6
Fitting straight lines

You can go pretty far in data science using relatively simple visual and numerical summaries of data sets—tables, scatter plots, bar plots, line graphs, boxplots, histograms, and so on. But 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 three that we'll explore in detail:

- to make a prediction;
- to summarize the trend in a data set;
- to make comparisons that adjust statistically for some systematic effect.

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

This is our first of many data sets where the response (price, Y) and predictor (food score, X) can be described by a linear regression model. We write the model in two parts as "$Y = \beta_0 + \beta_1 X + \text{error}$." The first part, the function $\beta_0 + \beta_1 X$, is called the *fitted value*: it's what our equation "expects" Y to be, given X.

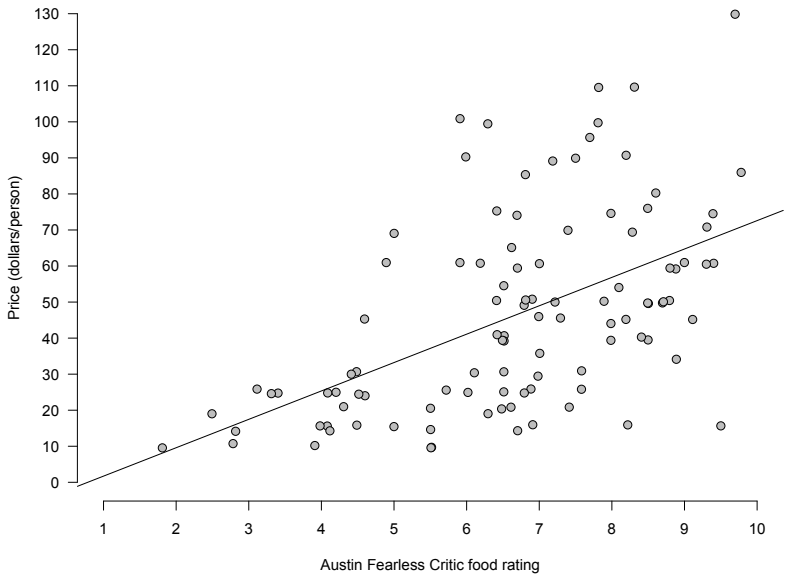


Figure 6.1: Price versus reviewer food rating for a sample of 104 restaurants near downtown Austin, Texas. The data are from a larger sample of 317 restaurants from across greater Austin, but downtown-area restaurants were chosen to hold location relatively constant. Data from Austin Fearless Critic, 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 $y = -6.2 + 7.9x$.

The second part, the error, 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:

$$y_i = \beta_0 + \beta_1 x_i + e_i \,. \tag{6.1}$$

Here the subscript i is just an index to denote which data point we're talking about: $i = 1$ for the first row of our data frame, $i = 9$ for the 9th, and so on.

An equation like (6.1) is our first example of a regression model. The *intercept* β_0 and the *slope* β_1 are called the *parameters* of the regression model. They provide a mathematical description of how price changes as a function of food score. The little e_i is called the error or the *residual* for the ith case—residual, because it's how much the line misses the ith 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.

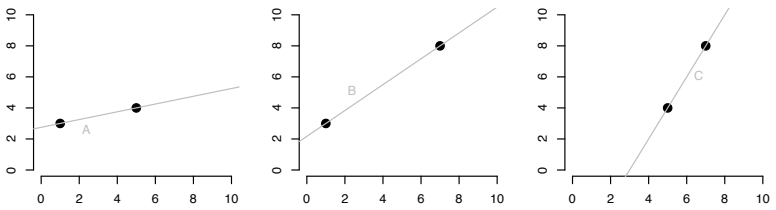
For every two points. . . .

A natural question is: how do we fit the parameters β_0 and β_1 to the observed data? Historically, the standard approach, still in widespread use today, is to use the method of least squares. This involves choosing β_0 and β_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 6.1.

The method of least squares is one of those ideas that, once you've encountered it, seems beautifully simple, almost to the point of being obvious. But it's worth pausing to consider its historical origins, for it was far from obvious to a large number of very bright 18th-century scientists.

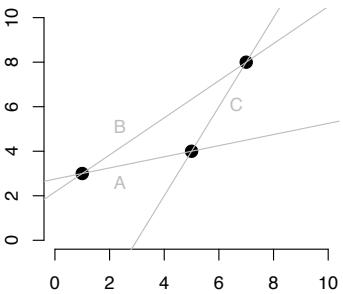
To see the issue, consider the following three simple data sets. Each has only two observations, and therefore little controversy about the best-fitting linear trend.



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

But things are more complicated if we observe three points.

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



Two unknowns, three equations. There is no solution for the parameters β_0 and β_1 that satisfies all three equations—and therefore no perfectly fitting linear trend exists. Seen graphically, at right, it is clear that no line can pass through all three points.

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

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

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

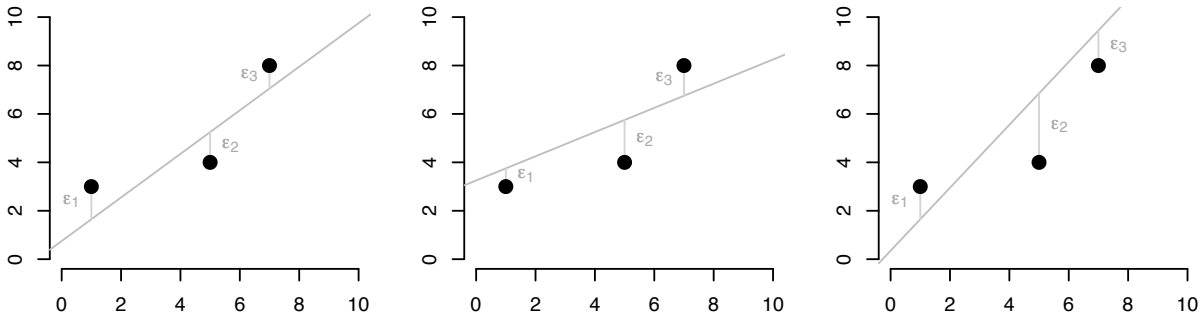


Figure 6.2: Three possible straight-line fits, each involving an attempt to distribute the "errors" among the observations.

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

But now we've created a different predicament. Before we added the e_i's to give us some wiggle room, there was no solution to our system of linear equations. Now we have three equations and five unknowns: an intercept, a slope, and three residuals. This system has infinitely many solutions. How are we to choose, for example, among the three lines in Figure 6.2? 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. Very rapidly, least squares became the dominant method for fitting equations 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 $(\widehat{\beta}_0, \widehat{\beta}_1)$:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}\,, \tag{6.3}$$

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

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 some calculus and matrix algebra. This is something that scientists of the nineteenth century could do via pen and paper—and that, happily, modern computers take care of for us nowadays.

Goals of regression analysis

The estimation of linear regression models by least squares is now entirely automatic using standard software for all but the very largest of data sets.[3] It's so ordinary, in fact, that the method is often abbreviated as OLS: ordinary least squares.

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

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

Least squares then and now: an historical aside

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

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

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

A cheap laptop computer bought today takes a second or less to solve the same problem.

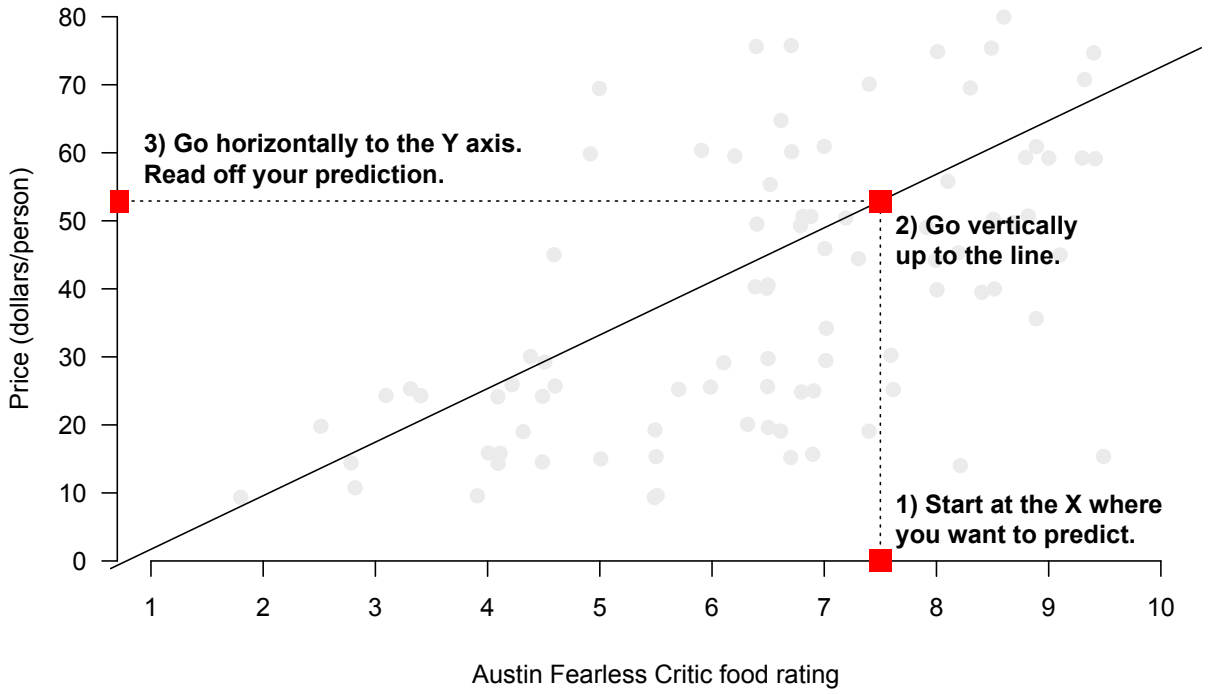


Figure 6.3: Using a regression model for plug-in prediction of the price of a meal, assuming a food rating of 7.5.

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

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

$$\hat{y}_i = \widehat{\beta}_0 + x_i \widehat{\beta}_1 \,. \tag{6.4}$$

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

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

For example, if we know that a new restaurant earned a food rating of 7.5, our best guess for the cost of the meal—knowing nothing else about the restaurant—would be to use the linear predictor: $\hat{y}^\star = -6.2 + 7.9 \cdot 7.5$, or \$53.05 per person. (See Figure 6.3). This, incidentally, is where the name *regression* comes from:

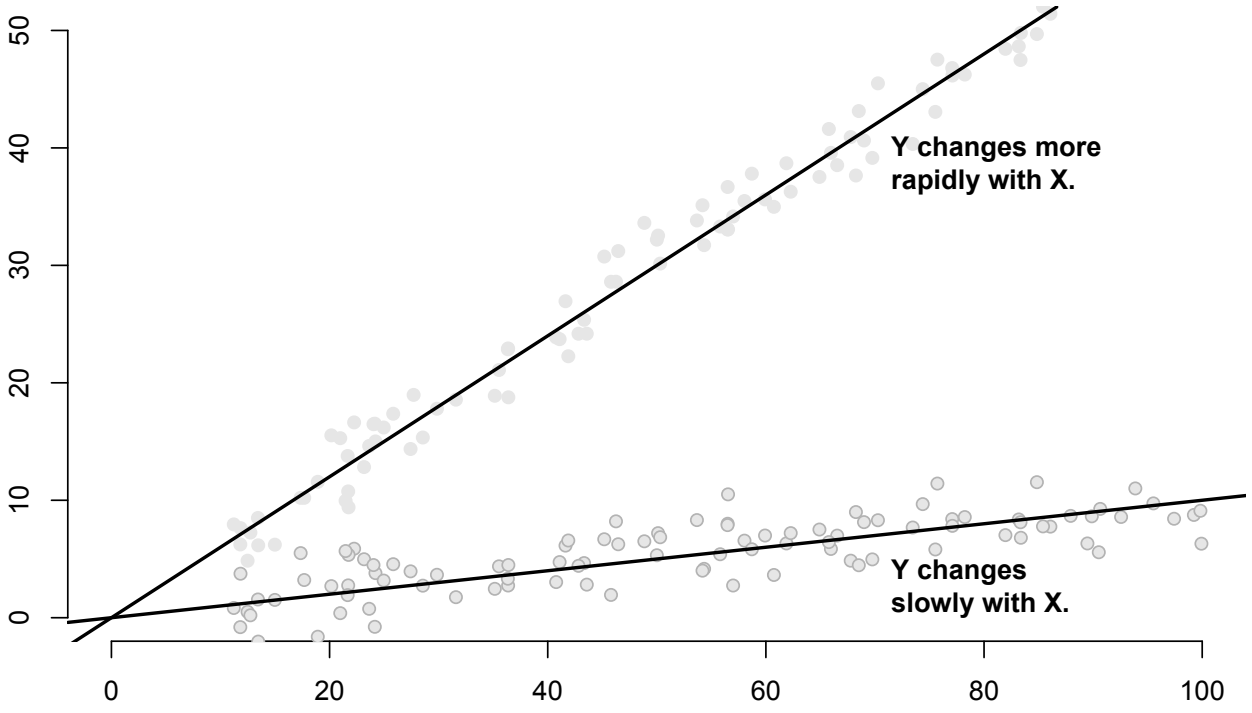


Figure 6.4: The slope of a regression model summarizes how fast the Y variable changes, as a function of X.

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 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 6.1, 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.

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 β_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 6.1, 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.

"Taking the X-ness out of Y" a bit of a weird phrase. But we like it for a good reason; bear with us.

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

$$\text{Observed } y \text{ value} = (\text{Fitted value}) + (\text{Residual}), \tag{6.5}$$

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—more like an error in our model, rather than an error in our measurement.

In Figure 6.1, 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

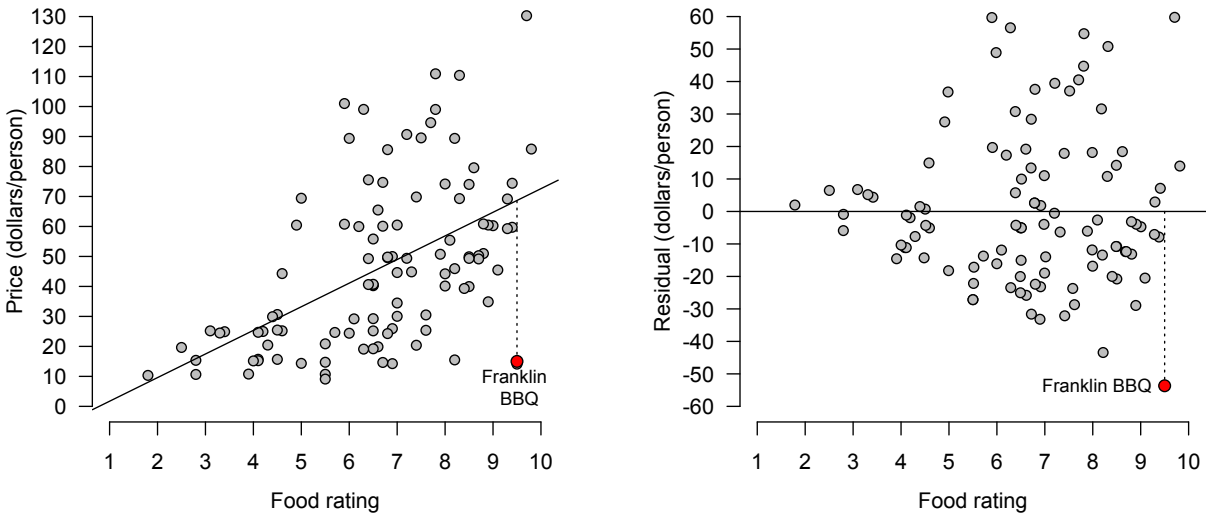


Figure 6.5: Left: the original data on price versus food rating. Right: the residuals from the least squares fit on the left. The residual for Franklin BBQ is the length of the dotted vertical line: $e_i = -\$53.85$.

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 6.5. 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 X. If they don't, then the model hasn't done its job and we need a different model.

You've just seen your first example of *statistical adjustment*. Notice the red dot sitting in the lower right of Figure 6.5, with a low price and a high food rating? This isn't the least expensive restaurant near downtown Austin in an absolute sense. But it is the least expensive *after we adjust for food rating*. To do this, we simply subtract off the fitted value from the observed value of y, leaving the residual—which, you'll recall, captures what's over in the response (price) after the predictor (food score) has been taken into account. The restaurant in question has a food rating of 9.5, good for *Fearless Critic*'s third best score in the entire city. For such delicious

food, you would expect to pay $\hat{y}^\star = -6.2 + 7.9 \cdot 9.5$, 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, and widely regarded as the most delicious residual in the city:

> Go to Austin and queue up at Franklin Barbecue by 10:30 a.m. When you get to the counter, Aaron Franklin will be waiting, knife in hand, ready to slice up his brisket. (Order the fatty end.) Grab a table, a few beers, and lots of napkins and dig in. Take a bite, and don't tell me you're not convinced you've reached the BBQ promised land.

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

R^2: quantifying the information content in the model.

The idea behind the Flu Prediction Project, run jointly by IBM Watson and the University of Osnabrück in Germany, is simple.[4] Researchers combine social-media and internet-search data, together with official data provided by government authorities, like the Centers for Disease Control (CDC) in the United States, to yield accurate real-time predictions about the spread of seasonal influenza. This kind of forecasting model allows public-health authorities to allocate resources (like antivirals and flu vaccines) using the most up-to-date information possible. After all, the official government data can usually tell you what flu activity was like two weeks ago. Social-media and internet-search data, if used correctly, have the potential to tell what you it's like right now.

[4] http://www.flu-prediction.com

To give you a sense of how strong the predictive signal from internet-search data can be, examine Figure 6.6, focusing first on the scatter plot in the left panel. Here each dot corresponds to a day. On the x-axis is a measure of Google search activity for the term "how long does flu last," where higher numbers mean that more people are searching for that term on that day.[5] On the y axis, we see a measure of actual flu activity on that day, constructed from data provided by the CDC.

[5] Specifically, it's a z score: how many standard deviations about the mean was the search frequency on that day for that particular term.

The search activity on a given day strongly predicts actual flu transmission, which makes sense: one of the first things that many

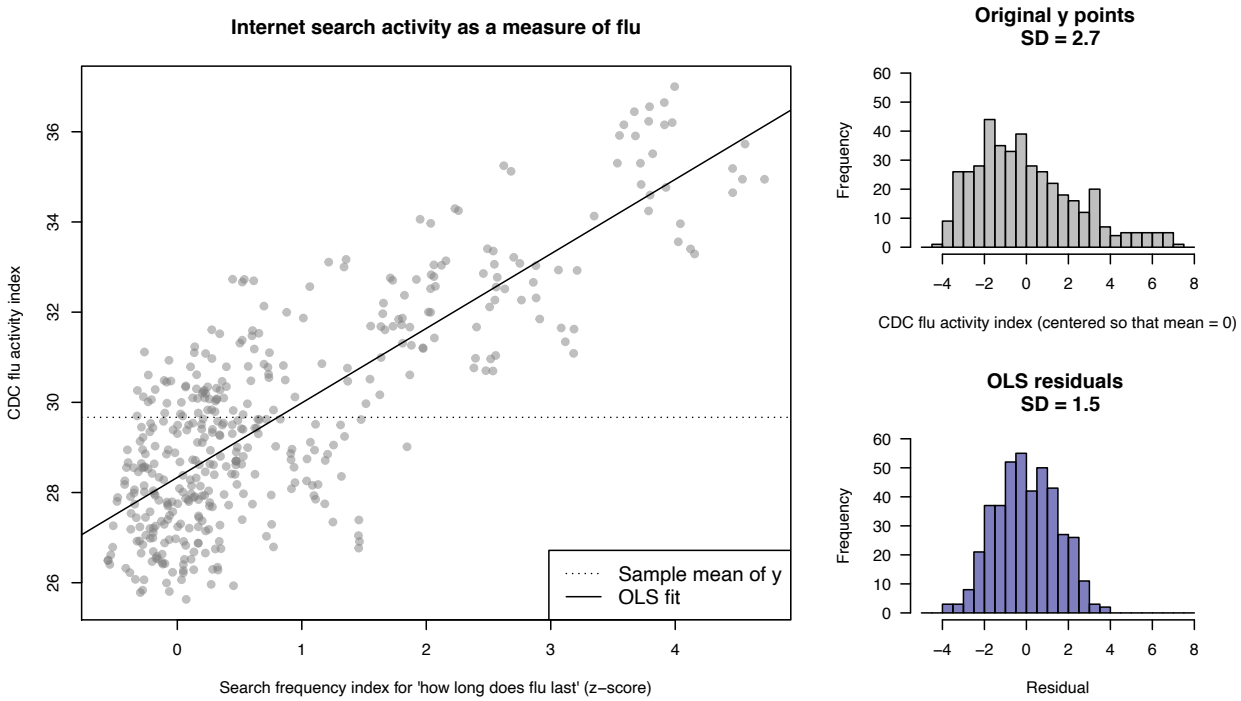


Figure 6.6: A scatter plot of the CDC's measure of flu activity versus Google search activity for the phrase "how long does flu last" (z score of search frequency). To the right of the scatter plot, we see two dot plots, both on the same scale: (1) the original deviations from the sample mean, $y_i - \bar{y}$; and (2) the residuals from the regresson equation, $y_i - \hat{y}_i$.

people do when they fall ill is to commiserate with a search engine about the depth and duration of their suffering. But just how much information about flu does the search activity for this single term—"how long does flu last"–convey?

Our regression model provides a way of measuring the "amount" of this information, because it allows us to compare our predictions of flu activity both with and without the x variable.

- Without knowing the predictor variable, our best guess for the outcome is just the sample mean, $\bar{y}$, and the prediction error for each case is $y_i - \bar{y}$. You can think of the sample mean as our "baseline" or "straw man" prediction; it is obviously a pretty simple baseline.

- With the predictor variable, our best guess is given by the regression model, $\hat{y}_i = \beta_0 + \beta_1 x_i$, and the prediction error for each case is the residual, $y_i - \hat{y}_i$.

In each case, we would expect these errors to be distributed around zero. The question is: how much smaller do the errors

of the regression model tend to be, compared with the errors we make by predicting the outcome using the sample mean alone? If our predictions errors get a lot smaller with the x variable than without it, then we'll know that this variable conveys a lot of information about response.

To answer this question, return to Figure 6.6. To the right of the scatter plot you see two histograms: (1) the original deviations $y_i - \bar{y}$, and (2) the residuals from the regresson model. You'll notice that some of the original variation has been absorbed by the regression model: the residuals are less variable (standard deviation 1.5) than the original y points (standard deviation 2.7).

This is how a regression model measures the information content of a predictor: information means reduction in prediction error for the response. The bigger this reduction in prediction uncertainty, the more informative the predictor. The typical way of summarizing these numbers is via a statistic called R^2:

$$R^2 = 1 - \left[\frac{\text{sd(residuals)}}{\text{sd(original } y \text{ values)}}\right]^2$$

R^2 is always between 0 and 1, with 1 meaning "perfectively informative" and 0 meaning "not informative at all." For our data in Figure 6.6, R^2 is approximately $1 - (1.5/2.7)^2 \approx 0.7$. This means that approximately 70% of the variation in the CDC flu activity index can be predicted using this single search term, while the remaining 30% cannot be predicted using this search term.