# 2
# *Fitting equations to data*

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

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

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

## Fitting straight lines

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

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

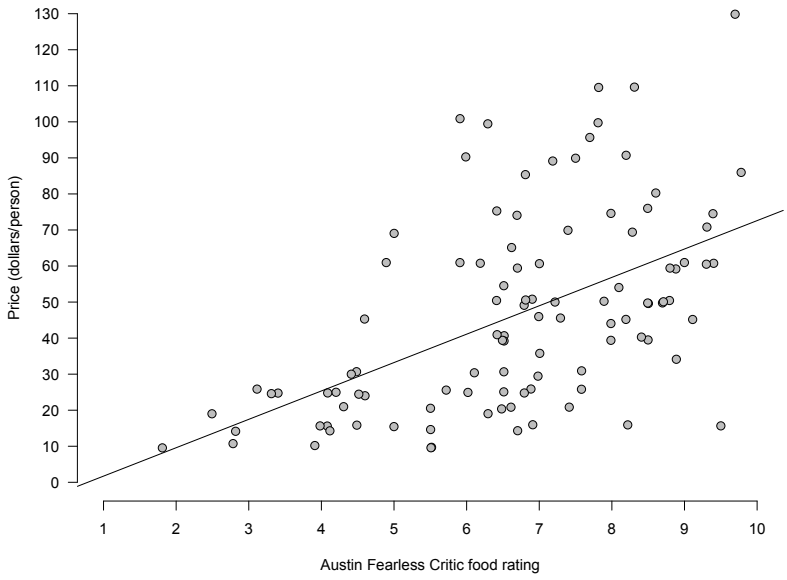



Figure 2.1: Price versus reviewer food rating for a sample of 104 restaurants near downtown Austin, Texas. The data are from a larger sample of 317 restaurants from across greater Austin, but downtown-area restaurants were chosen to hold location relatively constant. Data from Austin Fearless Critic,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

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

$$y_i = \beta_0 + \beta_1 x_i + e_i \,. \tag{2.1}$$

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

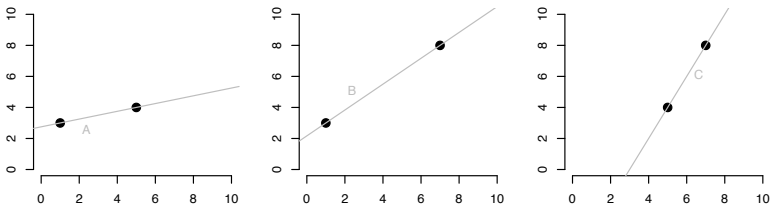
*For every two points. . . .*

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

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

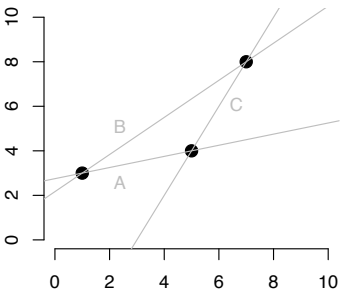
you've encountered it, seems beautifully simple, almost to the point of being obvious. But it's worth pausing to consider its historical origins, for it was far from obvious to a large number of very bright 18th-century scientists.

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




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

But things are more complicated if we observe three points.




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

Two unknowns, three equations. 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.

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 the residuals.

But now we've created a different predicament. Before we added the $e_i$'s to give us some wiggle room, there was no solution to our system of linear equations. Now we have three equations

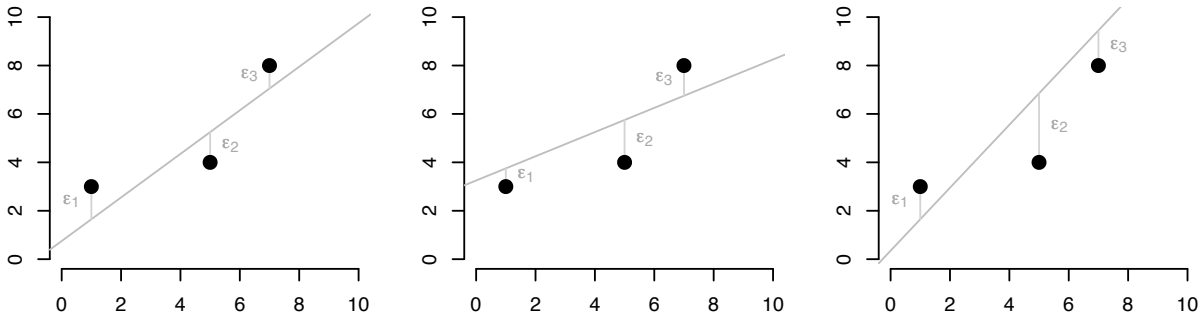



Figure 2.2: Three possible straight-line fits, each involving an attempt to distribute the "errors" among the observations.

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

Believe it or not, scientists of the 1700's struggled mightily with this question. Many of the central scientific problems of this era concerned the combination of astronomical or geophysical observations. Astronomy in particular was a hugely important subject for the major naval powers of the day, since their ships all navigated 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 fitting an equation to observations of the moon's orbit.

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

In fact, some thinkers of this period disputed that it was even a good idea to combine observations at all. Their reasoning was, roughly, that the "bad" observations in your sample would corrupt the "good" ones, resulting in an inferior final answer. To borrow

the phrase of Stephen Stigler, an historian of statistics, the "deceptively simple concept" that combining observations would improve accuracy, not compromise it, was very slow to catch on during the eighteenth century.[1]

[1] *The History of Statistics*, p. 15.

*The method of least squares*

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

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

> In most investigations where the object is to deduce the most accurate possible results from observational measurements, we are led to a system of equations of the form
>
> $$E = a + bx + cy + fz + \&c.,$$
>
> in which $a$, $b$, $c$, $f$, &$c$. are known coefficients, varying from one equation to the other, and $x$, $y$, $z$, &$c$. are unknown quantities, to be determined by the condition that each value of $E$ is reduced either to zero, or to a very small quantity. . . .
>
> Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of the squares of the errors a minimum. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.[2]

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

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

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

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}\,, \tag{2.3}$$

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 solving a simple mathematical problem involving differential calculus and matrix algebra—something that scientists of the nineteenth century could do easily, via pen and paper.

## Goals of regression analysis

WITH modern computers, the estimation of linear regression models by least squares is now entirely automatic. It's so ordinary, in fact, that the method is abbreviated as OLS: ordinary least squares.

Regression modeling is interesting, therefore, because of what we can do with it. Here are four kinds of stories one can tell with a regression model. Each is useful for a different purpose.

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

One way to interpret a regression model is as is a function $\hat{y} = f(x)$ that maps inputs to expected 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 $\hat{y}_i$:

$$\hat{y}_i = \widehat{\beta}_0 + x_i \widehat{\beta}_1\,. \tag{2.4}$$

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

### 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.*

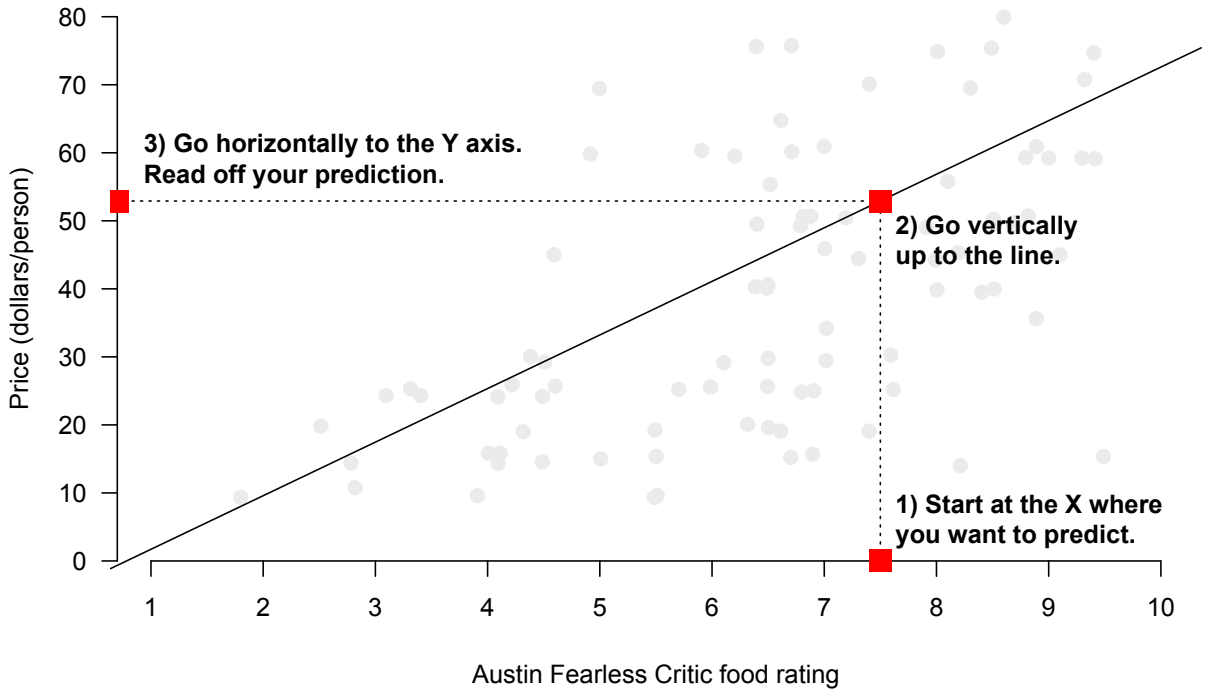



Figure 2.3: Using a regression model for plug-in prediction of the price of a meal, assuming a food rating of 7.5.

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. (See Figure 2.3). This, incidentally, is where the name *regression* comes from: we expect that future $y$'s will "regress to the mean" specified by the linear predictor.

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

The linear predictor tells you how $Y$ changes, on average, as a function of $X$. In particular, the slope $\beta_1$ tells you how the 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 2.1, the slope is $\beta_1 = 7.9$. On average,

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.

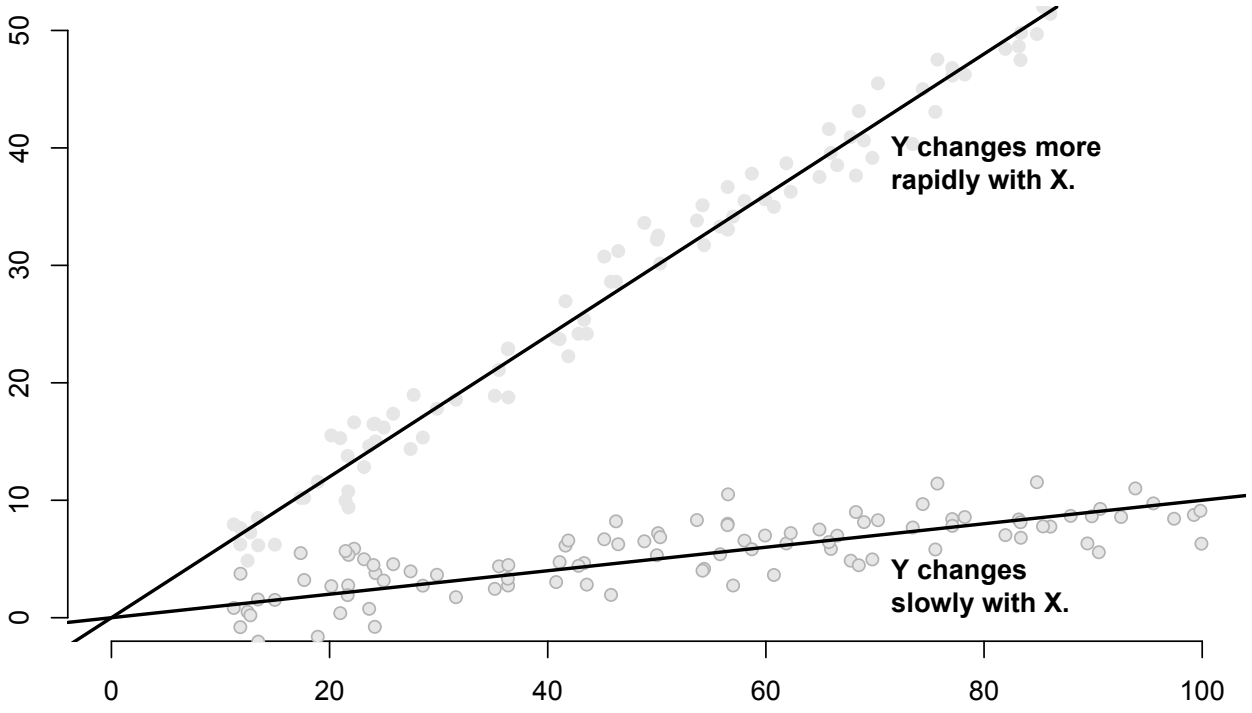



Figure 2.4: The slope of a regression model summarizes how fast the $Y$ variable changes, as a function of $X$.

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

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

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

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

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

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

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

or equivalently,

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

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

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

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

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

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

This is easily seen in our example by plotting the residual price ($e_i$) against food rating ($x_i$), side by side with the original data, as in Figure 2.5. In the right panel, there is no evident correlation between food rating and the residuals. This should always be true: a good regression model should take the $X$-ness out of $Y$, so that the residuals look independent of the predictor. If they don't, then the model hasn't done its job. Always plot your data.

You've just seen your first example of statistical adjustment. Notice the red dot sitting in the lower right of Figure 2.5, with a

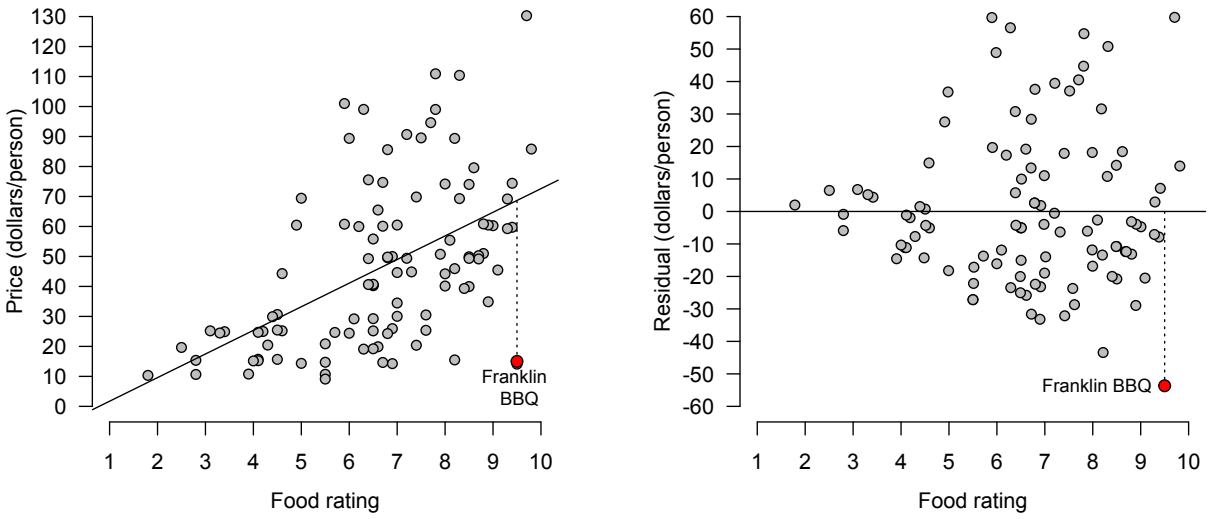



Figure 2.5: Left: the original data on price versus food rating. Right: the residuals from the least squares fit on the left. The residual for Franklin BBQ is the length of the dotted vertical line: $e_i = -\$53.85$.

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. That's the largest, in absolute value, of all the negative residuals.

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

> 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.[3]

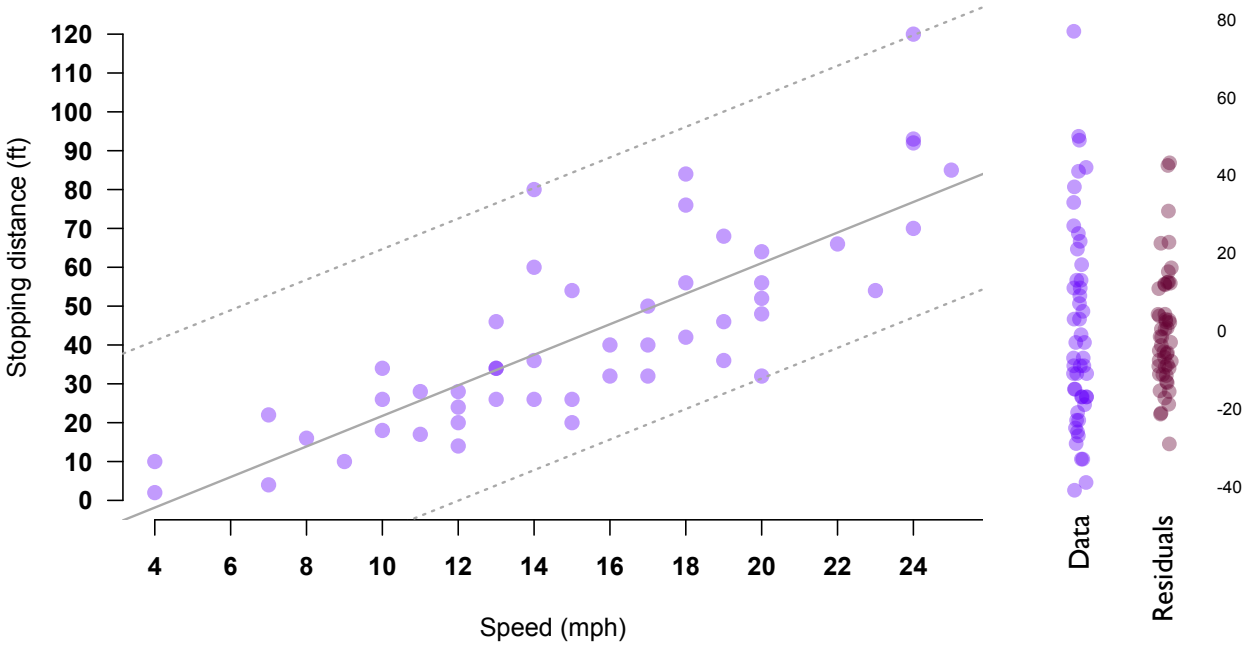
And undoubtedly the most delicious residual in the city.

[3] "A Day in the Life of a BBQ Genius." Andrew Knowlton, *Bon Appétit*, July 2011. These days (2016), queueing up at 10:30 would have you last in line.

*Story 4: A regression model 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 mathematically detailed answer to this question turns out to be surprisingly elegant, and cuts to the heart of an earlier question left unanswered: why measure variation using sums of squares? We'll consider this at length in the next chapter.

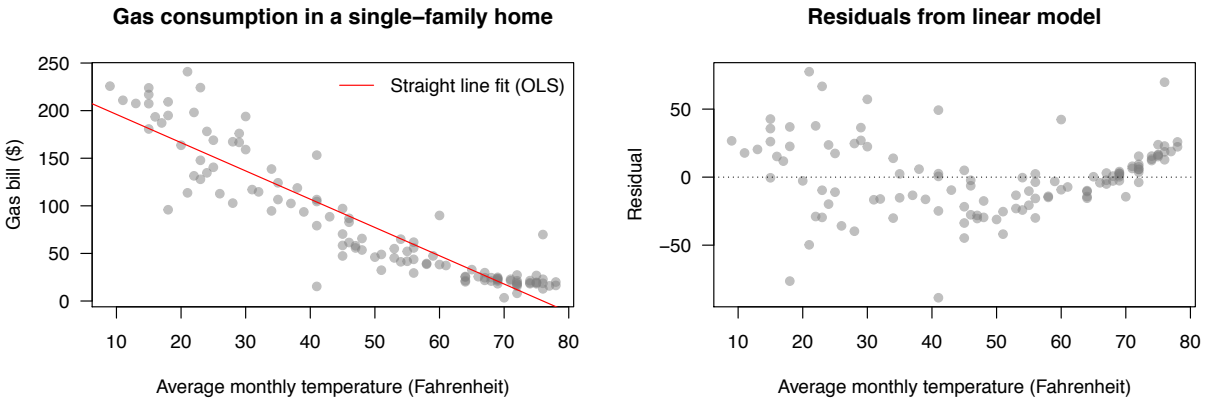



Figure 2.6: Left: a scatterplot of monthly gas consumption (measured in dollars) versus average monthly temperature at a single-family home in Minnesota, together with a linear regression model fit by ordinary least squares. Right: a plot of the residuals from the linear model versus temperature, showing the deficits of the straight line fit. Data source: Daniel T. Kaplan, *Statistical Modeling: A Fresh Approach*, 2009.

## Beyond straight lines

Up to this point, we've talked about fitting straight lines using the principle of least squares. For many data sets, however, a linear regression model doesn't provide an adequate description of what's going on. Consider, for example, the data on monthly gas consumption for a single-family home in Minnesota shown in the left panel Figure 2.6. As the temperature rises, the residents of the house use less gas for heating. But this trend is not well described using a straight line fit by least squares, in this case

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

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

The inadequacy of the linear model is revealed by the residual plot in the right panel. Here, the residuals $e_i$ from the linear fit in the left panel are plotted versus temperature. Remember, these residuals *should* be unrelated with the predictor if our regression model has done its job right. But here, this is clearly false:

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

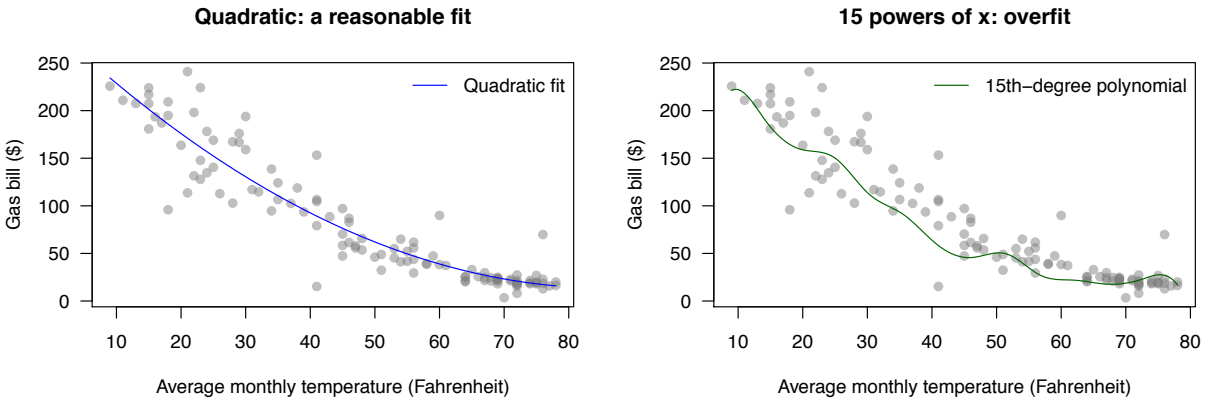



Figure 2.7: Left: the fit of a quadratic model (2nd-degree polynomial) estimated by least squares. Right: the fit of a 15th-degree polynomial. The model on the left provides an intuitively reasonable description of the underlying relationship, while the model on the right is a clear example of over-fitting.

most all positive, suggesting that the regression model made predictions that were systematically too low yet again.

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

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

(1) polynomial models (like quadratic or cubic equations); and

(2) models involving a simple mathematical transformation of the predictor, the response, or both.

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

*Polynomial regression models*

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

But we can fit other polynomials by least squares, too. For example, the left panel of Figure 2.7 shows the least-squares fit of a quadratic equation (another name for a second-degree polynomial) to the gas-consumption data set:

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

The quadratic model fits noticeably better than the straight line. In

particular, it captures the leveling-off in gas consumption at high temperatures that was missed by the linear model.

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

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

for an arbitrary choice of $K$.

To want to fit the data as well as possible is an understandable impulse. But for most data sets, if we venture beyond $K = 2$ (quadratic) or $K = 3$ (cubic), we rapidly get into dangerous over-fitting territory. *Over-fitting* occurs when a regression model starts to memorize the random noise in the data set, rather than describe the underlying relationship between predictor and response. We see a clear example of over-fitting in the right panel of 2.7, which shows the result of using least-squares to estimate a 15th-degree polynomial for gas bill versus temperature. The fitted curve exaggerates minor dips and rises in the data, leading to an absurdly complex function. There's no reason for us to think that gas consumption responds to temperature in the way implied by the green curve on the right of Figure 2.7. For example, why would consumption rise systematically between 45 and 50 degrees, but then drop again between 50 and 60 degrees?

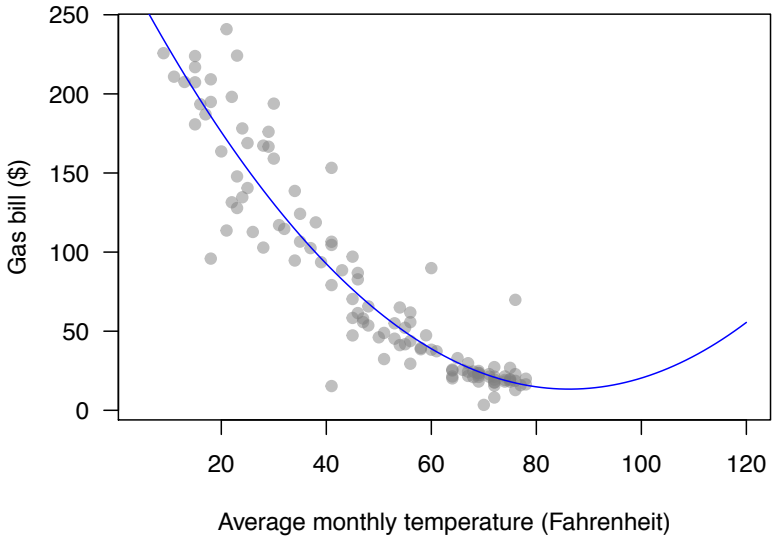
In regression modeling, we want to build models that are only as complex as they must be in order to describe the underlying relationship between predictor and response. But how do we know reliably when we've crossed this line from "fitting well" into over-fitting?

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

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

With apologies to Potter Stewart, when it comes to overfitting, you'll often know it when you see it. Always plot your data.

*Extrapolation.* Although the quadratic model fits the data well, its predictive abilities will rapidly deterioriate as we move above 80 degees (i.e. as we use the model to extrapolate further and further from past experience). That's because the fitted curve is a parabola: it turn upwards around 85 degrees, counterintuitively suggesting that gas bills would eventually rise with temperature:




This behavior is magnified dramatically with higher-order polynomials, which behave in unpredictable ways beyond the endpoints of your data. For this reason, never extrapolate using a polynomial regression model, unless you really know what you're doing.

*Exponential growth and decay*

Beginning in March 2014, West Africa experienced the largest outbreak of the Ebola virus in history. Guinea, Liberia, Niger, Sierra Leone, and Senegal were all hit hard by the epidemic. Figure 2.8 shows the number of laboratory-confirmed cases of Ebola in these five countries over time, beginning on March 25.

If we wanted to fit a model to describe how the number of Ebola infections grew over time, we might be tempted to fit a polynomial function (since a linear model clearly won't work well here). However, basic biology tells us that the transmission rate of a disease through a population is reasonably well described by an

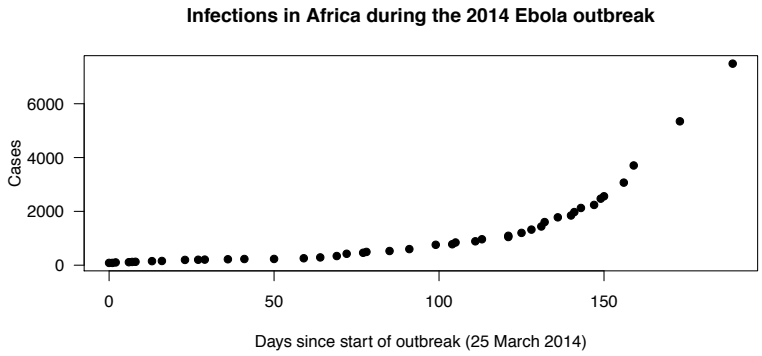



Figure 2.8: Cases of Ebola over time in West Africa, 2014. Compiled from CDC reports by Francis Smart, as described here.

exponential growth model: 1 infection leads to 2, which lead to 4, which lead to 8, to 16, and so on. The equation for an exponential-growth model is

$$y = \alpha \cdot e^{\beta t}\,, \tag{2.6}$$

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

It turns out that we can use least squares to fit an exponential growth model of this form, using a new trick: *take the logarithm of the response variable* and fit a linear model to this new transformed variable. We can see why this works if we take the logarithm of $y$ in the equation for exponential growth (labeled 2.6, above). To preserve equality, if we take the log of the left-hand side, we also have to take the log of the right-hand side:

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

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

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

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

Figure 2.9 shows the result of following these two steps for the Ebola data. The left panel shows the straight-line fit on the log

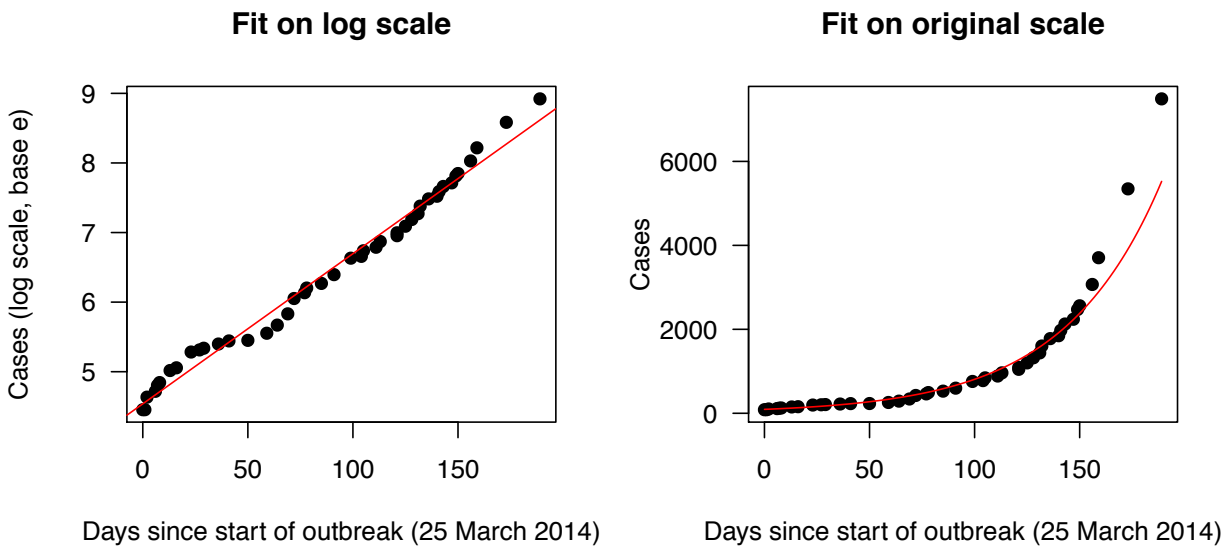



Figure 2.9: An exponential-growth model fit to the Ebola data by ordinary least squares, where the $y$ variable is shown on the log scale (left) and on the original scale (right).

scale:

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

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

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

The leading constant is calculated from the intercept on the log scale: $93.5 \approx e^{4.54}$. From Figure 2.9, we can see that the exponential-growth model fits adequately, although imperfectly: the rate of growth seems to be accelerating at the right of the picture, and the upward trajectory is visibly nonlinear on the log scale. (Remember: all models are wrong, but some models are useful.)

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

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

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

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

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

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

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

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

days during the spring and early summer of 2014.

In an exponential decay model (where $\beta_1 < 0$), a similar calculation would tell you the half life, not the doubling time.[4]

[4] Instead, solve the equation

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

for the different $t_2 - t_1$.

*Double log transformations*

In some cases, it may be best to take the log of both the predictor and the response, and to work on this doubly transformed scale. For example, in the upper left panel of Figure 2.10, we see a scatter plot of brain weight (in grams) versus body weight (in kilos) for 62 different mammalian species, ranging from the lesser short-tailed shrew (weight: 10 grams) to the African elephant (weight: 6000+ kilos). You can see that most species are scrunched up in a small box at the lower left of the plot. This happens because the observations span many orders of magnitude, and most are small in absolute terms.

But if we take the log of both body weight and brain weight, as in the top-right panel of Figure 2.10, the picture changes considerably. Notice that, in each of the top two panels, the red box encloses the same set of points. On the right, however, the double log transformation has stretched the box out in both dimensions, allowing us to see the large number of data points that, on the left, were all trying to occupy the same space. Meanwhile, the two points outside the box (the African and Asian elephants) have been forced to cede some real estate to the rest of Mammalia.

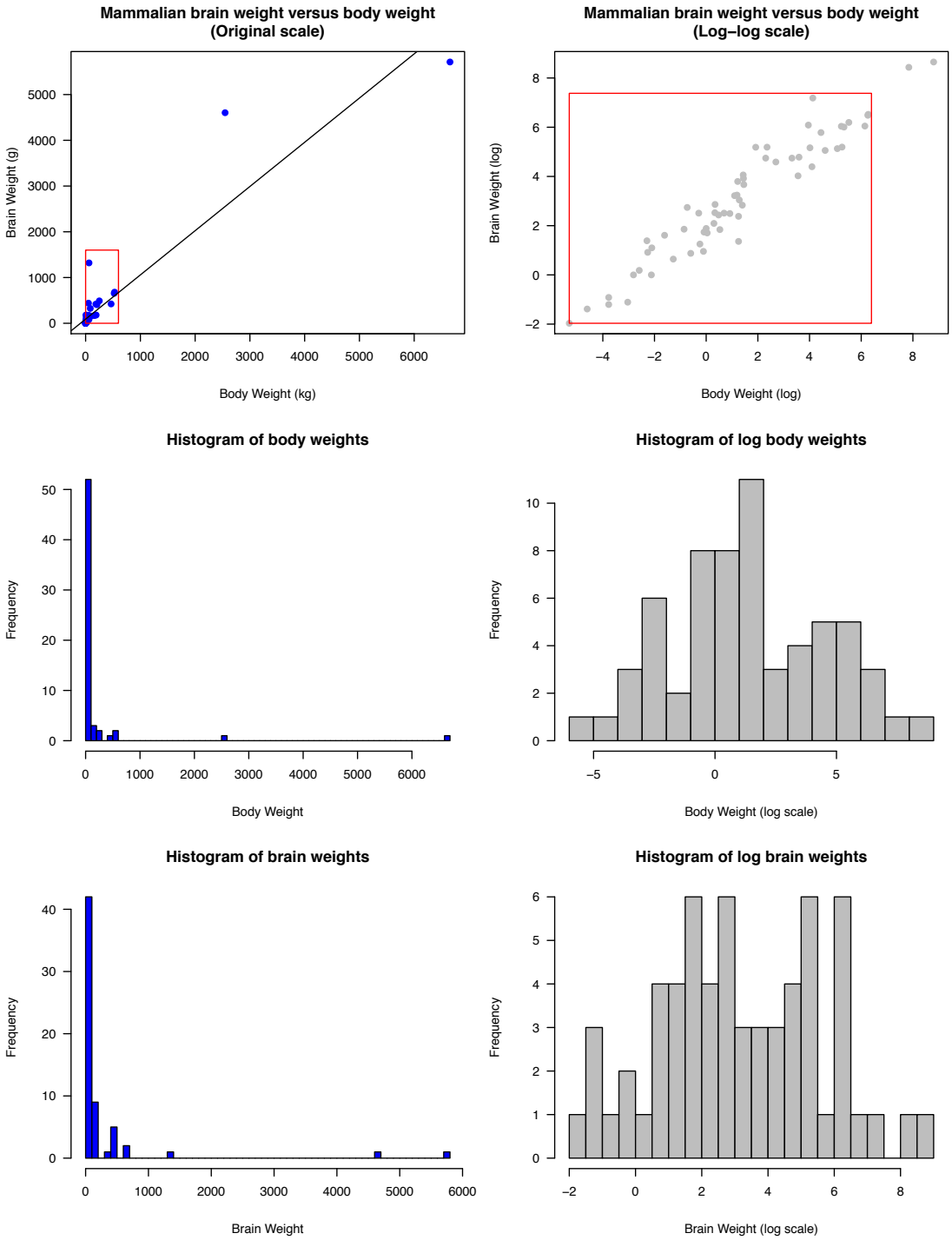



Figure 2.10: Brain weight versus body weight for 62 mammalian species, both on the original scale and the log scale. Notice how the log transformation "unsquishes" the points.

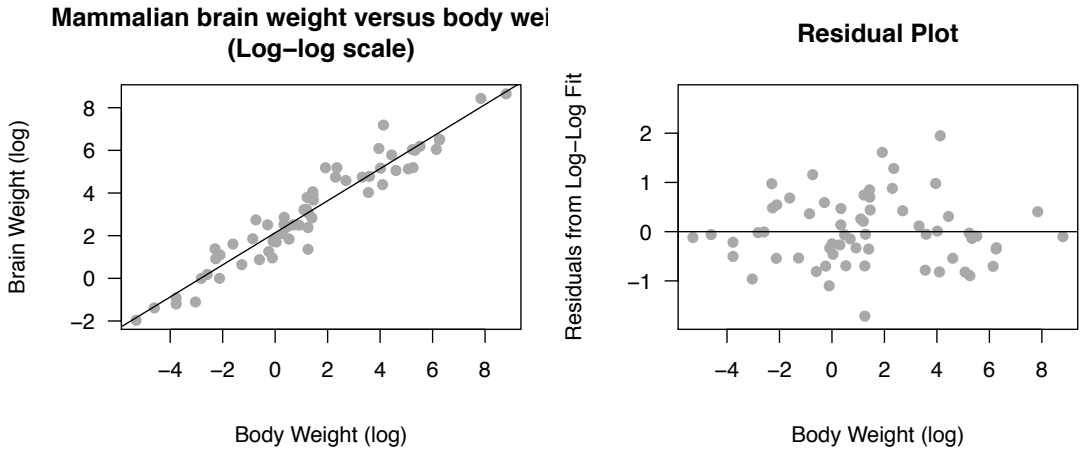



Figure 2.11: A straight-line fit to the mammalian brain weight data after a double log transformation.

This emphasizes the taking the log is an "unsquishing" operator. To see this explicitly, look at the histograms in the second and third row of panels in Figure 2.10. Whenever the histogram of a variable looks highly skewed right, as on the left, a log transformation is worth considering. It will yield a much more nicely spread-out distribution of points, as on the right.

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

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

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

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

Therefore, if $y$ and $x$ follow a power law, then $\log y$ and $\log x$ follow a linear relationship with intercept $\log \alpha$ and slope $\beta_1$. This implies that we can fit the parameters of the power law by applying the double log transformation and using ordinary least

squares. For our mammalian brain weight data, applying this recipe yields the fitted equation

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

or expressed as a power law on the original scale,

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

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

$$\log y_i = \log \alpha + \beta_1 \log x_i + e_i\,, \tag{2.7}$$

where $e_i$ is the amount by which the fitted line misses $\log y_i$. We suppressed the residuals before the lighten the algebra.

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

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

Therefore the residuals in a power law describe the percentage error on the original scale. Let's work through the calculations for two examples:

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

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

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

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \beta_1 \alpha x^{\beta_1 - 1}.$$

We can rewrite this as

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

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

$$\beta_1 = \frac{\mathrm{d}y/y}{\mathrm{d}x/x}. \tag{2.8}$$

Since the $\mathrm{d}y$ in the derivative means "change in $y$", the numerator is the rate at which the $y$ variable changes, as a fraction of its value. Similarly, since $\mathrm{d}x$ means "change in $x$", the denominator is the rate at which the $x$ variable changes, as a fraction of its value.

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

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

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

where $Q$ is the quantity demanded by consumers, $P$ is the price, and $\beta_1 < 0$. Economists would call $\beta_1$ the price elasticity of demand.[5]

[5] They actually define elasticity as the ratio in Equation 2.8, but as we've seen, this is mathematically equivalent to the regression coefficient you get when you fit the $x$–$y$ relationship using a power law.