn't going to cut it, let's turn back to the original data, where in each case we see the least-



squares line failing miserably to capture the underlying relationships between x and y. The reason is that these relationships are fundamentally nonlinear: we're asking a square peg to fit in a round hole.

So if not straight lines, what do the real regression functions look like? In the first case, oil production seems to follow an exponential growth curve as a function of time. In the second case, bacterial count seems to follow an exponential decay curve as a function of time. (See the illustrations to the right depicting exponential growth and exponential decay.)





Thus the real underlying relationships are probably best described not by lines, but by something like

$$Y \approx e^{\beta_0 + \beta_1 X}\,.$$

In the case where y is oil production and x is time, the growth gets faster as x increases, suggesting that β_1 is positive. And in the case where y is bacterial count and x is time, the decay gets slower as x increases, suggesting that β_1 is negative.

But so far we've only learned how to use least squares for fitting straight lines, not exponential curves. How are we to proceed? Let's try the following trick: take the logarithm of both sides in the above equation. Recalling that the log and the exponential are "inverse" functions of one another, we end up with:

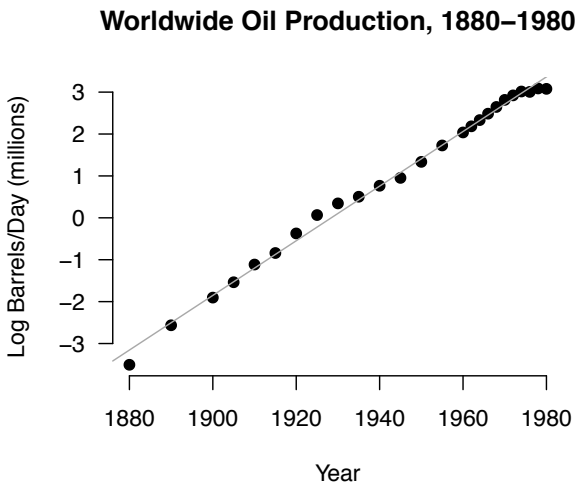
$$\log\{Y\} \approx \beta_0 + \beta_1 X\,.$$

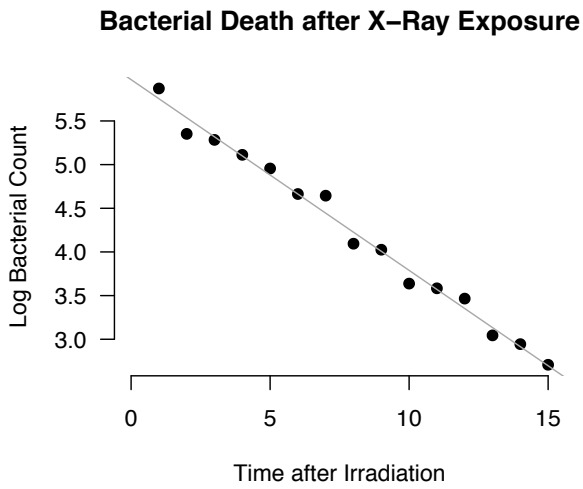
On the left, we have something denominated in the logarithm of whatever the y variable is. And on the right, we have a linear function of x. This should look comfortingly familiar. Let's follow the logic suggested by this equation and plot the natural logarithm (base e) of our y samples versus x, as on the bottom of the previous page.

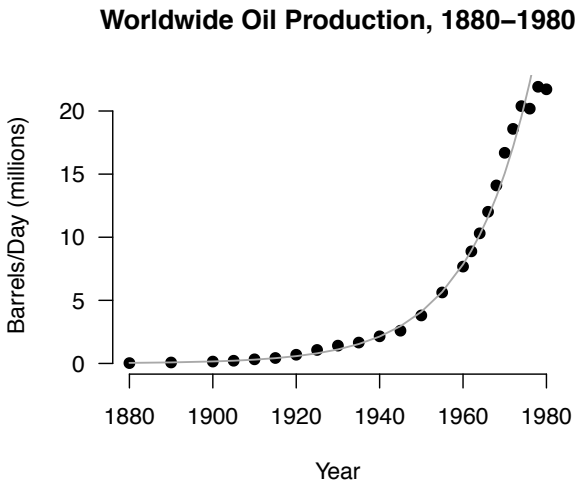
Clearly straight lines will be much more meaningful here than for the original data set. All we have to do to make the least-squares method work here is to define a new variable $z_i = \log y_i$, and write z as a linear function of x:

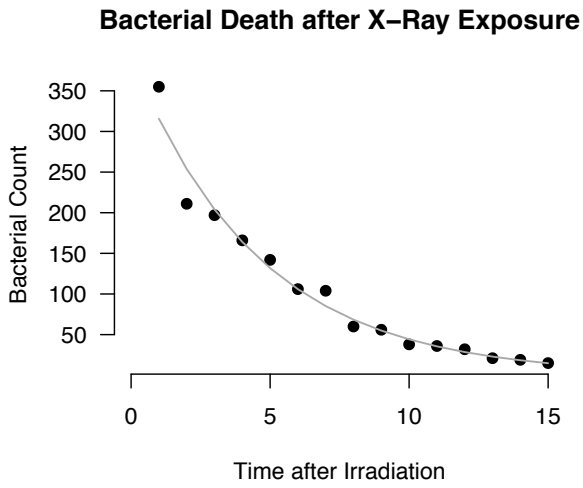
$$\log y_i = z_i = \beta_0 + \beta_1 x_i + \epsilon_i \,.$$

The residuals ϵ_i are needed here for the same reason they were before: because no straight line can fit the data without some









wiggle room to miss each point by a little bit. If we then use the least-squares criterion to let the data choose particular values for β_0 and β_1, we end up with:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ \widehat{\beta}_0 &= \bar{z}-\widehat{\beta}_1\bar{x}\\ \hat{z}_i &= \widehat{\beta}_0+\widehat{\beta}_1 x_i\,,\end{aligned}$$

where the $\hat{z}_i$'s are the fitted values of the transformed variable. We can see these lines fitted to the data in the plot above.

If we now want to return to the original scale of measurement for y, we compute the fitted values for y as $\hat{y}_i = e^{\hat{z}_i}$. By exponentiating, we "undo" the log transformation, producing a curve back on the original scale:

The fitted values now fall along exponential curves that have been determined entirely from the data. In the margin we see two tables depicting the "before" and "after" results. These allow us to compare how precise the new fits are, compared to the original linear fits.

This process is called a *transformation of variables*. In this case we performed a log transformation of the y variable. Transformations are a very general technique for making linear least-squares work for nonlinear relationships. In each of these two cases, the transformation buys a substantial boost in R^2 without introducing any extra parameters that must be fit from the data. Clearly the curves are better at predicting y than the straight lines are, at least within the confines of our sample.

Table 2.5: The oil-production data.

	Original	Transformed
$E(Y \mid X)$	$-442 + 0.23x_i$	$e^{-125+0.065x_i}$
R^2	0.743	0.995

Table 2.6: The bacterial decay data.

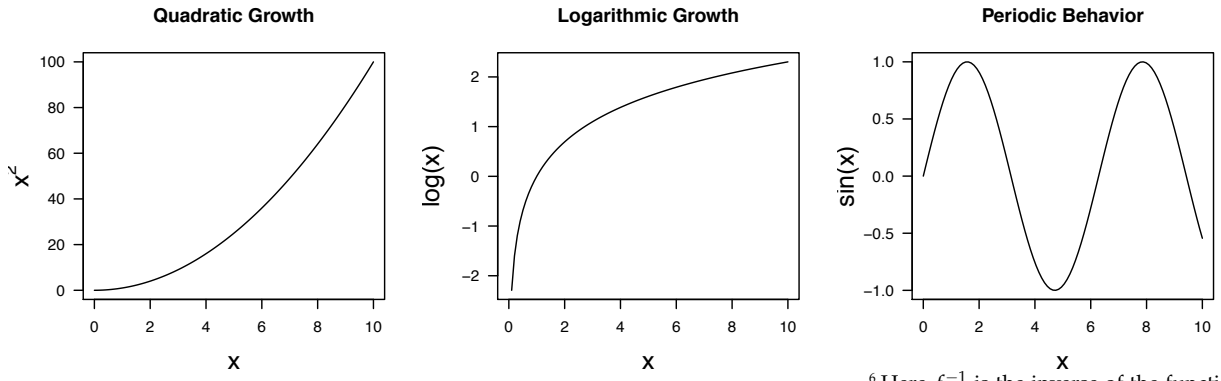
	Original	Transformed
$E(Y \mid X)$	$260 - 19.5x_i$	$e^{5.97-0.22x_i}$
R^2	0.823	0.988

When to use a transformation, and which ones to use

We've already encountered an obvious situation where a transformation should be used: when the y variable clearly looks like some nonlinear function of x. In addition to exponential growth and decay, some other possibilities for f include quadratic, logarithmic, and trigonometric functions.

To put it as generally as possible, if $y \approx f(x)$, then you will have to do one of two things to make linear least squares work:

(1) use $w_i = f(x_i)$ as a transformed predictor variable, and fit a line to y_i versus w_i.



(2) use $z_i = f^{-1}(y_i)$ as a transformed response variable, and fit a line to z_i versus x_i.[6]

The choice between these two options would seem, at first, to be a small matter.[7] For example, you might guess that, in fitting the oil-production or bacterial-decay data sets, we could instead have transformed the x variable, writing the model as $E(y \mid x) = \beta_0 + \beta_1 e^x$. This, too, is an exponential function of x, and so would seem to fit the bill. In this case we would use $w_i = e^{x_i}$ as our predictor variables, and fit a least-squares line for y_i given w_i.

The problem is that this is simply not a very flexible regression function. In particular, the rate of growth of this exponential curve is fixed at 1, whereas before the rate was estimated from the data (and was equal to β_1). Since in most exponential growth (or decay) scenarios, the relevant parameter to be estimated is the rate of growth (or decay), it is important to take the logarithm of y, rather than the exponential of x.

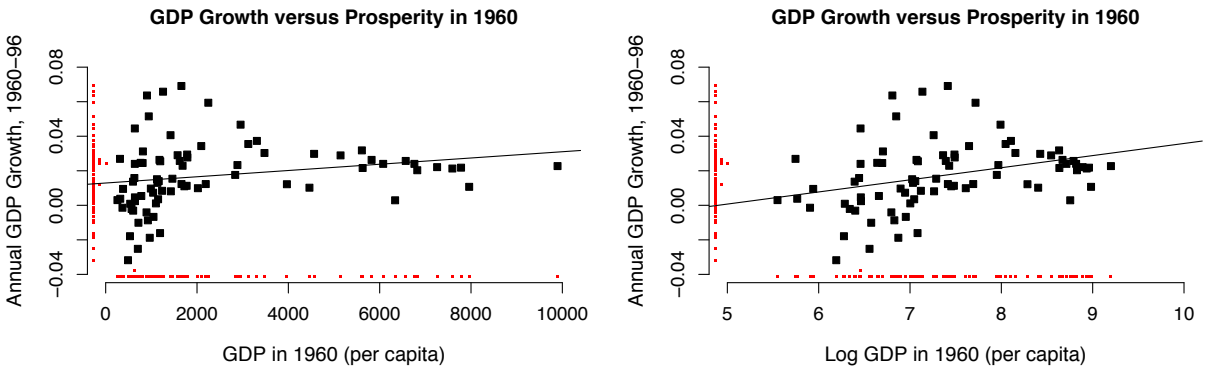
In most other cases, you should use option 1. We are usually interested in predicting y, after all, and it makes more sense to predict on the original scale. Also, some functions lack a unique inverse, making $f^{-1}(y_i)$ ambiguous. If this is the case, you have little choice but to use option 1. For example, the inverse of a quadratic function is a square root—but should we take the positive or negative square root? Inverse trigonometric functions fall peril to this problem, as well: $\arccos(1)$ could be 0, or 2π, or 4π, or. . . .

Log transformations of the x variable will be especially useful when the data are skewed to the right—that is, when there is a long right tail. Indeed, we've already seen an example of this, back when we plotted GDP growth versus starting GDP in 1960. Most

[6] Here f^{-1} is the inverse of the function that appears to describe the relationship between y and x. For example, the logarithm is the inverse of an exponential function, and the square-root is the inverse of a quadratic function.

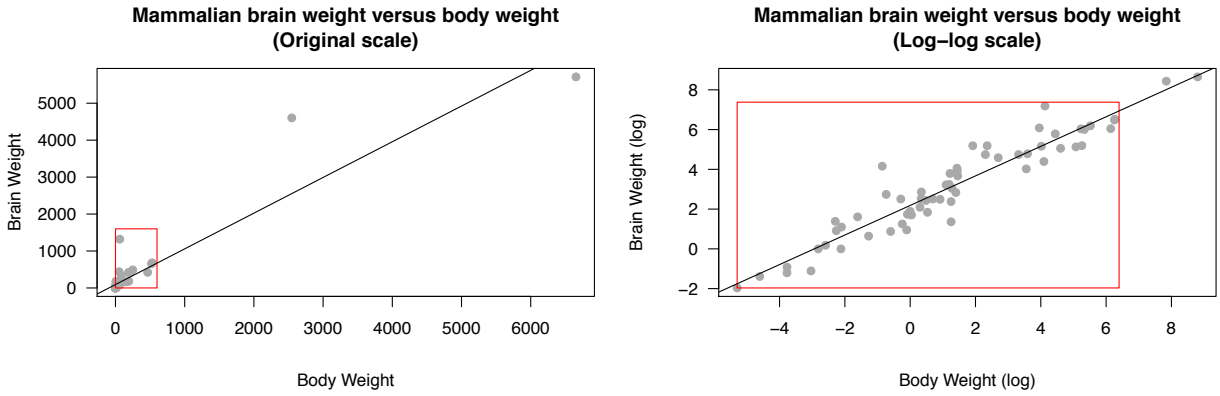
[7] Either approach is quite straightforward: it simply involves making a new variable in your data frame that applies the appropriate formula to the original variable, and using the new variable in your regression instead. In R, just define a new variable using the relevant function of the old one.

countries are bunched up near the bottom of the scale, and the few countries with higher GDP's tend to stretch out far to the right.



A log transformation of the x variable stretches out the lower end of the scale and squeezes in the upper end, resulting in a much more symmetric distribution along the x axis.

In some cases, it may be best to take logs of both variables:



This is the right tactic when both of the original axes need to be stretched at the lower end and squeezed at the upper end, as in the above data set relating mammalian brain weight and body weight. Notice that, in each of the two plots, the red box encircles 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 elephant and the blue whale) have been forced to cede some real estate to the rest of Mammalia!

Interpreting the coefficients under a transformed model

The downside of transforming either the x or the y variable is that it changes the intepretation of the regression coefficients. For example, for the data set where y is GDP growth and x is starting GDP in 1960, the original regression function is

$$E(y \mid x) = \beta_0 + \beta_1 x_i \,,$$

while the transformed function is

$$E(y \mid x) = \beta_0 + \beta_1 \log(x_i) \,.$$

In the first case β_1 multiplies GDP, whereas in the second case it multiplies the logarithm of GDP. Different units, different interpretations.

A little calculus can help us see the difference. In general, if $y = \beta_0 + \beta_1 f(x)$, then using the chain rule for taking derivatives, we have

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \beta_1 f'(x) \approx \beta_1 \frac{\Delta f(x)}{\Delta x} \,,$$

where Δ means "change in." Equivalently,

$$\beta_1 \approx \frac{\Delta y}{\Delta x} \frac{\Delta x}{\Delta f(x)} = \frac{\Delta y}{\Delta f(x)} \,.$$

Originally, β_1 expressed the rate of change of y for a given change in x. Now, after transforming the x variable, β_1 expresses the rate of change in y for a given change in $f(x)$. Thus the interpretation of the "slope" depends very much upon which transformation is used.

A very interesting case is when we take logs of both variables, leaving us with the following model:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i \,.$$

In this case, taking the derivative of both sides gives us:

$$\frac{\mathrm{d}y}{\mathrm{d}x} \frac{1}{y} = \beta_1 \frac{1}{x} \,.$$

Rearranging terms, we get

$$\beta_1 \approx \frac{\Delta y / y}{\Delta x / x} \,.$$

In other words, β_1 measures the ratio of percentage change in y to percentage change in x. Such a term is often called an *elasticity* parameter, especially in economics. For example, on the mammalian brain-weight data, the least-squares estimate of the slope on a log-log scale was $\widehat{\beta}_1 = 0.74$. This means that, among mammals, a 100% change in body weight is associated with a 74% expected change in brain weight. The bigger you are, it would seem, the smaller your brain gets—at least relatively speaking.

A dictionary of transformations

THE GOAL of this dictionary is to provide you with a basis for comparison, so that you may more easily diagnose and correct potential nonlinearities in any real data you encounter.

Logarithmic functions

If the true model that generated your data is a function of $\log x$, then the plot of y versus predictor j may look something like this:

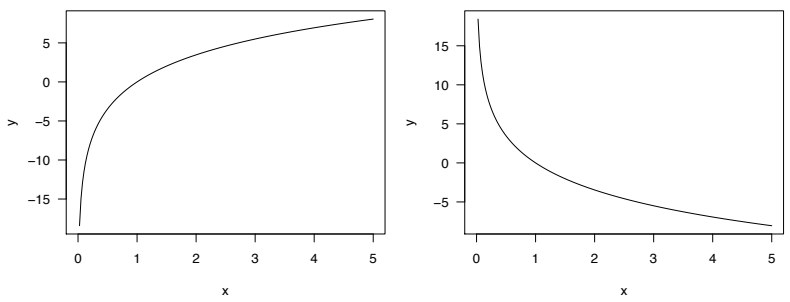


Figure 2.25: Graphs of $y = \beta \log x$ for $\beta > 0$ (left) and $\beta < 0$ (right).

These are graphs of $y = \beta \log x$. The graph on the left corresponds to $\beta > 0$, and the graph on the right corresponds to $\beta < 0$. If you think your data look like a logarithmic function of x, then there's a simple fix: try re-fitting the model with $\log x$ replacing x as a predictor. From a mathematical standpoint, it doesn't matter what base you use, but using the natural logarithm (base e) makes the resulting regression coefficient the easiest to interpret.

Power laws

If the true model that generated your data is a function of x to some power c, then the plot of Y versus X may look something like the following.

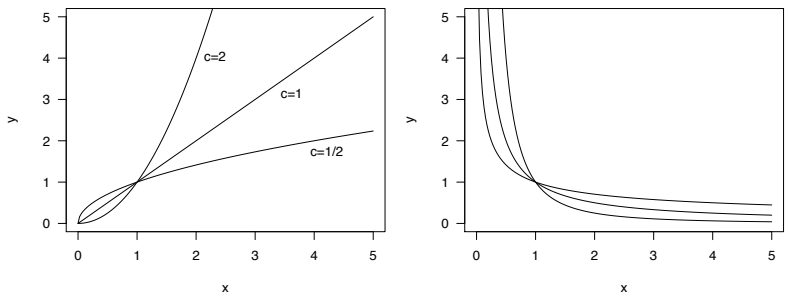


Figure 2.26: Graphs of $y = \beta x^c$ for different values of c, where $\beta > 0$ (left) and $\beta < 0$ (right).

In comparing these graphs to the previous set, you may find it hard to distinguish between two sets of cases. First, logarithmic growth versus fractional power-law growth can be hard to tell apart at first. The graphs of $y = \log x$ and $y = x^{1/2}$, for example, look similar in many respects. The key distinguishing feature is that $x^{1/2}$ never goes negative, while $\log x$ does (for values of x between 0 and 1).

It can also be tough to spot the difference between logarithmic decay and power-law decay—compare, for example, the graphs of $y = -\log x$ and $y = x^{-1}$. The key distinguishing feature is the same as above: the power-lay decay curve never goes negative, approaching 0 asymptotically. On the other hand, $y = -\log x$ keeps decreasing forever as x gets larger.

There are two possible fixes to make linear regression work for power laws.

Add powers of x. For example, if y looks like a quadratic function of x, then try adding x^2 to the regression. The one caveat here is: if you add higher powers of x, always make sure to retain the lower powers of x. You will never, for example, add x^3 to a model without also having x and x^2 in there, as well.

Take the logarithm of both x and y. This is generally the right way to go if you can't figure out what power of x to add. That's because the power itself becomes the regression parameter on

a log-log scale: if you fit $\log y = \beta_0 + \beta_1 \log x + \epsilon_i$, then some simple algebra shows that

$$E(y \mid x) = e^{\beta_0} \cdot x^{\beta_1} .$$

Exponential growth and decay

If the true model that generated your data is an exponential function of x, then the plot of y versus predictor j may look something like this:

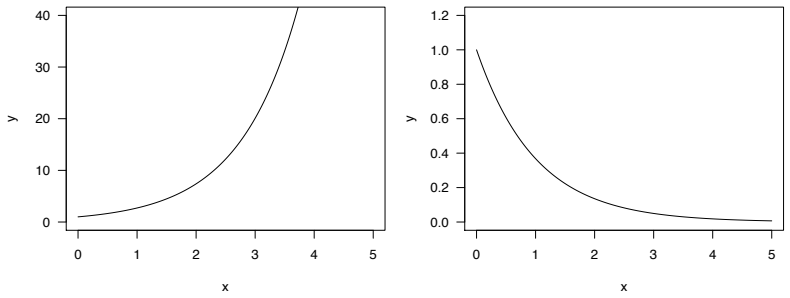


Figure 2.27: Graphs of $y = e^{\beta x}$, where $\beta > 0$ (left) and $\beta < 0$ (right).

Again, we have some challenging cases to tease apart. First of all, e^{-x} looks a bit like the power-law decay curves. The distinguishing feature is that e^{-x} is equal to 1 when $x = 0$, whereas x^{-1} and other power-law decay curves blow up to infinity in the limit as x approaches 0.

The hardest cases to tell apart are power-law growth and exponential growth. The main distinguishing feature is that exponential growth is simply much, much faster than power-law growth. You can see that x^3 and x^4 jump out ahead of e^x at first. But the exponential growth curve eventually catches up—and indeed would eventually catch up to any power of x, even those much bigger than x^4.

x	1	2	3	4	5	6	7	8	9	10
x^2	1	4	9	16	25	36	49	64	81	100
x^3	1	8	27	64	125	216	343	512	729	1000
x^4	1	16	81	256	625	1296	2401	4096	6561	10000
e^x	3	7	20	55	148	403	1097	2981	8103	22026

The right way to handle exponential growth and decay curves

is to take the logarithm of your y variable. This turns out to be a much more flexible way of fitting functions than taking the exponential function of x, because it allows us to infer the rate of growth or decay. Keep in mind, of course, that taking the logarithm of y doesn't just change the interpretation of the y–x relationship. It also changes the interpretation of the relationship between y and all of the other predictors that enter into the model.