CHAPTER 8

Linear Regression

WHO	Items on the Burger King menu
WHAT	Protein content and total fat content
UNITS	Grams of protein Grams of fat
HOW	Supplied by BK on request or at their Web site

The Whopper™ has been Burger King's signature sandwich since 1957. One Double Whopper with cheese provides 53 grams of protein—all the protein you need in a day. It also supplies 1020 calories and 65 grams of fat. The Daily Value (based on a 2000-calorie diet) for fat is 65 grams. So after a Double Whopper you'll want the rest of your calories that day to be fat-free.[1]

Of course, the Whopper isn't the only item Burger King sells. How are fat and protein related on the entire BK menu? The scatterplot of the *Fat* (in grams) versus the *Protein* (in grams) for foods sold at Burger King shows a positive, moderately strong, linear relationship.

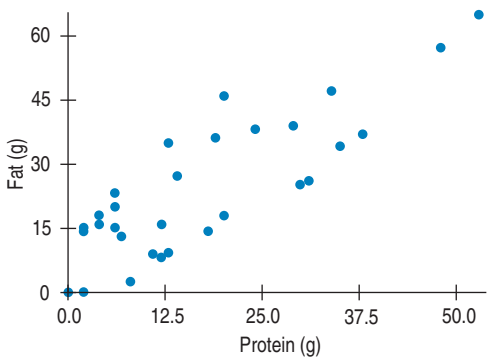


FIGURE 8.1
Total Fat *versus* Protein *for 30 items on the BK menu. The Double Whopper is in the upper right corner. It's extreme, but is it out of line?*

A S ***Video:* Manatees and Motorboats.** Are motorboats killing more manatees in Florida? Here's the story on video.

A S ***Activity:* Linear Equations.** For a quick review of linear equations, view this activity and play with the interactive tool.

If you want 25 grams of protein in your lunch, how much fat should you expect to consume at Burger King? The correlation between *Fat* and *Protein* is 0.83, a sign that the linear association seen in the scatterplot is fairly strong. But *strength* of the relationship is only part of the picture. The correlation says, "The linear association between these two variables is fairly strong," but it doesn't tell us *what the line is.*

[1] Sorry about the fries.

"Statisticians, like artists, have the bad habit of falling in love with their models."

—George Box, famous statistician

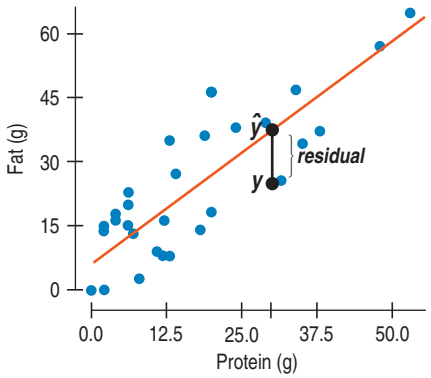
Now we *can* say more. We can **model** the relationship with a line and give its equation. The equation will let us predict the fat content for any Burger King food, given its amount of protein.

We met our first model in Chapter 6. We saw there that we can specify a Normal model with two parameters: its mean (μ) and standard deviation (σ).

For the Burger King foods, we'd choose a linear model to describe the relationship between *Protein* and *Fat*. The **linear model** is just an equation of a straight line through the data. Of course, no line can go through all the points, but a linear model can summarize the general pattern with only a couple of parameters. Like all models of the real world, the line will be wrong—wrong in the sense that it can't match reality *exactly*. But it can help us understand how the variables are associated.

A S ***Activity:*** **Residuals.** Residuals are the basis for fitting lines to scatterplots. See how they work.

Residuals



Not only can't we draw a line through all the points, the best line might not even hit *any* of the points. Then how can it be the "best" line? We want to find the line that somehow comes *closer* to all the points than any other line. Some of the points will be above the line and some below. For example, the line might suggest that a BK Broiler chicken sandwich with 30 grams of protein should have 36 grams of fat when, in fact, it actually has only 25 grams of fat. We call the estimate made from a model the **predicted value,** and write it as $\hat{y}$ (called *y-hat*) to distinguish it from the true value y (called, uh, y). The difference between the observed value and its associated predicted value is called the **residual.** The residual value tells us how far off the model's prediction is at that point. The BK Broiler chicken residual would be $y - \hat{y} = 25 - 36 = -11$ g of fat.

To find the residuals, we always subtract the predicted value from the observed one. The negative residual tells us that the actual fat content of the BK Broiler chicken is about 11 grams *less* than the model predicts for a typical Burger King menu item with 30 grams of protein.

Our challenge now is how to find the right line.

residual = observed value − predicted value
A *negative* residual means the predicted value is too big—an overestimate. And a *positive* residual shows that the model makes an underestimate. These may seem backwards until you think about them.

"Best Fit" Means Least Squares

A S ***Activity:*** **The Least Squares Criterion.** Does your sense of "best fit" look like the least squares line?

When we draw a line through a scatterplot, some residuals are positive and some negative. We can't assess how well the line fits by adding up all the residuals—the positive and negative ones would just cancel each other out. We faced the same issue when we calculated a standard deviation to measure spread. And we deal with it the same way here: by squaring the residuals. Squaring makes them all positive. Now we can add them up. Squaring also emphasizes the large residuals. After all, points near the line are consistent with the model, but we're more concerned about points far from the line. When we add all the squared residuals together, that sum indicates how well the line we drew fits the data—the smaller the sum, the better the fit. A different line will produce a different sum, maybe bigger, maybe smaller. The **line of best fit** is the line for which the sum of the squared residuals is smallest, the **least squares** line.

Who's on First

In 1805, Legendre was the first to publish the "least squares" solution to the problem of fitting a line to data when the points don't all fall exactly on the line. The main challenge was how to distribute the errors "fairly." After considerable thought, he decided to minimize the sum of the squares of what we now call the residuals. When Legendre published his paper, though, Gauss claimed he had been using the method since 1795. Gauss later referred to the "least squares" solution as "*our* method" (*principium nostrum*), which certainly didn't help his relationship with Legendre.

TI-nspire

Least squares. Try to minimize the sum of areas of residual squares as you drag a line across a scatterplot.

You might think that finding this line would be pretty hard. Surprisingly, it's not, although it was an exciting mathematical discovery when Legendre published it in 1805 (see margin note on previous page).

Correlation and the Line

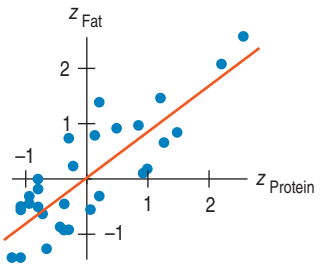


FIGURE 8.2
The Burger King scatterplot in z-scores.

If you suspect that what we know about correlation can lead us to the equation of the linear model, you're headed in the right direction. It turns out that it's not a very big step. In Chapter 7 we learned a lot about how correlation worked by looking at a scatterplot of the standardized variables. Here's a scatterplot of z_y (standardized *Fat*) vs. z_x (standardized *Protein*).

What line would you choose to model the relationship of the standardized values? Let's start at the center of the scatterplot. How much protein and fat does a *typical* Burger King food item provide? If it has average protein content, $\bar{x}$, what about its fat content? If you guessed that its fat content should be about average, $\bar{y}$, as well, then you've discovered the first property of the line we're looking for. The line must go through the point $(\bar{x}, \bar{y})$. In the plot of z-scores, then, the line passes through the origin (0, 0).

You might recall that the equation for a line that passes through the origin can be written with just a slope and no intercept:

$$y = mx.$$

The coordinates of our standardized points aren't written (x, y); their coordinates are z-scores: (z_x, z_y). We'll need to change our equation to show that. And we'll need to indicate that the point on the line corresponding to a particular z_x is $\hat{z}_y$, the model's estimate of the actual value of z_y. So our equation becomes

$$\hat{z}_y = mz_x.$$

NOTATION ALERT:

"Putting a hat on it" is standard Statistics notation to indicate that something has been predicted by a model. Whenever you see a hat over a variable name or symbol, you can assume it is the predicted version of that variable or symbol (and look around for the model).

Many lines with different slopes pass through the origin. Which one fits our data the best? That is, which slope determines the line that minimizes the sum of the squared residuals? It turns out that the best choice for m is the correlation coefficient itself, r! (You must really wonder where that stunning assertion comes from. Check the Math Box.)

Wow! This line has an equation that's about as simple as we could possibly hope for:

$$\hat{z}_y = rz_x.$$

Great. It's simple, but what does it tell us? It says that in moving one standard deviation from the mean in x, we can expect to move about r standard deviations away from the mean in y. Now that we're thinking about least squares lines, the correlation is more than just a vague measure of strength of association. It's a great way to think about what the model tells us.

Let's be more specific. For the sandwiches, the correlation is 0.83. If we standardize both protein and fat, we can write

$$\hat{z}_{Fat} = 0.83z_{Protein}.$$

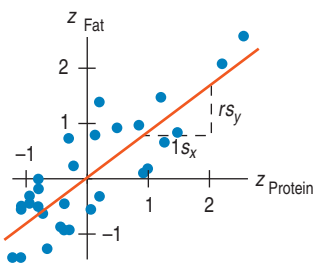


FIGURE 8.3
Standardized fat vs. standardized protein with the regression line. Each one standard deviation change in protein *results in a predicted change of r standard deviations in* fat.

This model tells us that for every standard deviation above (or below) the mean a sandwich is in protein, we'll predict that its fat content is 0.83 standard deviations above (or below) the mean fat content. A double hamburger has 31 grams of protein, about 1 SD above the mean. Putting 1.0 in for $z_{Protein}$ in the model gives a $\hat{z}_{Fat}$ value of 0.83. If you trust the model, you'd expect the fat content to be about 0.83 fat SDs above the mean fat level. Moving one standard deviation away from the mean in x moves our estimate r standard deviations away from the mean in y.

If $r = 0$, there's no linear relationship. The line is horizontal, and no matter how many standard deviations you move in x, the predicted value for y doesn't

change. On the other hand, if $r = 1.0$ or -1.0, there's a perfect linear association. In that case, moving any number of standard deviations in x moves exactly the same number of standard deviations in y. In general, moving any number of standard deviations in x moves r times that number of standard deviations in y.

How Big Can Predicted Values Get?

Sir Francis Galton was the first to speak of "regression," although others had fit lines to data by the same method.

Suppose you were told that a new male student was about to join the class, and you were asked to guess his height in inches. What would be your guess? A safe guess would be the mean height of male students. Now suppose you are also told that this student has a grade point average (*GPA*) of 3.9—about 2 SDs above the mean *GPA*. Would that change your guess? Probably not. The correlation between *GPA* and *height* is near 0, so knowing the *GPA* value doesn't tell you anything and doesn't move your guess. (And the equation tells us that as well, since it says that we should move 0×2 SDs from the mean.)

On the other hand, suppose you were told that, measured in centimeters, the student's height was 2 SDs above the mean. There's a perfect correlation between *height in inches* and *height in centimeters*, so you'd know he's 2 SDs above mean height in inches as well. (The equation would tell us to move 1.0×2 SDs from the mean.)

What if you're told that the student is 2 SDs above the mean in *shoe size*? Would you still guess that he's of average *height*? You might guess that he's taller than average, since there's a positive correlation between *height* and *shoe size*. But would you guess that he's 2 SDs above the mean? When there was no correlation, we didn't move away from the mean at all. With a perfect correlation, we moved our guess the full 2 SDs. Any correlation between these extremes should lead us to move somewhere between 0 and 2 SDs above the mean. (To be exact, the equation tells us to move $r \times 2$ standard deviations away from the mean.)

Notice that if x is 2 SDs above its mean, we won't ever guess more than 2 SDs away for y, since r can't be bigger than 1.0.[2] So, each predicted y tends to be closer to its mean (in standard deviations) than its corresponding x was. This property of the linear model is called **regression to the mean,** and the line is called the **regression line.**

The First Regression

Sir Francis Galton related the heights of sons to the heights of their fathers with a regression line. The slope of his line was less than 1. That is, sons of tall fathers were tall, but not as much above the average height as their fathers had been above their mean. Sons of short fathers were short, but generally not as far from their mean as their fathers. Galton interpreted the slope correctly as indicating a "regression" toward the mean height—and "regression" stuck as a description of the method he had used to find the line.

JUST CHECKING

A scatterplot of house *Price* (in thousands of dollars) vs. house *Size* (in thousands of square feet) for houses sold recently in Saratoga, NY shows a relationship that is straight, with only moderate scatter and no outliers. The correlation between house *Price* and house *Size* is 0.77.

1. You go to an open house and find that the house is 1 standard deviation above the mean in size. What would you guess about its price?
2. You read an ad for a house priced 2 standard deviations below the mean. What would you guess about its size?
3. A friend tells you about a house whose size in square meters (he's European) is 1.5 standard deviations above the mean. What would you guess about its size in square feet?

[2] In the last chapter we asserted that correlations max out at 1, but we never actually *proved* that. Here's yet another reason to check out the Math Box on the next page.

MATH BOX

Where does the equation of the line of best fit come from? To write the equation of any line, we need to know a point on the line and the slope. The point is easy. Consider the BK menu example. Since it is logical to predict that a sandwich with average protein will contain average fat, the line passes through the point $(\bar{x}, \bar{y})$.[3]

To think about the slope, we look once again at the z-scores. We need to remember a few things:

1. The mean of any set of z-scores is 0. This tells us that the line that best fits the z-scores passes through the origin (0,0).
2. The standard deviation of a set of z-scores is 1, so the variance is also 1. This means that $\frac{\sum(z_y - \bar{z}_y)^2}{n-1} = \frac{\sum(z_y - 0)^2}{n-1} = \frac{\sum z_y^2}{n-1} = 1$, a fact that will be important soon.
3. The correlation is $r = \frac{\sum z_x z_y}{n-1}$, also important soon.

Ready? Remember that our objective is to find the slope of the best fit line. Because it passes through the origin, its equation will be of the form $\hat{z}_y = mz_x$. We want to find the value for m that will minimize the sum of the squared residuals. Actually we'll divide that sum by $n - 1$ and minimize this "mean squared residual," or *MSR*. Here goes:

Minimize: $$MSR = \frac{\sum(z_y - \hat{z}_y)^2}{n-1}$$

Since $\hat{z}_y = mz_x$: $$MSR = \frac{\sum(z_y - mz_x)^2}{n-1}$$

Square the binomial: $$= \frac{\sum(z_y^2 - 2mz_xz_y + m^2z_x^2)}{n-1}$$

Rewrite the summation: $$= \frac{\sum z_y^2}{n-1} - 2m\frac{\sum z_xz_y}{n-1} + m^2\frac{\sum z_x^2}{n-1}$$

4. Substitute from (2) and (3): $$= 1 - 2mr + m^2$$

Wow! That simplified nicely! And as a bonus, the last expression is quadratic. Remember parabolas from algebra class? A parabola in the form $y = ax^2 + bx + c$ reaches its minimum at its turning point, which occurs when $x = \frac{-b}{2a}$. We can minimize the mean of squared residuals by choosing $m = \frac{-(-2r)}{2(1)} = r$.

Wow, again! The slope of the best fit line for z-scores is the correlation, r. This stunning fact immediately leads us to two important additional results, listed below. As you read on in the text, we explain them in the context of our continuing discussion of Burger King foods.

- A slope of r for z-scores means that for every increase of 1 standard deviation in z_x there is an increase of r standard deviations in $\hat{z}_y$. "Over one, up r," as you probably said in algebra class. Translate that back to the original x and y values: "Over one standard deviation in x, up r standard deviations in $\hat{y}$."

  That's it! In x- and y-values, the slope of the regression line is $b = \frac{rs_y}{s_x}$.

[3] It's actually not hard to prove this too.

- We know choosing $m = r$ minimizes the sum of the squared residuals, but how small does that sum get? Equation (4) told us that the mean of the squared residuals is $1 - 2mr + m^2$. When $m = r$, $1 - 2mr + m^2 = 1 - 2r^2 + r^2 = 1 - r^2$. This is the variability *not* explained by the regression line. Since the variance in z_y was 1 (Equation 2), the percentage of variability in y that is explained by x is r^2. This important fact will help us assess the strength of our models.

And there's still another bonus. Because r^2 is the percent of variability explained by our model, r^2 is at most 100%. If $r^2 \leq 1$, then $-1 \leq r \leq 1$, proving that correlations are always between -1 and $+1$. (Told you so!)

The Regression Line in Real Units

Why Is Correlation "r"?
In his original paper on correlation, Galton used r for the "index of correlation" that we now call the correlation coefficient. He calculated it from the regression of y on x or of x on y after standardizing the variables, just as we have done. It's fairly clear from the text that he used r to stand for (standardized) regression.

A S ***Simulation:*** **Interpreting Equations.** This demonstrates how to use and interpret linear equations.

When you read the Burger King menu, you probably don't think in z-scores. But you might want to know the fat content in grams for a specific amount of protein in grams.

How much fat should we predict for a double hamburger with 31 grams of protein? The mean protein content is near 17 grams and the standard deviation is 14, so that item is 1 SD above the mean. Since $r = 0.83$, we predict the fat content will be 0.83 SDs above the mean fat content. Great. How much fat is that? Well, the mean fat content is 23.5 grams and the standard deviation of fat content is 16.4, so we predict that the double hamburger will have $23.5 + 0.83 \times 16.4 = 37.11$ grams of fat.

We can always convert both x and y to z-scores, find the correlation, use $\hat{z}_y = rz_x$, and then convert $\hat{z}_y$ back to its original units so that we can understand the prediction. But can't we do this more simply?

Yes. Let's write the equation of the line for protein and fat—that is, the actual x and y values rather than their z-scores. In Algebra class you may have once seen lines written in the form $y = mx + b$. Statisticians do exactly the same thing, but with different notation:

$$\hat{y} = b_0 + b_1 x.$$

In this equation, b_0 is the ***y*-intercept,** the value of y where the line crosses the y-axis, and b_1 is the **slope**.[4]

Protein	Fat
$\bar{x} = 17.2$ g	$\bar{y} = 23.5$ g
$s_x = 14.0$ g	$s_y = 16.4$ g
$r = 0.83$	

First we find the slope, using the formula we developed in the Math Box.[5] Remember? We know that our model predicts that for each increase of one standard deviation in protein we'll see an increase of about 0.83 standard deviations in fat.

In other words, the slope of the line in original units is

$$b_1 = \frac{rs_y}{s_x} = \frac{0.83 \times 16.4 \text{ g fat}}{14 \text{ g protein}} = 0.97 \text{ grams of fat per gram of protein.}$$

Slope
$$b_1 = \frac{rs_y}{s_x}$$

Next, how do we find the y-intercept, b_0? Remember that the line has to go through the mean-mean point $(\bar{x}, \bar{y})$. In other words, the model predicts $\bar{y}$ to be the value that corresponds to $\bar{x}$. We can put the means into the equation and write $\bar{y} = b_0 + b_1\bar{x}$.

Solving for b_0, we see that the intercept is just $b_0 = \bar{y} - b_1\bar{x}$.

Intercept
$$b_0 = \bar{y} - b_1\bar{x}$$

[4] We changed from $mx + b$ to $b_0 + b_1x$ for a reason—not just to be difficult. Eventually we'll want to add more x's to the model to make it more realistic and we don't want to use up the entire alphabet. What would we use after m? The next letter is n, and that one's already taken. o? See our point? Sometimes subscripts are the best approach.

[5] Several important results popped up in that Math Box. Check it out!

For the Burger King foods, that comes out to

$$b_0 = 23.5 \text{ g fat} - 0.97 \frac{\text{g fat}}{\text{g protein}} \times 17.2 \text{ g protein} = 6.8 \text{ g fat}.$$

Putting this back into the regression equation gives

$$\widehat{fat} = 6.8 + 0.97\, protein.$$

Units of *y* per unit of *x*

Get into the habit of identifying the units by writing down "*y*-units per *x*-unit," with the unit names put in place. You'll find it'll really help you to Tell about the line in context.

What does this mean? The slope, 0.97, says that an additional gram of protein is associated with an additional 0.97 grams of fat, on average. Less formally, we might say that Burger King sandwiches pack about 0.97 grams of fat per gram of protein. Slopes are always expressed in *y*-units per *x*-unit. They tell how the *y*-variable changes (in its units) for a one-unit change in the *x*-variable. When you see a phrase like "students per teacher" or "kilobytes per second" think slope.

Changing the units of the variables doesn't change the *correlation*, but for the *slope*, units do matter. We may know that age and height in children are positively correlated, but the *value* of the slope depends on the units. If children grow an average of 3 inches per year, that's the same as 0.21 millimeters per day. For the slope, it matters whether you express age in days or years and whether you measure height in inches or millimeters. How you choose to express x and y—what units you use—affects the slope directly. Why? We know changing units doesn't change the correlation, but does change the standard deviations. The slope introduces the units into the equation by multiplying the correlation by the ratio of s_y to s_x. The units of the **slope** are always the units of y per unit of x.

How about the **intercept** of the BK regression line, 6.8? Algebraically, that's the value the line takes when x is zero. Here, our model predicts that even a BK item with no protein would have, on average, about 6.8 grams of fat. Is that reasonable? Well, the apple pie, with 2 grams of protein, has 14 grams of fat, so it's not impossible. But often 0 is not a plausible value for x (the year 0, a baby born weighing 0 grams, …). Then the intercept serves only as a starting value for our predictions and we don't interpret it as a meaningful predicted value.

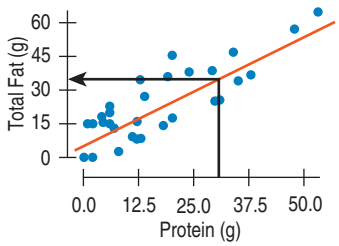


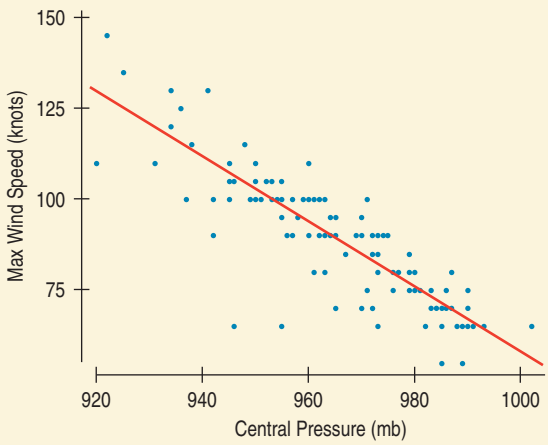
FIGURE 8.4
Burger King menu items in their natural units with the regression line.

FOR EXAMPLE A regression model for hurricanes

In Chapter 7 we looked at the relationship between the central pressure and maximum wind speed of Atlantic hurricanes. We saw that the scatterplot was straight enough, and then found a correlation of −0.879, but we had no model to describe how these two important variables are related or to allow us to predict wind speed from pressure. Since the conditions we need to check for regression are the same ones we checked before, we can use technology to find the regression model. It looks like this:

$$\widehat{MaxWindSpeed} = 955.27 - 0.897 CentralPressure$$

Question: Interpret this model. What does the slope mean in this context? Does the intercept have a meaningful interpretation?



The negative slope says that as *CentralPressure* falls, *MaxWindSpeed* increases. That makes sense from our general understanding of how hurricanes work: Low central pressure pulls in moist air, driving the rotation and the resulting destructive winds. The slope's value says that, on average, the maximum wind speed increases by about 0.897 knots for every 1-millibar drop in central pressure.

It's not meaningful, however, to interpret the intercept as the wind speed predicted for a central pressure of 0—that would be a vacuum. Instead, it is merely a starting value for the model.

With the estimated linear model, $\widehat{fat} = 6.8 + 0.97\ protein$, it's easy to predict fat content for any menu item we want. For example, for the BK Broiler chicken sandwich with 30 grams of *protein*, we can plug in 30 grams for the amount of *protein* and see that the *predicted fat* content is $6.8 + 0.97(30) = 35.9$ grams of fat. Because the BK Broiler chicken sandwich actually has 25 grams of fat, its residual is

$$fat - \widehat{fat} = 25 - 35.9 = -10.9 \text{ g}.$$

To use a regression model, we should check the same conditions for regressions as we did for correlation: the **Quantitative Variables Condition,** the **Straight Enough Condition,** and the **Outlier Condition.**

JUST CHECKING

Let's look again at the relationship between house *Price* (in thousands of dollars) and house *Size* (in thousands of square feet) in Saratoga. The regression model is

$$\widehat{Price} = -3.117 + 94.454\ Size.$$

4. What does the slope of 94.454 mean?
5. What are the units of the slope?
6. Your house is 2000 sq ft bigger than your neighbor's house. How much more do you expect it to be worth?
7. Is the *y*-intercept of −3.117 meaningful? Explain.

STEP-BY-STEP EXAMPLE Calculating a Regression Equation

Wildfires are an ongoing source of concern shared by several government agencies. In 2004, the Bureau of Land Management, Bureau of Indian Affairs, Fish and Wildlife Service, National Park Service, and USDA Forest Service spent a combined total of $890,233,000 on fire suppression, down from nearly twice that much in 2002. These government agencies join together in the National Interagency Fire Center, whose Web site (www.nifc.gov) reports statistics about wildfires.

Question: Has the annual number of wildfires been changing, on average? If so, how fast and in what way?

Plan State the problem.

I want to know how the number of wildfires in the continental United States has changed in the past two decades.

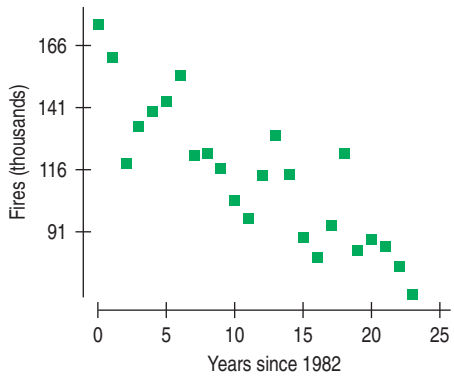
Variables Identify the variables and report the W's.

I have data giving the number of wildfires for each year (in thousands of fires) from 1982 to 2005.

Check the appropriate assumptions and conditions.

✓ ***Quantitative Variables Condition:*** Both the number of fires and the year are quantitative.

Just as we did for correlation, check the conditions for a regression by making a picture. Never fit a regression without looking at the scatterplot first.



✓ **Straight Enough Condition:** The scatterplot shows a strong linear relationship with a negative association.

✓ **Outlier Condition:** No outliers are evident in the scatterplot.

Because these conditions are satisfied, it is OK to model the relationship with a regression line.

Note: It's common (and usually simpler) not to use four-digit numbers to identify years. Here we have chosen to number the years beginning in 1982, so 1982 is represented as year 0 and 2005 as year 23.

SHOW

Mechanics Find the equation of the regression line. Summary statistics give the building blocks of the calculation.

(We generally report summary statistics to one more digit of accuracy than the data. We do the same for intercept and predicted values, but for slopes we usually report an additional digit. Remember, though, not to round off until you finish computing an answer.)[6]

Year:

$\bar{x} = 11.5$ (representing 1993.5)

$s_x = 7.07$ years

Fires:

$\bar{y} = 114.098$ fires

$s_y = 28.342$ fires

Correlation:

$r = -0.862$

Find the slope, b_1.

$$b_1 = \frac{rs_y}{s_x} = \frac{-0.862(28.342)}{7.07}$$

$$= -3.4556 \text{ fires per year}$$

Find the intercept, b_0.

$$b_0 = \bar{y} - b_1\bar{x} = 114.098 - (-3.4556)11.5$$

$$= 153.837$$

Write the equation of the model, using meaningful variable names.

So the least squares line is

$$\hat{y} = 153.837 - 3.4556x, \text{ or}$$

$$\widehat{Fires} = 153.837 - 3.4556\ year$$

[6] We warned you in Chapter 6 that we'll round in the intermediate steps of a calculation to show the steps more clearly. If you repeat these calculations yourself on a calculator or statistics program, you may get somewhat different results. When calculated with more precision, the intercept is 153,809 and the slope is −3.453.

TELL

Conclusion Interpret what you have found in the context of the question. Discuss in terms of the variables and their units.

During the period from 1982 to 2005, the annual number of fires declined at an average rate of about 3,456 (3.456 thousand) fires per year. For prediction, the model uses a base estimation of 153,837 fires in 1982.

A S

***Activity:* Find a Regression Equation.** Now that we've done it by hand, try it with technology using the statistics package paired with your version of *ActivStats*.

Residuals Revisited

Why *e* for "Residual"?
The flip answer is that *r* is already taken, but the truth is that *e* stands for "error." No, that doesn't mean it's a mistake. Statisticians often refer to variability not explained by a model as error.

The linear model we are using assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled. We can write

$$Data = Model + Residual$$

or, equivalently,

$$Residual = Data - Model.$$

Or, in symbols,

$$e = y - \hat{y}.$$

When we want to know how well the model fits, we can ask instead what the model missed. To see that, we look at the residuals.

FOR EXAMPLE Katrina's residual

Recap: The linear model relating hurricanes' wind speeds to their central pressures was

$$\widehat{MaxWindSpeed} = 955.27 - 0.897CentralPressure$$

Let's use this model to make predictions and see how those predictions do.

Question: Hurricane Katrina had a central pressure measured at 920 millibars. What does our regression model predict for her maximum wind speed? How good is that prediction, given that Katrina's actual wind speed was measured at 110 knots?

Substituting 920 for the central pressure in the regression model equation gives

$$\widehat{MaxWindSpeed} = 955.27 - 0.897(920) = 130.03$$

The regression model predicts a maximum wind speed of 130 knots for Hurricane Katrina.

The residual for this prediction is the observed value minus the predicted value:

$$110 - 130 = -20kts.$$

In the case of Hurricane Katrina, the model predicts a wind speed 20 knots higher than was actually observed.

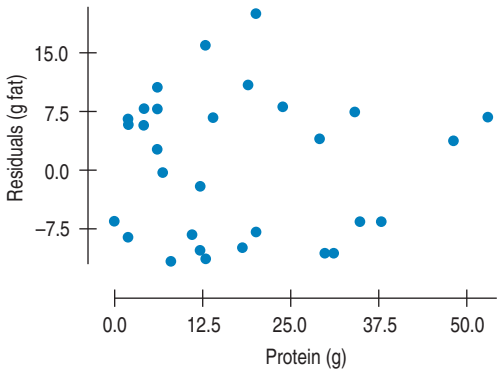


FIGURE 8.5

The residuals for the BK menu regression look appropriately boring.

Residuals help us to see whether the model makes sense. When a regression model is appropriate, it should model the underlying relationship. Nothing interesting should be left behind. So after we fit a regression model, we usually plot the residuals in the hope of finding . . . nothing.

A scatterplot of the residuals versus the x-values should be the most boring scatterplot you've ever seen. It shouldn't have any interesting features, like a direction or shape. It should stretch horizontally, with about the same amount of scatter throughout. It should show no bends, and it should have no outliers. If you see any of these features, find out what the regression model missed.

Most computer statistics packages plot the residuals against the predicted values $\hat{y}$, rather than against x. When the slope is negative, the two versions are mirror images. When the slope is positive, they're virtually identical except for the axis labels. Since all we care about is the patterns (or, better, lack of patterns) in the plot, it really doesn't matter which way we plot the residuals.

JUST CHECKING

Our linear model for Saratoga homes uses the *Size* (in thousands of square feet) to estimate the *Price* (in thousands of dollars): $\widehat{Price} = -3.117 + 94.454 Size$. Suppose you're thinking of buying a home there.

8. Would you prefer to find a home with a negative or a positive residual? Explain.

9. You plan to look for a home of about 3000 square feet. How much should you expect to have to pay?

10. You find a nice home that size selling for $300,000. What's the residual?

The Residual Standard Deviation

If the residuals show no interesting pattern when we plot them against x, we can look at how big they are. After all, we're trying to make them as small as possible. Since their mean is always zero, though, it's only sensible to look at how much they vary. The standard deviation of the residuals, s_e, gives us a measure of how much the points spread around the regression line. Of course, for this summary to make sense, the residuals should all share the same underlying spread, so we check to make sure that the residual plot has about the same amount of scatter throughout.

This gives us a new assumption: the **Equal Variance Assumption.** The associated condition to check is the **Does the Plot Thicken? Condition.** We check to make sure that the spread is about the same all along the line. We can check that either in the original scatterplot of y against x or in the scatterplot of residuals.

We estimate the standard deviation of the residuals in almost the way you'd expect:

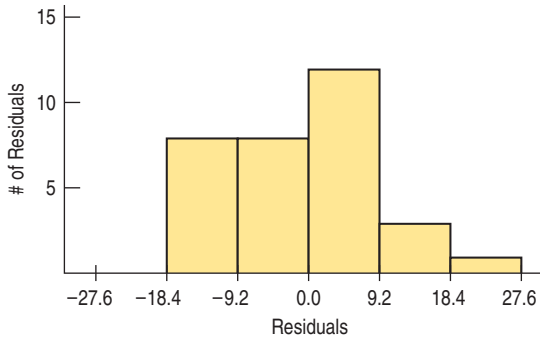
$$s_e = \sqrt{\frac{\Sigma e^2}{n - 2}}$$

We don't need to subtract the mean because the mean of the residuals $\bar{e} = 0$.

For the Burger King foods, the standard deviation of the residuals is 9.2 grams of fat. That looks about right in the scatterplot of residuals. The residual for the BK Broiler chicken was −11 grams, just over one standard deviation.

Why $n - 2$ rather than $n - 1$? We used $n - 1$ for s when we estimated the mean. Now we're estimating both a slope and an intercept. Looks like a pattern—and it is. We subtract one more for each parameter we estimate.

It's a good idea to make a histogram of the residuals. If we see a unimodal, symmetric histogram, then we can apply the 68–95–99.7 Rule to see how well the regression model describes the data. In particular, we know that 95% of the residuals should be no larger in size than $2s_e$. The Burger King residuals look like this:



Sure enough, almost all are less than 2(9.2), or 18.4, g of fat in size.

R^2—The Variation Accounted For

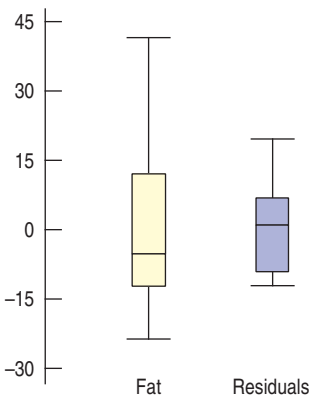


FIGURE 8.6

Compare the variability of total Fat *with the residuals from the regression. The means have been subtracted to make it easier to compare spreads. The variation left in the residuals is unaccounted for by the model, but it's less than the variation in the original data.*

TI-*nspire*

Understanding R^2. Watch the unexplained variability decrease as you drag points closer to the regression line.

The variation in the residuals is the key to assessing how well the model fits. Let's compare the variation of the response variable with the variation of the residuals. The total *Fat* has a standard deviation of 16.4 grams. The standard deviation of the residuals is 9.2 grams. If the correlation were 1.0 and the model predicted the *Fat* values perfectly, the residuals would all be zero and have no variation. We couldn't possibly do any better than that.

On the other hand, if the correlation were zero, the model would simply predict 23.5 grams of *Fat* (the mean) for all menu items. The residuals from that prediction would just be the observed *Fat* values minus their mean. These residuals would have the same variability as the original data because, as we know, just subtracting the mean doesn't change the spread.

How well does the BK regression model do? Look at the boxplots. The variation in the residuals is smaller than in the data, but certainly bigger than zero. That's nice to know, but how much of the variation is still left in the residuals? If you had to put a number between 0% and 100% on the fraction of the variation left in the residuals, what would you say?

All regression models fall somewhere between the two extremes of zero correlation and perfect correlation. We'd like to gauge where our model falls. As we showed in the Math Box,[7] the squared correlation, r^2, gives the fraction of the data's variation accounted for by the model, and $1 - r^2$ is the fraction of the original variation left in the residuals. For the Burger King model, $r^2 = 0.83^2 = 0.69$, and $1 - r^2$ is 0.31, so 31% of the variability in total *Fat* has been left in the residuals. How close was that to your guess?

All regression analyses include this statistic, although by tradition, it is written with a capital letter, $\boldsymbol{R^2}$, and pronounced "*R*-squared." An R^2 of 0 means that none of the variance in the data is in the model; all of it is still in the residuals. It would be hard to imagine using that model for anything.

[7] Have you looked yet? Please do.

Is a correlation of 0.80 twice as strong as a correlation of 0.40? Not if you think in terms of R^2. A correlation of 0.80 means an R^2 of $0.80^2 = 64\%$. A correlation of 0.40 means an R^2 of $0.40^2 = 16\%$—only a quarter as much of the variability accounted for. A correlation of 0.80 gives an R^2 *four* times as strong as a correlation of 0.40 and accounts for four times as much of the variability.

Because R^2 is a fraction of a whole, it is often given as a percentage.[8] For the Burger King data, R^2 is 69%. When interpreting a regression model, you need to *Tell* what R^2 means. According to our linear model, 69% of the variability in the fat content of Burger King sandwiches is accounted for by variation in the protein content.

How can we see that R^2 is really the fraction of variance accounted for by the model? It's a simple calculation. The variance of the fat content of the Burger King foods is $16.4^2 = 268.42$. If we treat the residuals as data, the variance of the residuals is 83.195.[9] As a fraction, that's $83.195/268.42 = 0.31$, or 31%. That's the fraction of the variance that is not accounted for by the model. The fraction that is accounted for is $100\% - 31\% = 69\%$, just the value we got for R^2.

FOR EXAMPLE Interpreting R^2

Recap: Our regression model that predicts maximum wind speed in hurricanes based on the storm's central pressure has $R^2 = 77.3\%$.

Question: What does that say about our regression model?

An R^2 of 77.3% indicates that 77.3% of the variation in maximum wind speed can be accounted for by the hurricane's central pressure. Other factors, such as temperature and whether the storm is over water or land, may explain some of the remaining variation.

JUST CHECKING

Back to our regression of house *Price* (in thousands of $) on house *Size* (in thousands of square feet). The R^2 value is reported as 59.5%, and the standard deviation of the residualsis 53.79.

11. What does the R^2 value mean about the relationship of *Price* and *Size*?

12. Is the correlation of *Price* and *Size* positive or negative? How do you know?

13. If we measure house *Size* in square meters instead, would R^2 change? Would the slope of the line change? Explain.

14. You find that your house in Saratoga is worth $100,000 more than the regression model predicts. Should you be very surprised (as well as pleased)?

How Big Should R^2 Be?

R^2 is always between 0% and 100%. But what's a "good" R^2 value? The answer depends on the kind of data you are analyzing and on what you want to do with it. Just as with correlation, there is no value for R^2 that automatically determines

[8] By contrast, we usually give correlation coefficients as decimal values between −1.0 and 1.0.
[9] This isn't quite the same as squaring the s_e that we discussed on the previous page, but it's very close. We'll deal with the distinction in Chapter 27.

Some Extreme Tales

One major company developed a method to differentiate between proteins. To do so, they had to distinguish between regressions with R^2 of 99.99% and 99.98%. For this application, 99.98% was not high enough.

The president of a financial services company reports that although his regressions give R^2 below 2%, they are highly successful because those used by his competition are even lower.

that the regression is "good." Data from scientific experiments often have R^2 in the 80% to 90% range and even higher. Data from observational studies and surveys, though, often show relatively weak associations because it's so difficult to measure responses reliably. An R^2 of 50% to 30% or even lower might be taken as evidence of a useful regression. The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.

As we've seen, an R^2 of 100% is a perfect fit, with no scatter around the line. The s_e would be zero. All of the variance is accounted for by the model and none is left in the residuals at all. This sounds great, but it's too good to be true for real data.[10]

Along with the slope and intercept for a regression, you should always report R^2 so that readers can judge for themselves how successful the regression is at fitting the data. Statistics is about variation, and R^2 measures the success of the regression model in terms of the fraction of the variation of y accounted for by the regression. R^2 is the first part of a regression that many people look at because, along with the scatterplot, it tells whether the regression model is even worth thinking about.

Regression Assumptions and Conditions

Make a Picture

To use regression, first check that

- the scatterplot is straight enough.

After you've fit the regression, make a residual plot and check that there are no obvious patterns. In particular, check that

- there are no obvious bends,
- the spread of the residuals is about the same throughout, and
- there are no obvious outliers.

The linear regression model is perhaps the most widely used model in all of Statistics. It has everything we could want in a model: two easily estimated parameters, a meaningful measure of how well the model fits the data, and the ability to predict new values. It even provides a self-check in plots of the residuals to help us avoid silly mistakes.

Like all models, though, linear models don't apply all the time, so we'd better think about whether they're reasonable. It makes no sense to make a scatterplot of categorical variables, and even less to perform a regression on them. Always check the **Quantitative Variables Condition** to be sure a regression is appropriate.

The linear model makes several assumptions. First, and foremost, is the **Linearity Assumption**—that the relationship between the variables is, in fact, linear. You can't verify an assumption, but you can check the associated condition. A quick look at the scatterplot will help you check the **Straight Enough Condition.** You don't need a *perfectly* straight plot, but it must be straight enough for the linear model to make sense. If you try to model a curved relationship with a straight line, you'll usually get exactly what you deserve.

If the scatterplot is not straight enough, stop here. You can't use a linear model for *any* two variables, even if they are related. They must have a *linear* association, or the model won't mean a thing.

For the standard deviation of the residuals to summarize the scatter, all the residuals should share the same spread, so we need the **Equal Variance Assumption.** The **Does the Plot Thicken? Condition** checks for changing spread in the scatterplot.

Check the **Outlier Condition.** Outlying points can dramatically change a regression model. Outliers can even change the sign of the slope, misleading us about the underlying relationship between the variables. We'll see examples in the next chapter.

Even though we've checked the conditions in the scatterplot of the data, a scatterplot of the residuals can sometimes help us see any violations even more

[10] If you see an R^2 of 100%, it's a good idea to figure out what happened. You may have discovered a new law of Physics, but it's much more likely that you accidentally regressed two variables that measure the same thing.

clearly. And examining the residuals is the best way to look for additional patterns and interesting quirks in the data.

A Tale of Two Regressions

Regression slopes may not behave exactly the way you'd expect at first. Our regression model for the Burger King sandwiches was $\widehat{fat} = 6.8 + 0.97\ protein$. That equation allowed us to estimate that a sandwich with 30 grams of protein would have 35.9 grams of fat. Suppose, though, that we knew the fat content and wanted to predict the amount of protein. It might seem natural to think that by solving our equation for *protein* we'd get a model for predicting *protein* from *fat*. But that doesn't work.

Our original model is $\hat{y} = b_0 + b_1 x$, but the new one needs to evaluate an $\hat{x}$ based on a value of y. There's no y in our original model, only $\hat{y}$, and that makes all the difference. Our model doesn't fit the BK data values perfectly, and the least squares criterion focuses on the vertical errors the model makes in using to model y—not on horizontal errors related to x.

A quick look at the equations reveals why. Simply solving our equation for x would give a new line whose slope is the reciprocal of ours. To model y in terms of x, our slope is $b_1 = \frac{rs_y}{s_x}$. To model x in terms of y, we'd need to use the slope $b_1 = \frac{rs_x}{s_y}$. Notice that it's *not* the reciprocal of ours.

Protein	Fat
$\bar{x} = 17.2$ g	$\bar{y} = 23.5$ g
$s_x = 14.0$ g	$s_y = 16.4$ g
$r = 0.83$	

If we want to predict *protein* from *fat*, we need to create that model. The slope is $b_1 = \frac{(0.83)(14.0)}{16.4} = 0.709$ grams of protein per gram of fat. The equation turns out to be $\widehat{protein} = 0.55 + 0.709\ fat$, so we'd predict that a sandwich with 35.9 grams of fat should have 26.0 grams of protein—not the 30 grams that we used in the first equation.

Moral of the story: *Think.* (Where have you heard *that* before?) Decide which variable you want to use (x) to predict values for the other (y). Then find the model that does that. If, later, you want to make predictions in the other direction, you'll need to start over and create the other model from scratch.

STEP-BY-STEP EXAMPLE — Regression

Even if you hit the fast food joints for lunch, you should have a good breakfast. Nutritionists, concerned about "empty calories" in breakfast cereals, recorded facts about 77 cereals, including their *Calories* per serving and *Sugar* content (in grams).

Question: How are calories and sugar content related in breakfast cereals?

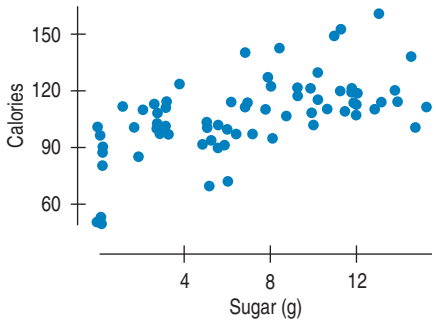
Plan State the problem and determine the role of the variables.

I am interested in the relationship between sugar content and calories in cereals. I'll use *Sugar* to estimate *Calories*.

Variables Name the variables and report the W's.

✓ **Quantitative Variables Condition:** I have two quantitative variables, *Calories* and *Sugar* content per serving, measured on 77 breakfast cereals. The units of measurement are calories and grams of sugar, respectively.

Check the conditions for a regression by making a picture. Never fit a regression without looking at the scatterplot first.



- ✓ **Outlier Condition:** There are no obvious outliers or groups.
- ✓ The **Straight Enough Condition** is satisfied; I will fit a regression model to these data.
- ✓ The **Does the Plot Thicken? Condition** is satisfied. The spread around the line looks about the same throughout.

SHOW

Mechanics If there are no clear violations of the conditions, fit a straight line model of the form $\hat{y} = b_0 + b_1 x$ to the data. Summary statistics give the building blocks of the calculation.

Calories

$\bar{y} = 107.0$ calories

$s_y = 19.5$ calories

Sugar

$\bar{x} = 7.0$ grams

$s_x = 4.4$ grams

Correlation

$r = 0.564$

Find the slope.

$$b_1 = \frac{rs_y}{s_x} = \frac{0.564(19.5)}{4.4}$$

$= 2.50$ calories per gram of sugar.

Find the intercept.

$b_0 = \bar{y} - b_1\bar{x} = 107 - 2.50(7) = 89.5$ calories.

Write the equation, using meaningful variable names.

So the least squares line is

$$\hat{y} = 89.5 + 2.50\,x \text{ or}$$

$$\widehat{\text{Calories}} = 89.5 + 2.50\,\text{Sugar}.$$

State the value of R^2.

Squaring the correlation gives

$$R^2 = 0.564^2 = 0.318 \text{ or } 31.8\%.$$

TELL

Conclusion Describe what the model says in words and numbers. Be sure to use the names of the variables and their units.

The scatterplot shows a positive, linear relationship and no outliers. The slope of the least squares regression line suggests that cereals have about 2.50 Calories more per additional gram of Sugar.

The key to interpreting a regression model is to start with the phrase "b_1 y-units per x-unit," substituting the estimated value of the slope for b_1 and the names of the

respective units. The intercept is then a starting or base value.

The intercept predicts that sugar-free cereals would average about 89.5 calories.

R^2 gives the fraction of the variability of y accounted for by the linear regression model.

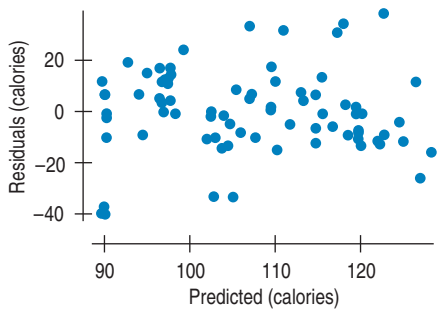
The R^2 says that 31.8% of the variability in Calories is accounted for by variation in Sugar content.

Find the standard deviation of the residuals, s_e, and compare it to the original s_y.

s_e = 16.2 calories. That's smaller than the original SD of 19.5, but still fairly large.

THINK AGAIN

Check Again Even though we looked at the scatterplot *before* fitting a regression model, a plot of the residuals is essential to any regression analysis because it is the best check for additional patterns and interesting quirks in the data.



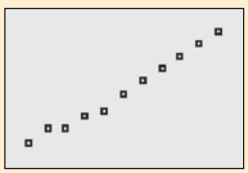
The residuals show a horizontal direction, a shapeless form, and roughly equal scatter for all predicted values. The linear model appears to be appropriate.

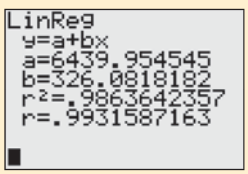
TI-*nspire*

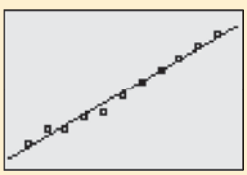
Residuals plots. See how the residuals plot changes as you drag points around in a scatterplot.

TI Tips

Regression lines and residuals plots







By now you will not be surprised to learn that your calculator can do it all: scatterplot, regression line, and residuals plot. Let's try it using the Arizona State tuition data from the last chapter. (TI Tips, p. 149) You should still have that saved in lists named YR and TUIT. First, recreate the scatterplot.

1. Find the equation of the regression line.

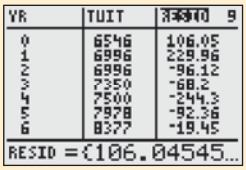
Actually, you already found the line when you used the calculator to get the correlation. But this time we'll be a little fancier so that we can display the line on our scatterplot. We want to tell the calculator to do the regression and save the equation of the model as a graphing variable.

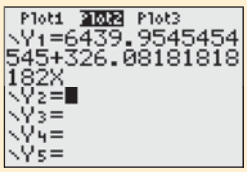
- Under STAT CALC choose LinReg(a+bx).
- Specify that x and y are YR and TUIT, as before, but . . .
- Now add a comma and one more specification. Press VARS, go to the Y-VARS menu, choose 1:Function, and finally(!) choose Y1.
- Hit ENTER.

There's the equation. The calculator tells you that the regression line is $\widehat{tuit} = 6440 + 326\ year$. Can you explain what the slope and y-intercept mean?

2. Add the line to the plot.

When you entered this command, the calculator automatically saved the equation as Y1. Just hit GRAPH to see the line drawn across your scatterplot.





3. Check the residuals.

Remember, you are not finished until you check to see if a linear model is appropriate. That means you need to see if the residuals appear to be randomly distributed. To do that, you need to look at the residuals plot.

This is made easy by the fact that the calculator has already placed the residuals in a list named RESID. Want to see them? Go to STAT EDIT and look through the lists. (If RESID is not already there, go to the first blank list and import the name RESID from your LIST NAMES menu. The residuals should appear.) Every time you have the calculator compute a regression analysis, it will automatically save this list of residuals for you.

4. Now create the residuals plot.

- Set up STAT PLOT Plot2 as a scatterplot with Xlist:YR and Ylist:RESID.
- Before you try to see the plot, go to the Y= screen. By moving the cursor around and hitting ENTER in the appropriate places you can turn off the regression line and Plot1, and turn on Plot2.
- ZoomStat will now graph the residuals plot.

Uh-oh! See the curve? The residuals are high at both ends, low in the middle. Looks like a linear model may not be appropriate after all. Notice that the residuals plot makes the curvature much clearer than the original scatterplot did.

Moral: Always check the residuals plot!

So a linear model might not be appropriate here. What now? The next two chapters provide techniques for dealing with data like these.

Reality Check: Is the Regression Reasonable?

Adjective, Noun, or Verb

You may see the term *regression* used in different ways. There are many ways to fit a line to data, but the term "regression line" or "regression" without any other qualifiers always means least squares. People also use *regression* as a verb when they speak of *regressing* a *y*-variable on an *x*-variable to mean fitting a linear model.

Statistics don't come out of nowhere. They are based on data. The results of a statistical analysis should reinforce your common sense, not fly in its face. If the results are surprising, then either you've learned something new about the world or your analysis is wrong.

Whenever you perform a regression, think about the coefficients and ask whether they make sense. Is a slope of 2.5 calories per gram of sugar reasonable? That's hard to say right off. We know from the summary statistics that a typical cereal has about 100 calories and 7 grams of sugar. A gram of sugar contributes some calories (actually, 4, but you don't need to know that), so calories should go up with increasing sugar. The direction of the slope seems right.

To see if the *size* of the slope is reasonable, a useful trick is to consider its order of magnitude. We'll start by asking if deflating the slope by a factor of 10 seems reasonable. Is 0.25 calories per gram of sugar enough? The 7 grams of sugar found in the average cereal would contribute less than 2 calories. That seems too small.

Now let's try inflating the slope by a factor of 10. Is 25 calories per gram reasonable? Then the average cereal would have 175 calories from sugar alone. The average cereal has only 100 calories per serving, though, so that slope seems too big.

We have tried inflating the slope by a factor of 10 and deflating it by 10 and found both to be unreasonable. So, like Goldilocks, we're left with the value in the middle that's just right. And an increase of 2.5 calories per gram of sugar is certainly *plausible.*

The small effort of asking yourself whether the regression equation is plausible is repaid whenever you catch errors or avoid saying something silly or absurd about the data. It's too easy to take something that comes out of a computer at face value and assume that it makes sense.

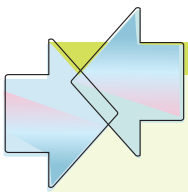
Always be skeptical and ask yourself if the answer is reasonable.

WHAT CAN GO WRONG?

There are many ways in which data that appear at first to be good candidates for regression analysis may be unsuitable. And there are ways that people use regression that can lead them astray. Here's an overview of the most common problems. We'll discuss them at length in the next chapter.

- **Don't fit a straight line to a nonlinear relationship.** Linear regression is suited only to relationships that are, well, *linear.* Fortunately, we can often improve the linearity easily by using re-expression. We'll come back to that topic in Chapter 10.
- **Beware of extraordinary points.** Data points can be extraordinary in a regression in two ways: They can have y-values that stand off from the linear pattern suggested by the bulk of the data, or extreme x-values. Both kinds of extraordinary points require attention.
- **Don't extrapolate beyond the data.** A linear model will often do a reasonable job of summarizing a relationship in the narrow range of observed x-values. Once we have a working model for the relationship, it's tempting to use it. But beware of predicting y-values for x-values that lie outside the range of the original data. The model may no longer hold there, so such *extrapolations* too far from the data are dangerous.
- **Don't infer that *x* causes *y* just because there is a good linear model for their relationship.** When two variables are strongly correlated, it is often tempting to assume a causal relationship between them. Putting a regression line on a scatterplot tempts us even further, but it doesn't make the assumption of causation any more valid. For example, our regression model predicting hurricane wind speeds from the central pressure was reasonably successful, but the relationship is very complex. It is reasonable to say that low central pressure at the eye is responsible for the high winds because it draws moist, warm air into the center of the storm, where it swirls around, generating the winds. But as is often the case, things aren't quite that simple. The winds themselves also contribute to lowering the pressure at the center of the storm as it becomes a hurricane. Understanding causation requires far more work than just finding a correlation or modeling a relationship.
- **Don't choose a model based on R^2 alone.** Although R^2 measures the *strength* of the linear association, a high R^2 does not demonstrate the *appropriateness* of the regression. A single outlier, or data that separate into two groups rather than a single cloud of points, can make R^2 seem quite large when, in fact, the linear regression model is simply inappropriate. Conversely, a low R^2 value may be due to a single outlier as well. It may be that most of the data fall roughly along a straight line, with the exception of a single point. Always look at the scatterplot.

R^2 does not mean that protein accounts for 69% of the fat in a BK food item. It is the *variation* in fat content that is accounted for by the linear model.



CONNECTIONS

We've talked about the importance of models before, but have seen only the Normal model as an example. The linear model is one of the most important models in Statistics. Chapter 7 talked about the assignment of variables to the y- and x-axes. That didn't matter to correlation, but it does matter to regression because y is predicted by x in the regression model.

The connection of R^2 to correlation is obvious, although it may not be immediately clear that just by squaring the correlation we can learn the fraction of the variability of y accounted for by a regression on x. We'll return to this in subsequent chapters.

We made a big fuss about knowing the units of your quantitative variables. We didn't need units for correlation, but without the units we can't define the slope of a regression. A regression makes no sense if you don't know the *Who,* the *What,* and the *Units* of both your variables.

We've summed squared deviations before when we computed the standard deviation and variance. That's not coincidental. They are closely connected to regression.

When we first talked about models, we noted that deviations away from a model were often interesting. Now we have a formal definition of these deviations as residuals.

WHAT HAVE WE LEARNED?

We've learned that when the relationship between quantitative variables is fairly straight, a linear model can help summarize that relationship and give us insights about it:

- The regression (best fit) line doesn't pass through all the points, but it is the best compromise in the sense that the sum of squares of the residuals is the smallest possible.

We've learned several things the correlation, *r*, tells us about the regression:

- The slope of the line is based on the correlation, adjusted for the units of x and y:

$$b_1 = \frac{rs_y}{s_x}$$

We've learned to interpret that slope in context:

- For each SD of x that we are away from the x mean, we expect to be r SDs of y away from the y mean.
- Because r is always between -1 and $+1$, each predicted y is fewer SDs away from its mean than the corresponding x was, a phenomenon called regression to the mean.
- The square of the correlation coefficient, R^2, gives us the fraction of the variation of the response accounted for by the regression model. The remaining $1 - R^2$ of the variation is left in the residuals.

The residuals also reveal how well the model works:

- If a plot of residuals against predicted values shows a pattern, we should re-examine the data to see why.
- The standard deviation of the residuals, s_e, quantifies the amount of scatter around the line.

Of course, the linear model makes no sense unless the **Linearity Assumption** is satisfied. We check the **Straight Enough Condition** and **Outlier Condition** with a scatterplot, as we did for correlation, and also with a plot of residuals against either the x or the predicted values. For the standard deviation of the residuals to make sense as a summary, we have to make the **Equal Variance Assumption**. We check it by looking at both the original scatterplot and the residual plot for the **Does the Plot Thicken? Condition.**

Terms

Model 172. An equation or formula that simplifies and represents reality.

Linear model 172. A linear model is an equation of a line. To interpret a linear model, we need to know the variables (along with their W's) and their units.

Predicted value 172. The value of $\hat{y}$ found for a given x-value in the data. A predicted value is found by substituting the x-value in the regression equation. The predicted values are the values on the fitted line; the points $(x, \hat{y})$ all lie exactly on the fitted line.

Residuals 172. Residuals are the differences between data values and the corresponding values predicted by the regression model—or, more generally, values predicted by any model.

$$\text{Residual} = \text{observed value} - \text{predicted value} = e = y - \hat{y}$$

Least squares 172. The least squares criterion specifies the unique line that minimizes the variance of the residuals or, equivalently, the sum of the squared residuals.

Regression to the mean 174. Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding x was from its mean. This is called regression to the mean.

Regression line
Line of best fit 174. The particular linear equation

$$\hat{y} = b_0 + b_1 x$$

that satisfies the least squares criterion is called the least squares regression line. Casually, we often just call it the regression line, or the line of best fit.

Slope

176. The slope, b_1, gives a value in "y-units *per* x-unit." Changes of one unit in x are associated with changes of b_1 units in predicted values of y. The slope can be found by

$$b_1 = \frac{rs_y}{s_x}.$$

Intercept

176. The intercept, b_0, gives a starting value in y-units. It's the $\hat{y}$-value when x is 0. You can find it from $b_0 = \bar{y} - b_1\bar{x}$.

s_e

181. The standard deviation of the residuals is found by $s_e = \sqrt{\frac{\Sigma e^2}{n-2}}$. When the assumptions and conditions are met, the residuals can be well described by using this standard deviation and the 68–95–99.7 Rule.

R^2

- 182. R^2 is the square of the correlation between y and x.
- R^2 gives the fraction of the variability of y accounted for by the least squares linear regression on x.
- R^2 is an overall measure of how successful the regression is in linearly relating y to x.

Skills

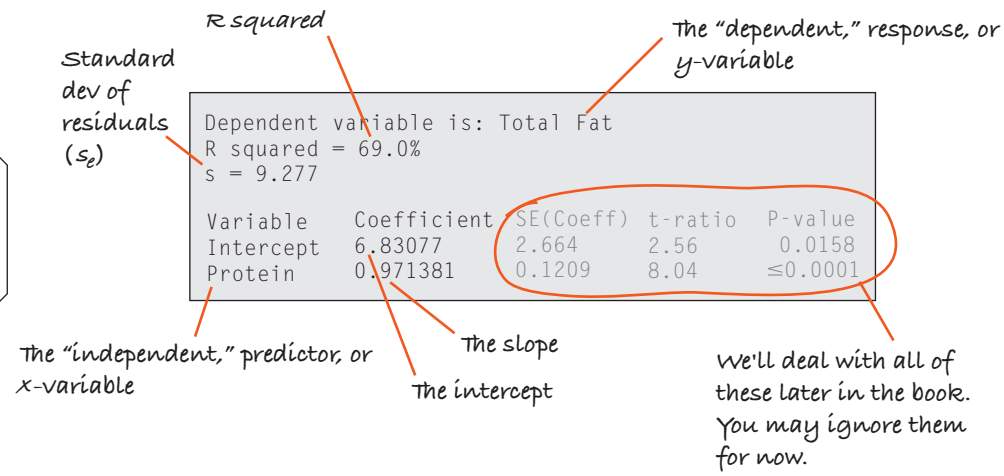
- Be able to identify response (y) and explanatory (x) variables in context.
- Understand how a linear equation summarizes the relationship between two variables.
- Recognize when a regression should be used to summarize a linear relationship between two quantitative variables.
- Be able to judge whether the slope of a regression makes sense.
- Know how to examine your data for violations of the **Straight Enough Condition** that would make it inappropriate to compute a regression.
- Understand that the least squares slope is easily affected by extreme values.
- Know that residuals are the differences between the data values and the corresponding values predicted by the line and that the *least squares criterion* finds the line that minimizes the sum of the squared residuals.
- Know how to use a plot of residuals against predicted values to check the **Straight Enough Condition**, the **Does the Plot Thicken? Condition,** and the **Outlier Condition.**
- Understand that the standard deviation of the residuals, s_e, measures variability around the line. A large s_e means the points are widely scattered; a small s_e means they lie close to the line.

- Know how to find a regression equation from the summary statistics for each variable and the correlation between the variables.
- Know how to find a regression equation using your statistics software and how to find the slope and intercept values in the regression output table.
- Know how to use regression to predict a value of y for a given x.
- Know how to compute the residual for each data value and how to display the residuals.

- Be able to write a sentence explaining what a linear equation says about the relationship between y and x, basing it on the fact that the slope is given in *y-units per x-unit.*
- Understand how the correlation coefficient and the regression slope are related. Know how R^2 describes how much of the variation in y is accounted for by its linear relationship with x.
- Be able to describe a prediction made from a regression equation, relating the predicted value to the specified x-value.
- Be able to write a sentence interpreting s_e as representing typical errors in predictions—the amounts by which actual y-values differ from the $\hat{y}$'s estimated by the model.

REGRESSION ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:



Finding Least Squares Lines. We almost always use technology to find regressions. Practice now—just in time for the exercises.

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labeled with the name of the x-variable, and the intercept is labeled "Intercept" or "Constant." So the regression equation shown here is

$$\widehat{Fat} = 6.83077 + 0.97138 Protein.$$

It is not unusual for statistics packages to give many more digits of the estimated slope and intercept than could possibly be estimated from the data. (The original data were reported to the nearest gram.) Ordinarily, you should round most of the reported numbers to one digit more than the precision of the data, and the slope to two. We will learn about the other numbers in the regression table later in the book. For now, all you need to be able to do is find the coefficients, the s_e, and the R^2 value.

EXERCISES

1. **Cereals.** For many people, breakfast cereal is an important source of fiber in their diets. Cereals also contain potassium, a mineral shown to be associated with maintaining a healthy blood pressure. An analysis of the amount of fiber (in grams) and the potassium content (in milligrams) in servings of 77 breakfast cereals produced the regression model $\widehat{Potassium} = 38 + 27 Fiber$. If your cereal provides 9 grams of fiber per serving, how much potassium does the model estimate you will get?

2. **Horsepower.** In Chapter 7's Exercise 33 we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model $\widehat{mpg} = 46.87 - 0.084 HP$. If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?

3. **More cereal.** Exercise 1 describes a regression model that estimates a cereal's potassium content from the amount of fiber it contains. In this context, what does it mean to say that a cereal has a negative residual?

4. **Horsepower, again.** Exercise 2 describes a regression model that uses a car's horsepower to estimate its fuel economy. In this context, what does it mean to say that a certain car has a positive residual?

5. **Another bowl.** In Exercise 1, the regression model $\widehat{Potassium} = 38 + 27 Fiber$ relates fiber (in grams) and potassium content (in milligrams) in servings of breakfast cereals. Explain what the slope means.

6. **More horsepower.** In Exercise 2, the regression model $\widehat{mpg} = 46.87 - 0.084 HP$ relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.