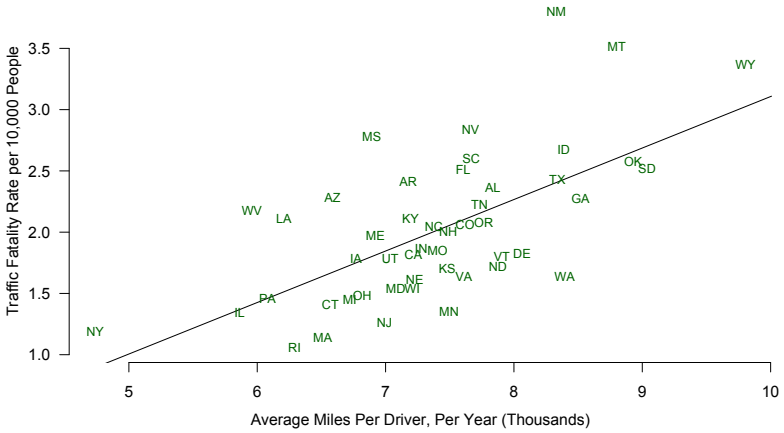
3
Predictable and Unpredictable Variation

Residuals as part of the model

In 1983, roughly 2 Americans in every 10,000 died in a traffic accident. But this rate varied considerably across the states—almost four times higher in New Mexico (3.79), for example, than in Rhode Island (1.04). As the figure below shows, much of this variation can be described by differences in the number of miles logged by drivers in each state.



But the data points are dispersed about the least-squares line; there are clearly other factors at work. (The severity of sentences for drunk drivers? Speed limits? Taxes on alcohol? Safer cars? Socio-economic predictors?) The natural question is: after adjusting for miles driven, how much variation remains to be explained by these other factors?

Similar questions arise in almost all statistical models:

- Mammals more keenly in danger of predation tend to dream fewer hours. But there is still residual variation that practically begs for some kind of Zen proverb. (Why does the

water rat dream at length? Why does the wolverine not?)

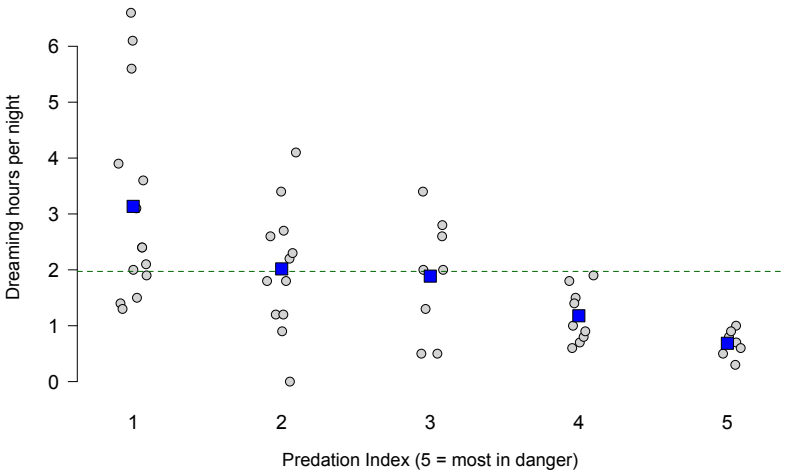


Figure 3.1: Dreaming hours per night versus danger of predation for 50 mammalian species. In this and in Figure 3.2, the blue squares show the group-wise means, while the dotted green line shows the grand mean for the entire data set.

- 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. How precisely could the power company forecast peak demand, using only time of year as a predictor?

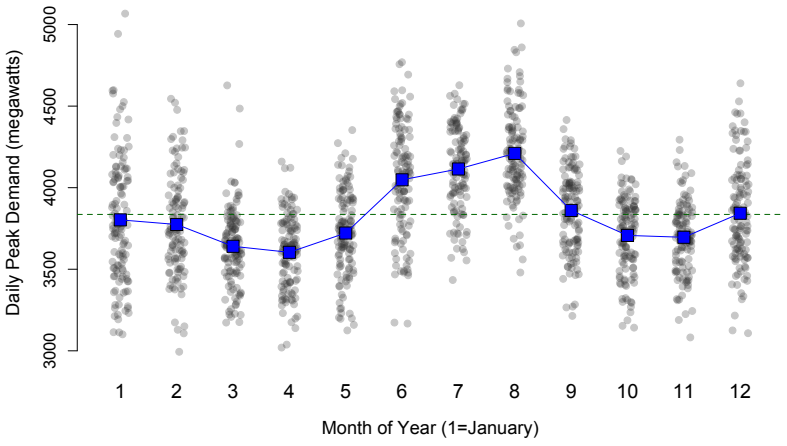
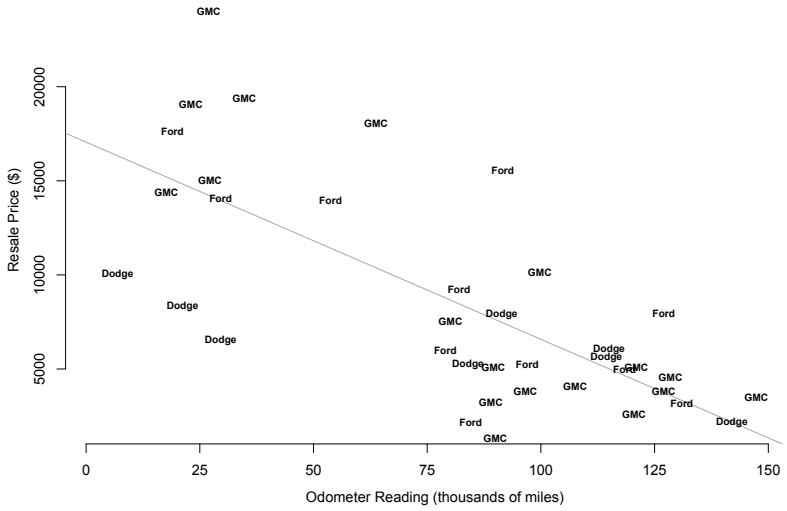


Figure 3.2: Daily peak demand for electricity versus month of the year in Raleigh, NC from 2006–2009.

- Among pickup trucks advertised for sale on Craigslist, those with higher odometer readings tend to have lower asking prices, just as you'd expect:



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, on average. If the owner is asking $11,000, is this reasonable, or drastically out of line with the market? Does your assessment change depending on whether the truck is a Ford, GMC, or Dodge?

In all of these cases, one must remember that the fitted values from a statistical model are generalizations about a typical case, where "typical" takes into account the information from the predictors. 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 just a recipe for calculating the fitted values, and to think that the residuals are what's "left over" from the model. This is a conceptual error: we'll have a richer picture if we see the residuals as part of the model. If you've ignored them, or don't have a sense for how big they could be, then you haven't specified a complete statistical model.

The crucial 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 as the examples above convey, they are rarely an end to the story.

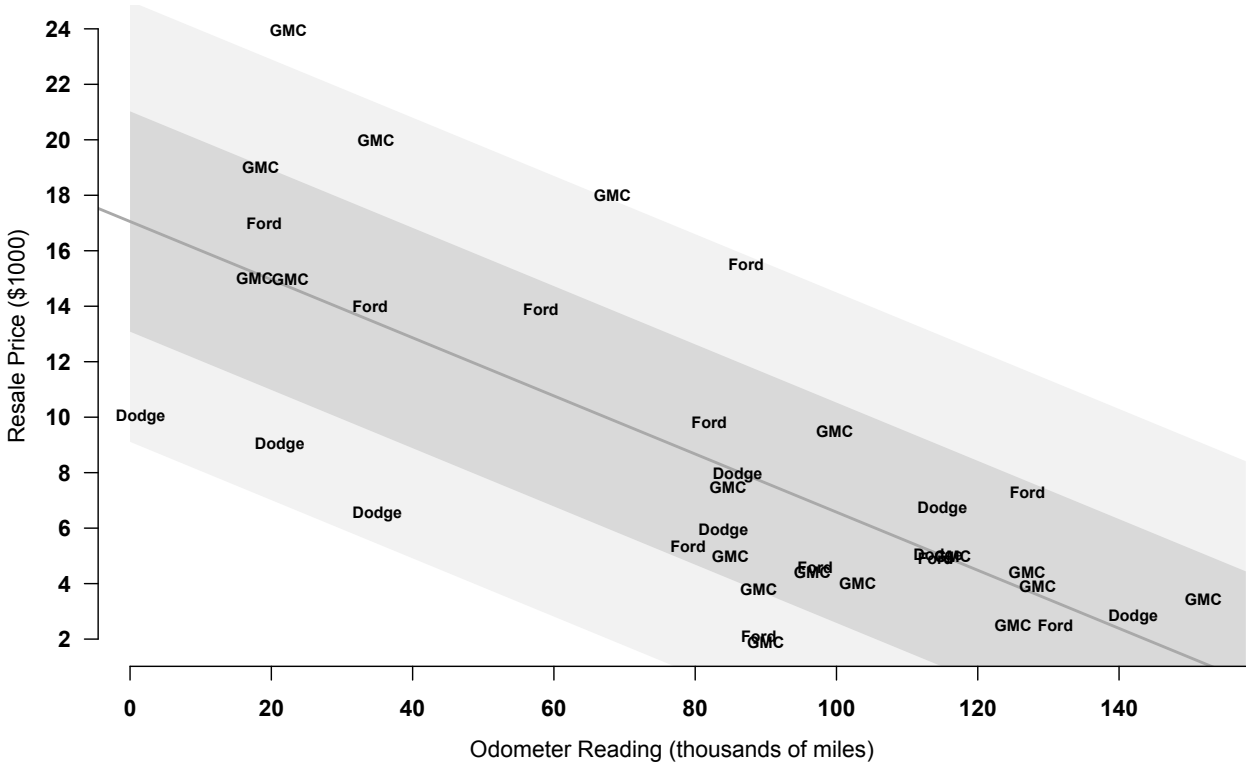
Naïve prediction intervals

We have already learned a handful of tools for measuring the variation of a typical case in a data set—the sample variance and standard deviation, box plots, histograms, dot plots, and so forth. All of these tools can be applied directly to the residuals, which will allow us to answer the question: "By how much does a typical case vary from the prediction of the model?"

To see how this works, let's take another look at the data set of pickup trucks advertised on Craigslist. Below, the dark grey line bisecting the point cloud is the least-squares fit: $Y = 17054 - 0.105\ x$. But the residuals are part of the model, too! Their sample standard deviation is \$3,971, compared to a "raw" standard deviation of \$5,584 for the observed truck price—that is, a "typical" truck deviates from the sample mean by about \$5,584, and from the least-squares line by about \$3,971.[1]

The two grey strips below depict this uncertainty visually. The medium grey strip extends to 1 residual standard deviation (line ±

[1] You will have noticed that the sample standard deviation of the residuals (3971) is smaller than the standard deviation of the raw truck prices (5584). This is as it should be; the predictor contains information about truck prices, and this extra information reduces our uncertainty about the likely value of a truck's price. In fact, this suggests a natural way to measure the information content of a predictor in a statistical model—the more a predictor reduces our uncertainty, so the thinking goes, the more information it contains. We'll build on this notion in the next section.



$3,971) on either side of the line, while the light grey strip extends to 2 residual standard deviations (line ± $7,942). These grey strips not only summarize the typical degree of variation from the line, but can be used as interval estimates for a future case. 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 the one-standard-deviation interval estimate is $8,672 ± $3,971, or the interval $(4701, 12643)$. Our hypothetical asking price of $11,000 is well within this interval.

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 *naïve 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 $(8892, 16834)$.

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" or "is an element of of."

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

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

or more concisely, $y \in \hat{y} \pm \alpha \cdot s_e$. Here s_e is the residual standard deviation and α is a chosen multiple that characterizes the width of the intervals.[2] As our discussion above hints, typical values for α are 1 or 2. To quantify the accuracy of the interval, we look at the empirical coverage: that is, what fraction of examples in our original data set are contained within their corresponding interval.

[2] There is a clear trade-off here: larger choices of α mean wider intervals mean more uncertainty, but greater coverage.

We call these prediction intervals "naïve" because they ignore uncertainty about the parameters of the model itself, and only account 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, they actually understate the total amount of uncertainty that we'd ideally like to incorporate into our interval estimate. We'll soon learn how to quantify these additional forms of uncertainty. But imperfections aside, even a naïve prediction interval is more useful than a point estimate.

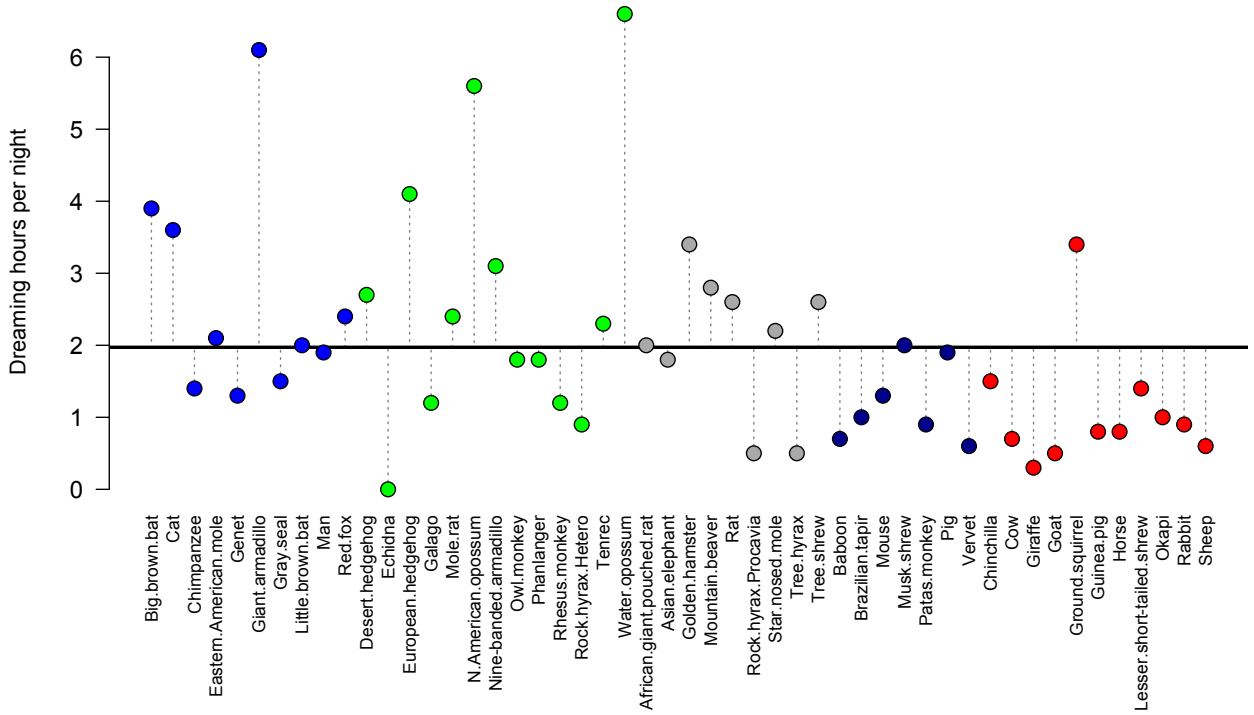


Figure 3.3: Dreaming hours by species, along with the grand mean. For reference, the colors denote the predation index, ordered from left to right in increasing order of danger (1–5). The vertical dotted lines show the deviations from the grand mean: $y_i - \bar{y}$.

Partitioning sums of squares

QUANTIFYING the information content of a predictor brings us straight back to a question we posed earlier: what's so great about sums of squares for measuring variation? To jump straight to the punch line: 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.

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 3.3, 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, $y_i - \bar{y}$.

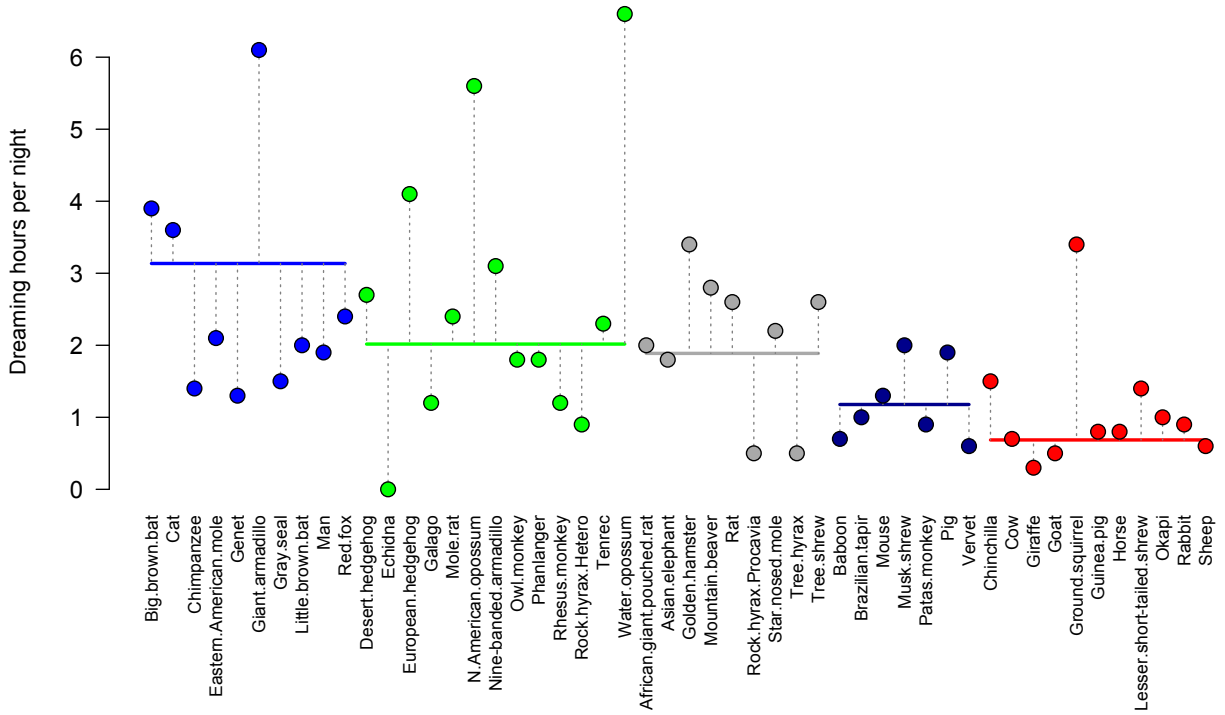


Figure 3.4: Dreaming hours by species, along with the group means stratified by predation index. The vertical dotted lines show the residuals from the group-wise model "Dreaming hours ~ predation index."

To account for the information in the predictor, we fit the model "dreaming hours ~ 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.
- 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 3.4 and again as diamonds in Figure 3.5. 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.

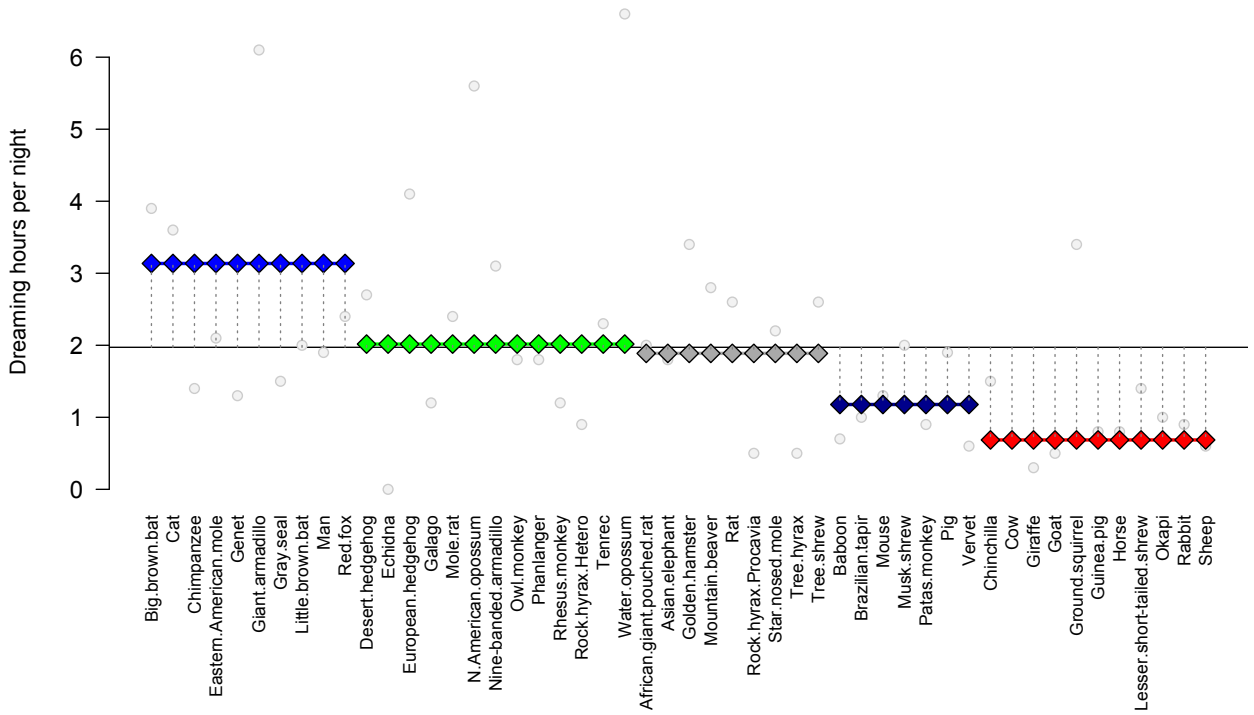


Figure 3.5: Dreaming hours by species (in grey), along with the fitted values (colored diamonds) from the group-wise model using predation index as a predictor. The vertical lines depict the differences $\hat{y}_i - \bar{y}$.

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:

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

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

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

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

- 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:

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

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

$$102.1 = 36.4 + 65.7\,,$$

so that TV = PV + UV! It appears that 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:

$$\begin{aligned}\sum_{i=1}^{n} |y_i - \bar{y}| &= 53.0\\ \sum_{i=1}^{n} |\hat{y}_i - \bar{y}| &= 33.7\\ \sum_{i=1}^{n} |y_i - \hat{y}_i| &= 42.5\,.\end{aligned}$$

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:

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

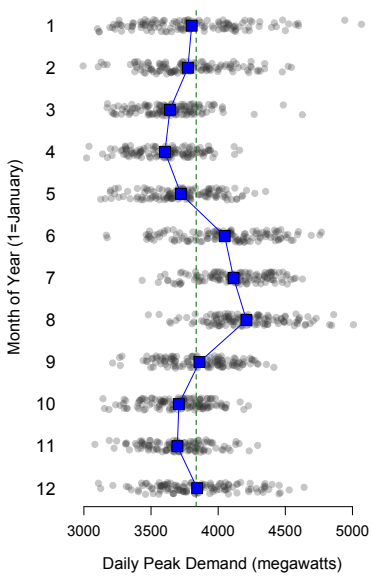
Next, we sum up the squared deviations of the fitted values. For each observation, the fitted value is just the group-wise mean for the corresponding month, given by the blue dots at right:

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

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

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

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?

$$\begin{aligned}\sum_{i=1}^{n} |y_i - \bar{y}| &= 397{,}887.7 \\ \sum_{i=1}^{n} |\hat{y}_i - \bar{y}| &= 220{,}382.1 \\ \sum_{i=1}^{n} |y_i - \hat{y}_i| &= 325{,}409.0\,.\end{aligned}$$

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.

What's more, a similar decomposition also holds for linear regression models. In Figure 3.6 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!

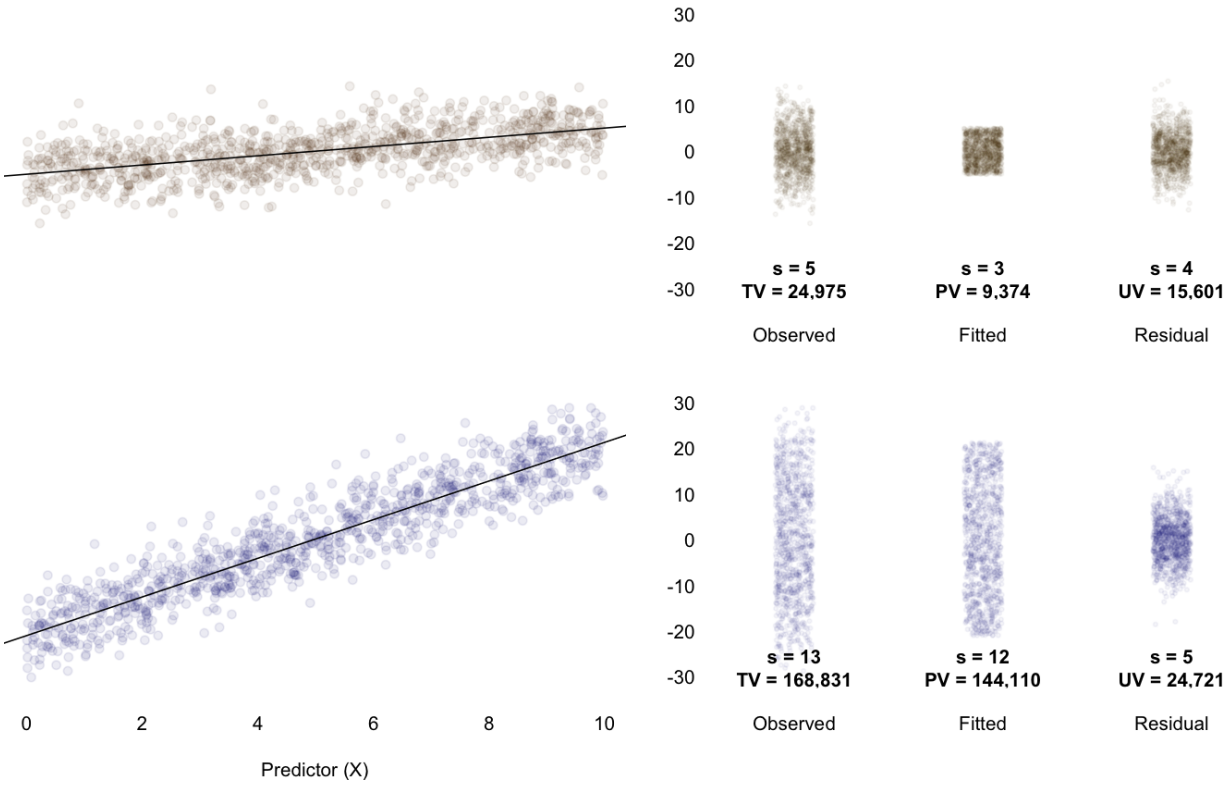


Figure 3.6: Two imaginary data sets, along with their least squares lines.

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:

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

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!

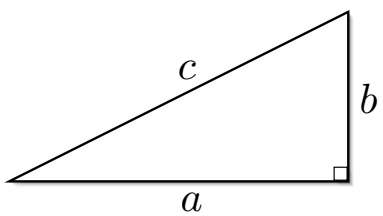
As we've repeatedly mentioned, it would be perfectly sensible to measure variation using sums of absolute values $|y_i - \hat{y}_i|$ instead, or even something else entirely. But if we were to do this, the analogous "TV = PV + UV" decomposition would not hold as a general rule:

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

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 both beautiful and, as you'll soon discover, very powerful. Yet it will probably strike you as 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 DVD's that you own in common. It makes no sense to say: "Well, there are 10,000 (100^2) squared-DVD's in total, so I'll take 3,600 (60^2) squared-DVD's, and you take the remaining 1,600 (40^2)." Not only is the statement itself barely interpretable—what the heck is a squared DVD?—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.



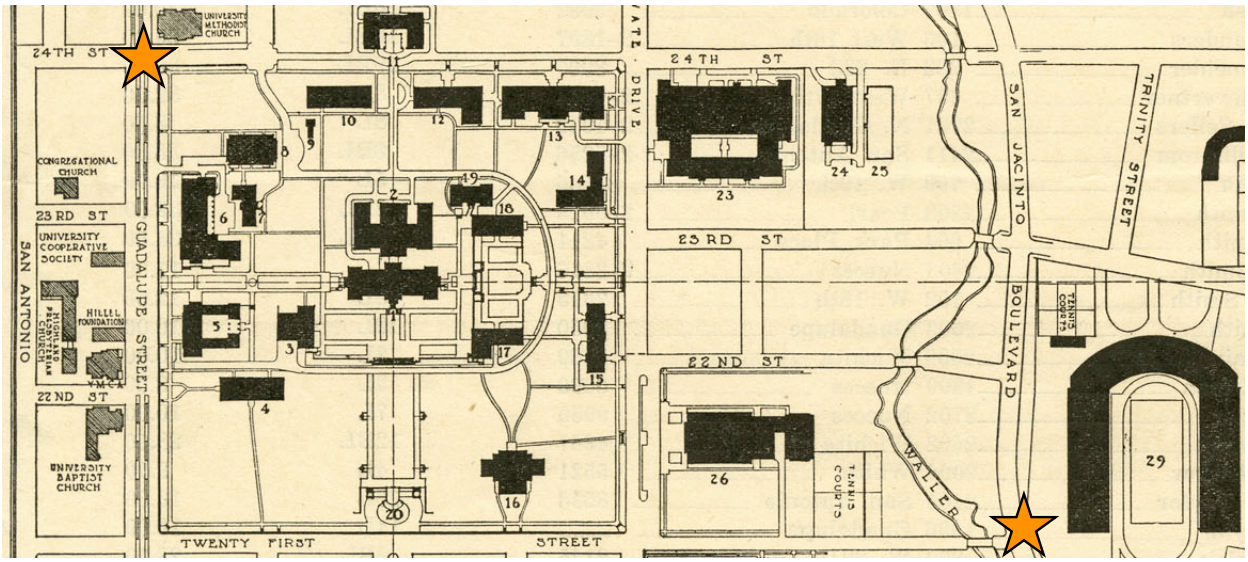
Pythagoras' famous theorem says that $c^2 = a^2 + b^2$, where c is the hypotenuse of a right triangle, and a and b are the legs. Notice that Pythagoras *doesn't* have anything interesting to say about the actual numbers: $c \neq a + b$. It's the squares of the numbers that matter.

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! Below, 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.

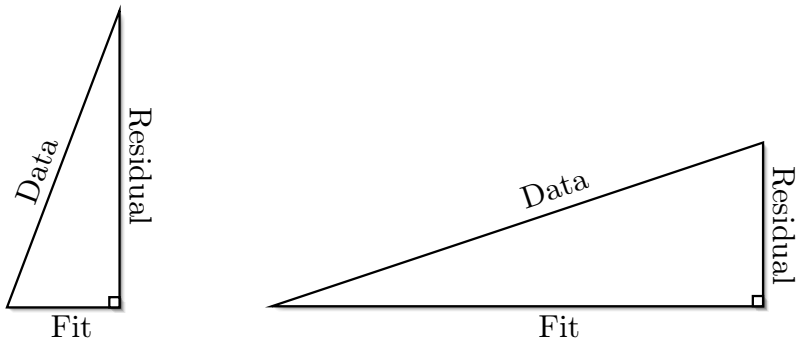
A student who made this journey on November 26th of the year this map was drawn (1930) would have witnessed a 26–0 victory over Texas A&M. Just by the by.

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, metaphorically speaking, perpendicular to one another: one direction that can be predicted by the model, and one direc-



tion that can't. The total variation is then like the hypotenuse of the right triangle so formed:



I personally think of the "predictable" direction as east–west, because the predictor in a scatter plot usually gets plotted on the horizontal axis. But don't take that aspect of the metaphor too literally.

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

One final note on sums of squares: I've been talking total rubbish about one crucial point. It turns out that this story about the fitted values and residuals pointing in perpendicular directions isn't a metaphor at all. 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. 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:

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

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

Always remember that 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 $\text{TV} = 102.1$ squared hours in total variation, and were left with $\text{UV} = 65.7$ squared hours in unpredictable variation after fitting the group-wise model based on the predation index. Therefore $R^2 = \text{PV}/\text{TV} \approx 0.36$, meaning that the model predicts 36% of the variation in dreaming hours.

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.

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.

$\widehat{\beta}_1$, 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:

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

or predictable variation divided by total variation.

These are different quantities: the slope β_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 β_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, β_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 β_1 is dollars per year of education.
- If X is seconds and Y is meters, then β_1 is meters per second.
- If X is bits and Y is druthers, then β_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.[3] 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, $\widehat{\beta}_1$, and R^2—can be related to each other in a single line of equations:

$$\{r(y,x)\}^2 = r^2 = R^2 = \{r(y,\hat{y})\}^2 = \{r(y, \widehat{\beta}_0 + x\widehat{\beta}_1)\}^2 .$$

If you understand all those links, you're doing brilliantly!

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 1.11, shown again

[3] To see this algebraically, note that

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

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

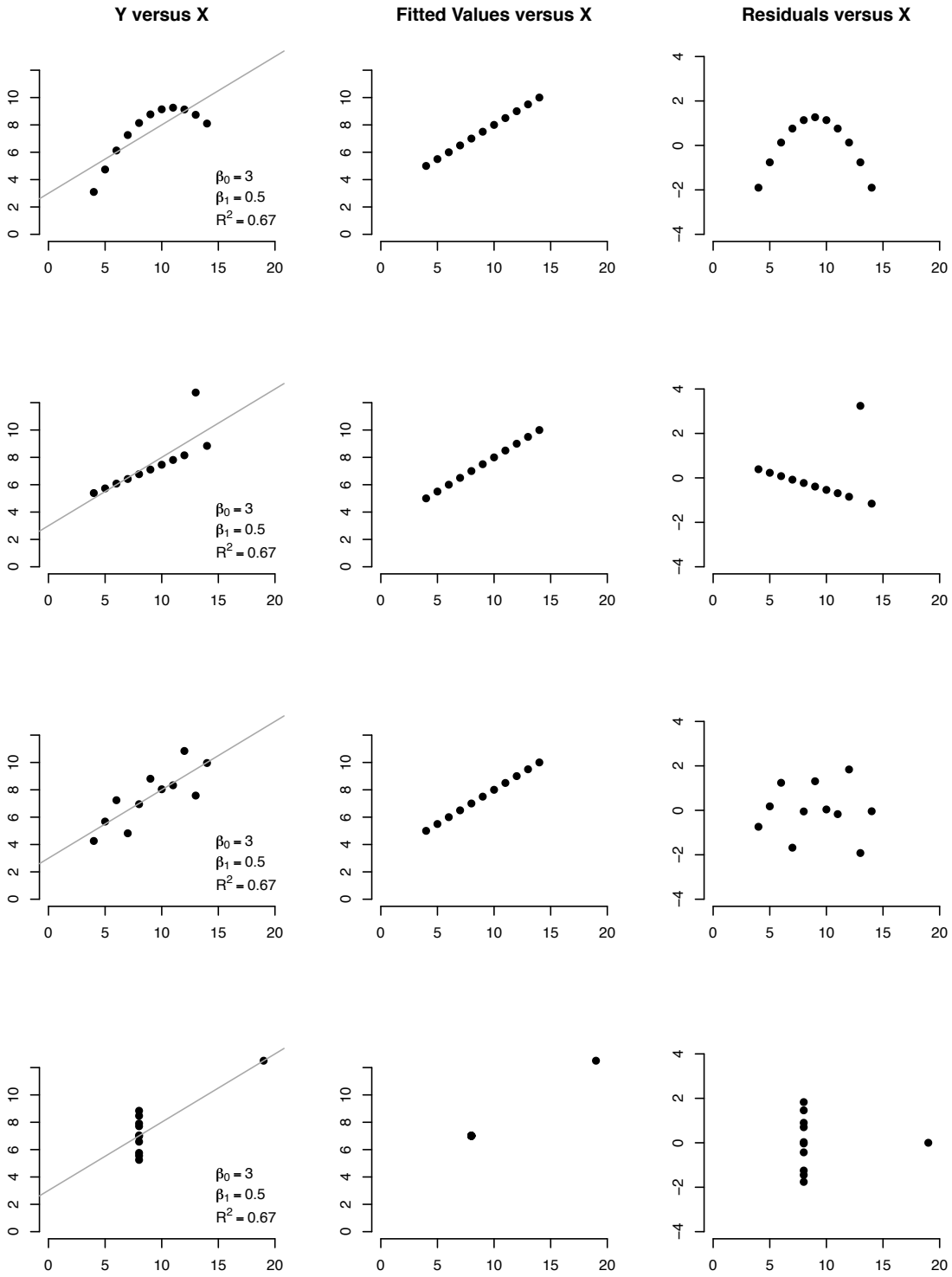


Figure 3.7: These four data sets have the same least-squares line.

on the following page. 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: remarkably, the four data sets all have the same least-squares line and the same value of R^2, too! In Figure 3.7 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.[4]

[4] *Data Analysis for Politics and Policy*, p. 72.

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 $\widehat{\beta}_1 = -26,920$ 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. Clearly as the nation moved from George to Dan to Al, innovation blossomed.

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.

Nothing has been "explained" here at all, the high R^2 notwithstanding: garbage in, garbage out. 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. If you want to disabuse yourself of this notion, try reversing every regression you run, and fit X versus Y instead. If the two variables are related, you'll find that you also do a pretty good job at predicting the "predictor" using the "response".