CHAPTER

27

Inferences for Regression

WHO	250 male subjects
WHAT	Body fat and waist size
UNITS	% Body fat and inches
WHEN	1990s
WHERE	United States
WHY	Scientific research

Three percent of a man's body is essential fat. (For a woman, the percentage is closer to 12.5%.) As the name implies, essential fat is necessary for a normal, healthy body. Fat is stored in small amounts throughout your body. Too much body fat, however, can be dangerous to your health. For men between 18 and 39 years old, a healthy percent body fat ranges from 8% to 19%. (For women of the same age, it's 21% to 32%.)

Measuring body fat can be tedious and expensive. The "standard reference" measurement is by dual-energy X-ray absorptiometry (DEXA), which involves two low-dose X-ray generators and takes from 10 to 20 minutes.

How close can we get to a useable prediction of body fat from easily measurable variables such as *Height*, *Weight*, or *Waist* size? Here's a scatterplot of *%Body Fat* plotted against *Waist* size for a sample of 250 males of various ages.

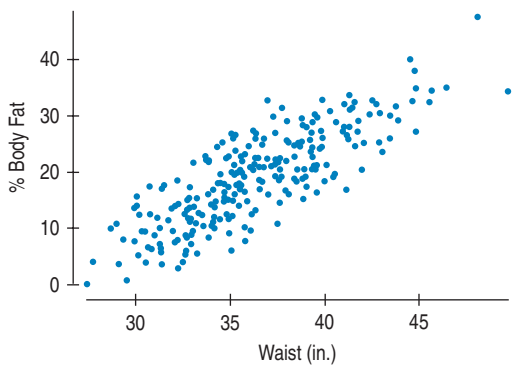


FIGURE 27.1
Percent Body Fat *vs.* Waist *size for 250 men of various ages. The scatterplot shows a strong, positive, linear relationship.*

Back in Chapter 8 we modeled relationships like this by fitting a least squares line. The plot is clearly straight, so we can find that line. The equation of the least squares line for these data is

$$\widehat{\%Body\ Fat} = -42.7 + 1.7\,Waist.$$

The slope says that, on average, *%Body Fat* is greater by 1.7 percent for each additional inch around the waist.

How useful is this model? When we fit linear models before, we used them to describe the relationship between the variables and we interpreted the slope and intercept as descriptions of the data. Now we'd like to know what the regression model can tell us beyond the 250 men in this study. To do that, we'll want to make confidence intervals and test hypotheses about the slope and intercept of the regression line.

The Population and the Sample

When we found a confidence interval for a mean, we could imagine a single, true underlying value for the mean. When we tested whether two means or two proportions were equal, we imagined a true underlying difference. But what does it mean to do inference for regression? We know better than to think that even if we knew every population value, the data would line up perfectly on a straight line. After all, even in our sample, not all men who have 38-inch waists have the same *%Body Fat*. In fact, there's a whole distribution of *%Body Fat* for these men:

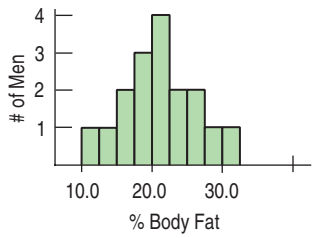


FIGURE 27.2
The distribution of %Body Fat *for men with a* Waist *size of 38 inches is unimodal and symmetric.*

This is true at each *Waist* size. In fact, we could depict the distribution of *%Body Fat* at different *Waist* sizes like this:

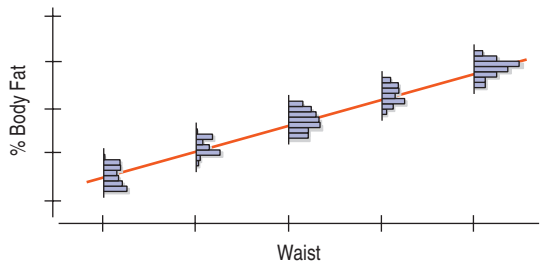


FIGURE 27.3
There's a distribution of %Body Fat *for each value of* Waist *size. We'd like the means of these distributions to line up.*

But we want to *model* the relationship between *%Body Fat* and *Waist* size for all men. To do that, we imagine an idealized regression line. The model assumes that the *means* of the distributions of *%Body Fat* for each *Waist* size fall along the line, even though the individuals are scattered around it. We know that this model is not a perfect description of how the variables are associated, but it may be useful for predicting *%Body Fat* and for understanding how it's related to *Waist* size.

If only we had all the values in the population, we could find the slope and intercept of this *idealized regression line* explicitly by using least squares. Following our usual conventions, we write the idealized line with Greek letters and consider the coefficients (the slope and intercept) to be *parameters:* β_0 is the intercept and β_1 is the slope. Corresponding to our fitted line of $\hat{y} = b_0 + b_1 x$, we write

$$\mu_y = \beta_0 + \beta_1 x.$$

Why μ_y instead of $\hat{y}$? Because this is a model. There is a distribution of *%Body Fat* for each *Waist* size. The model places the *means* of the distributions of *%Body Fat* for each *Waist* size on the same straight line.

NOTATION ALERT:

This time we used up only one Greek letter for two things. Lower-case Greek β (beta) is the natural choice to correspond to the b's in the regression equation. We used β before for the probability of a Type II error, but there's little chance of confusion here.

Of course, not all the individual y's are at these means. (In fact, the line will miss most—and quite possibly all—of the plotted points.) Some individuals lie above and some below the line, so, like all models, this one makes **errors.** Lots of them. In fact, one at each point. These errors are random and, of course, can be positive or negative. They are model errors, so we use a Greek letter and denote them by ε.

When we put the errors into the equation, we can account for each individual y:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

This equation is now true for each data point (since there is an ε to soak up the deviation), so the model gives a value of y for any value of x.

For the body fat data, an idealized model such as this provides a summary of the relationship between *%Body Fat* and *Waist* size. Like all models, it simplifies the real situation. We know there is more to predicting body fat than waist size alone. But the advantage of a model is that the simplification might help us to think about the situation and assess how well *%Body Fat* can be predicted from simpler measurements.

We estimate the β's by finding a regression line, $\hat{y} = b_0 + b_1 x$, as we did in Chapter 8. The residuals, $e = y - \hat{y}$, are the sample-based versions of the errors, ε. We'll use them to help us assess the regression model.

We know that least squares regression will give reasonable estimates of the parameters of this model from a random sample of data. Our challenge is to account for our uncertainty in how well they do. For that, we need to make some assumptions about the model and the errors.

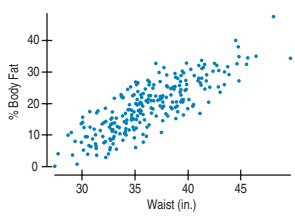
Assumptions and Conditions

A S ***Activity:* Conditions for Regression Inference.** View an illustrated discussion of the conditions for regression inference.

Back in Chapter 8 when we fit lines to data, we needed to check only the Straight Enough Condition. Now, when we want to make inferences about the coefficients of the line, we'll have to make more assumptions. Fortunately, we can check conditions to help us judge whether these assumptions are reasonable for our data. And as we've done before, we'll make some checks *after* we find the regression equation.

Also, we need to be careful about the order in which we check conditions. If our initial assumptions are not true, it makes no sense to check the later ones. So now we number the assumptions to keep them in order.

Check the scatterplot.
The shape must be linear or we can't use linear regression at all.



1. Linearity Assumption

If the true relationship is far from linear and we use a straight line to fit the data, our entire analysis will be useless, so we always check this first.

The **Straight Enough Condition** is satisfied if a scatterplot looks straight. It's generally not a good idea to draw a line through the scatterplot when checking. That can fool your eyes into seeing the plot as more straight. Sometimes it's easier to see violations of the Straight Enough Condition by looking at a scatterplot of the residuals against x or against the predicted values, $\hat{y}$. That plot will have a horizontal direction and should have no pattern if the condition is satisfied.

If the scatterplot is straight enough, we can go on to some assumptions about the errors. If not, stop here, or consider re-expressing the data (see Chapter 10) to make the scatterplot more nearly linear. For the *%Body Fat* data, the scatterplot is beautifully linear. Of course, the data must be quantitative for this to make sense. Check the **Quantitative Data Condition.**

Check the residuals plot (1).
The residuals should appear to be randomly scattered.

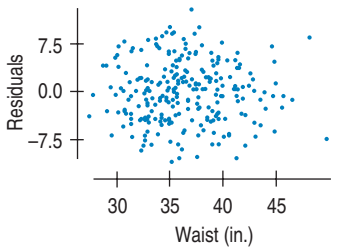


FIGURE 27.4
The residuals show only random scatter when plotted against Waist *size.*

2. Independence Assumption

Independence Assumption: The errors in the true underlying regression model (the ε's) must be mutually independent. As usual, there's no way to be sure that the Independence Assumption is true.

Usually when we care about inference for the regression parameters, it's because we think our regression model might apply to a larger population. In such cases, we can check a **Randomization Condition** that the individuals are a representative sample from that population.

We can also check displays of the regression residuals for evidence of patterns, trends, or clumping, any of which would suggest a failure of independence. In the special case when the x-variable is related to time, a common violation of the Independence Assumption is for the errors to be correlated. (The error our model makes today may be similar to the one it made for yesterday.) This violation can be checked by plotting the residuals against the x-variable and looking for patterns.

The *%Body Fat* data were collected on a sample of men taken to be representative. The subjects were not related in any way, so we can be pretty sure that their measurements are independent. The residuals plot shows no pattern.

Check the residuals plot (2).
The vertical spread of the residuals should be roughly the same everywhere.

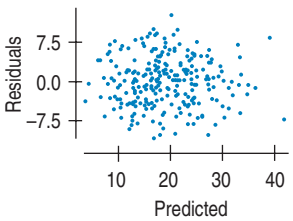


FIGURE 27.5
A scatterplot of residuals against predicted values can help check for plot thickening. Note that this plot looks identical to the plot of residuals against Waist *size. For a regression of one response variable on one predictor, these plots differ only in the labels on the x-axis.*

3. Equal Variance Assumption

The variability of y should be about the same for all values of x. In Chapter 8 we looked at the standard deviation of the residuals (s_e) to measure the size of the scatter. Now we'll need this standard deviation to build confidence intervals and test hypotheses. The standard deviation of the residuals is the building block for the standard errors of all the regression parameters. But it makes sense only if the scatter of the residuals is the same everywhere. In effect, the standard deviation of the residuals "pools" information across all of the individual distributions at each x-value, and pooled estimates are appropriate only when they combine information for groups with the same variance.

Practically, what we can check is the **Does the Plot Thicken? Condition.** A scatterplot of y against x offers a visual check. Fortunately, we've already made one. Make sure the spread around the line is nearly constant. Be alert for a "fan" shape or other tendency for the variation to grow or shrink in one part of the scatterplot. Often it is better to look at the residuals plotted against the predicted values, $\hat{y}$. With the slope of the line removed, it's easier to see patterns left behind. For the body fat data, the spread of *%Body Fat* around the line is remarkably constant across *Waist* sizes from 30 inches to about 45 inches.

If the plot is straight enough, the data are independent, and the plot doesn't thicken, you can now move on to the final assumption.

4. Normal Population Assumption

We assume the errors around the idealized regression line at each value of x follow a Normal model. We need this assumption so that we can use a Student's t-model for inference.

As we have at other times when we've used Student's t, we'll settle for the residuals satisfying the **Nearly Normal Condition** and the **Outlier Condition.** Look at a histogram or Normal probability plot of the residuals.[1]

[1] *This* is why we have to check the conditions in order. We have to check that the residuals are independent and that the variation is the same for all x's so that we can lump all the residuals together for a single check of the Nearly Normal Condition.

Check a histogram of the residuals.
The distribution of the residuals should be unimodal and symmetric.

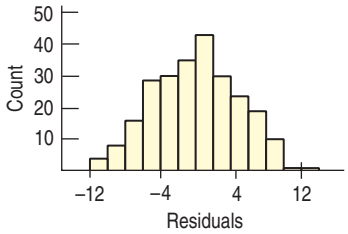


FIGURE 27.6
A histogram of the residuals is one way to check whether they are nearly Normal. Alternatively, we can look at a Normal probability plot.

The histogram of residuals in the *%Body Fat* regression certainly looks nearly Normal. As we have noted before, the Normality Assumption becomes less important as the sample size grows, because the model is about means and the Central Limit Theorem takes over.

If all four assumptions were true, the idealized regression model would look like this:

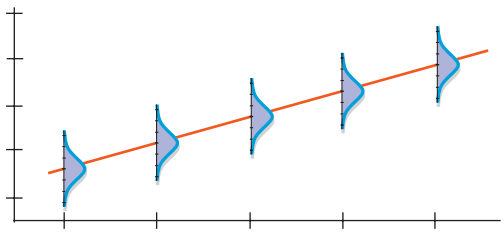


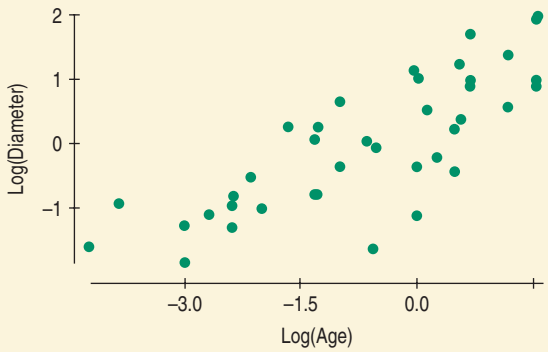
FIGURE 27.7
The regression model has a distribution of y-values for each x-value. These distributions follow a normal model with means lined up along the line and with the same standard deviations.

At each value of x there is a distribution of y-values that follows a Normal model, and each of these Normal models is centered on the line and has the same standard deviation. Of course, we don't expect the assumptions to be exactly true, and we know that all models are wrong, but the linear model is often close enough to be very useful.

FOR EXAMPLE Checking assumptions and conditions

Look at the moon with binoculars or a telescope, and you'll see craters formed by thousands of impacts. The earth, being larger, has been hit even more often. Meteor Crater in Arizona was the first recognized impact crater and was identified as such only in the 1920s. With the help of satellite images, more and more craters have been identified; now more than 180 are known. These, of course, are only a small sample of all the impacts the earth has experienced: Only 29% of earth's surface is land, and many craters have been covered or eroded away. Astronomers have recognized a roughly 35 million-year cycle in the frequency of cratering, although the cause of this cycle is not fully understood. Here's a scatterplot of the known impact craters from the most recent 35 million years.[2] We've taken logs of both age (in millions of years ago) and diameter (km) to make the relationship simpler. (See Chapter 10.)

WHO	39 impact craters
WHAT	Diameter and age
UNITS	km and millions of years ago
WHEN	Past 35 million years
WHERE	Worldwide
WHY	Scientific research



Question: Are the assumptions and conditions satisfied for fitting a linear regression model to these data?

✓ **Linearity Assumption:** The scatterplot satisfies the Straight Enough Condition.

✓ **Independence Assumption:** Sizes of impact craters are likely to be generally independent.

(*continued*)

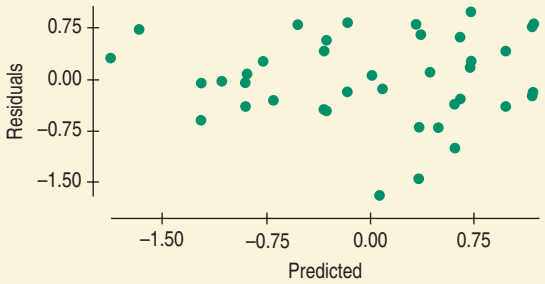
[2] Data, pictures, and much more information at the Earth Impact Database found at http://www.unb.ca.

For Example (*continued*)

✓ **Randomization Condition:** These are the only known craters, and may differ from others that have disappeared or not yet been found. I'll need to be careful not to generalize my conclusions too broadly.

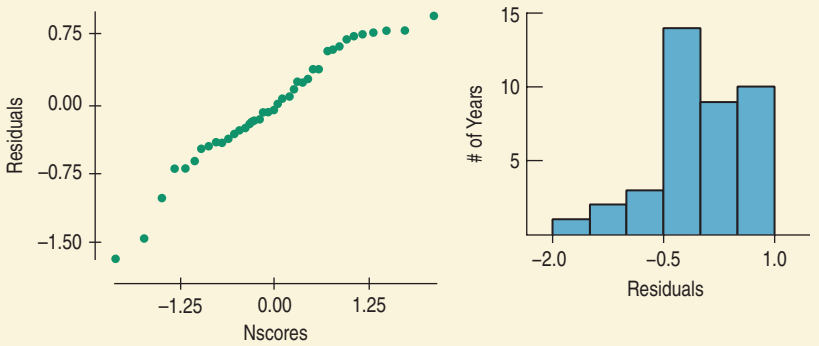
✓ **Does the Plot Thicken? Condition:** After fitting a linear model, I find the residuals shown.

Two points seem to give the impression that the residuals may be more variable for higher predicted values than for lower ones, but this doesn't seem to be a serious violation of the Equal Variance Assumption.



✓ **Nearly Normal Condition:** A Normal probability plot suggests a bit of skewness in the distribution of residuals, and the histogram confirms that.

There are no violations severe enough to stop my regression analysis, but I'll be cautious about my conclusions.



Which Come First: the Conditions or the Residuals?

"Truth will emerge more readily from error than from confusion."

—Francis Bacon (1561–1626)

In regression, there's a little catch. The best way to check many of the conditions is with the residuals, but we get the residuals only *after* we compute the regression. Before we compute the regression, however, we should check at least one of the conditions.

So we work in this order:

1. Make a scatterplot of the data to check the Straight Enough Condition. (If the relationship is curved, try re-expressing the data. Or stop.)
2. If the data are straight enough, fit a regression and find the residuals, e, and predicted values, $\hat{y}$.
3. Make a scatterplot of the residuals against x or the predicted values. This plot should have no pattern. Check in particular for any bend (which would suggest that the data weren't all that straight after all), for any thickening (or thinning), and, of course, for any outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions—one with and one without the outliers.)
4. If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not independent.
5. If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
6. If all the conditions seem to be reasonably satisfied, go ahead with inference.

STEP-BY-STEP EXAMPLE Regression Inference

If our data can jump through all these hoops, we're ready to do regression inference. Let's see how much more we can learn about body fat and waist size from a regression model.

Questions: What is the relationship between *%Body Fat* and *Waist* size in men?
What model best predicts body fat from waist size, and how well does it do the job?

THINK

Plan Specify the question of interest.

Name the variables and report the W's.

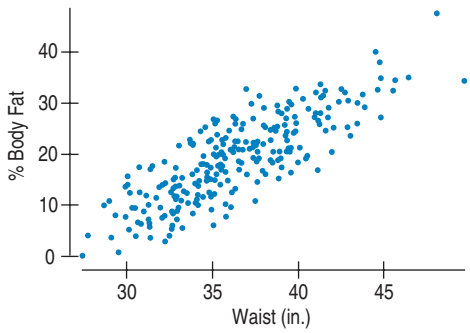
Identify the parameters you want to estimate.

Model Think about the assumptions and check the conditions.

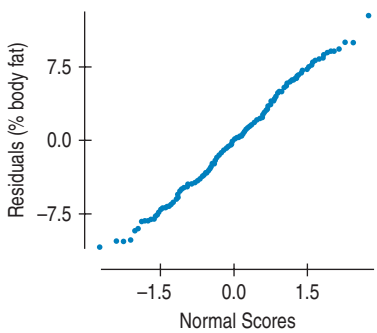
Make pictures. For regression inference, you'll need a scatterplot, a residuals plot, and either a histogram or a Normal probability plot of the residuals.

(We've seen plots of the residuals already. See Figures 27.5 and 27.6.)

I have quantitative body measurements on 250 adult males from the BYU Human Performance Research Center. I want to understand the relationship between *%Body Fat* and *Waist* size.



- ✓ **Straight Enough Condition:** There's no obvious bend in the original scatterplot of the data or in the plot of residuals against predicted values.
- ✓ **Independence Assumption:** These data are not collected over time, and there's no reason to think that the *%Body Fat* of one man influences the *%Body Fat* of another.
- ✓ **Does the Plot Thicken? Condition:** Neither the original scatterplot nor the residual scatterplot shows any changes in the spread about the line.
- ✓ **Nearly Normal Condition, Outlier Condition:** A histogram of the residuals is unimodal and symmetric. The Normal probability plot of the residuals is quite straight, indicating that the Normal model is reasonable for the errors.



Choose your method.

Under these conditions a **regression model** is appropriate.

SHOW

Mechanics Let's just "push the button" and see what the regression looks like.

The formula for the regression equation can be found in Chapter 8, and the standard error formulas will be shown a bit later, but regressions are almost always computed with a computer program or calculator.

Write the regression equation.

Here's the computer output for this regression:

Dependent variable is: %BF
R-squared = 67.8%
s = 4.713 with 250 − 2 = 248 degrees of freedom

Variable	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	−42.734	2.717	−15.7	<0.0001
Waist	1.70	0.0743	22.9	<0.0001

The estimated regression equation is

$$\widehat{\%\text{Body Fat}} = -42.73 + 1.70\ \text{Waist}.$$

TELL

Conclusion Interpret your results in context.

The R^2 for the regression is 67.8%. Waist size seems to account for about 2/3 of the %Body Fat variation in men. The slope of the regression says that %Body Fat increases by about 1.7 percentage points per inch of Waist size, on average.

More Interpretation We haven't worked it out in detail yet, but the output gives us numbers labeled as *t*-statistics and corresponding P-values, and we have a general idea of what those mean.

(Now it's time to learn more about regression inference so we can figure out what the rest of the output means.)

The standard error of 0.07 for the slope is much smaller than the slope itself, so it looks like the estimate is reasonably precise. And there are a couple of t-ratios and P-values given. Because the P-values are small, it appears that some null hypotheses can be rejected.

Intuition About Regression Inference

A S

***Simulation:* Simulate the Sampling Distribution of a Regression Slope.** Draw samples repeatedly to see for yourself how slope can vary from sample to sample. This simulation experiment lets you build up a histogram to see the sampling distribution.

Wait a minute! We've just pulled a fast one. We've pushed the "regression button" on our computer or calculator but haven't discussed where the standard errors for the slope or intercept come from. We know that if we had collected similar data on a different random sample of men, the slope and intercept would be different. Each sample would have produced its own regression line, with slightly different b_0's and b_1's. This sample-to-sample variation is what generates the sampling distributions for the coefficients.

There's only one regression model; each sample regression is trying to estimate the same parameters, β_0 and β_1. We expect any sample to produce a b_1 whose expected value is the true slope, β_1. What about its standard deviation? What aspects of the data affect how much the slope (and intercept) vary from sample to sample?

▶ **Spread around the line.** Here are two situations in which we might do regression. Which situation would yield the more consistent slope? That is, if we were to sample over and over from the two underlying populations that these samples come from and compute all the slopes, which group of slopes would vary less?

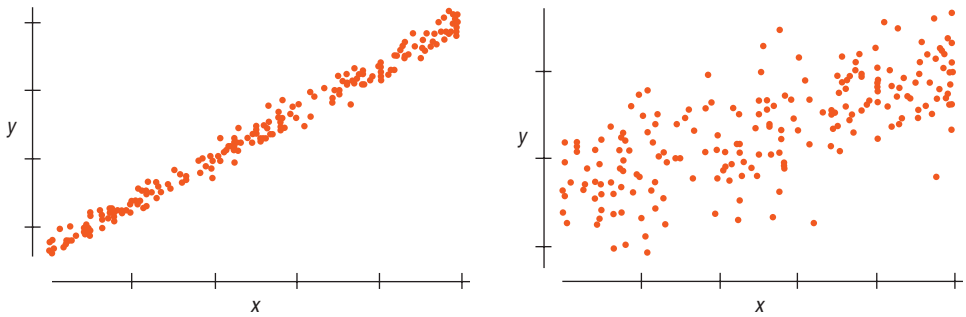


FIGURE 27.8
Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from its underlying population?

Clearly, data like those in the left plot give more consistent slopes.

Less scatter around the line means the slope will be more consistent from sample to sample. The spread around the line is measured with the **residual standard deviation,** s_e. You can always find s_e in the regression output, often just labeled s. You're probably not going to calculate the residual standard deviation by hand. As we noted when we first saw this formula in Chapter 8, it looks a lot like the standard deviation of y, only now subtracting the predicted values rather than the mean and dividing by $n - 2$ instead of $n - 1$:

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}.$$

The less scatter around the line, the smaller the residual standard deviation and the stronger the relationship between x and y.

Some people prefer to assess the strength of a regression by looking at s_e rather than R^2. After all, s_e has the same units as y, and because it's the standard deviation of the errors around the line, it tells you how close the data are to our model. By contrast, R^2 is the proportion of the variation of y accounted for by x. We say, why not look at both?

$n - 2$?

For standard deviation (in Chapter 4), we divided by $n - 1$ because we didn't know the true mean and had to estimate it. Now it's later in the course and there's even more we don't know. Here we don't know *two* things: the slope and the intercept. If we knew them both, we'd divide by n and have n degrees of freedom. When we estimate both, however, we adjust by subtracting 2, so we divide by $n - 2$ and (as we will see soon) have 2 fewer degrees of freedom.

▶ **Spread of the x's:** Here are two more situations. Which of these would yield more consistent slopes?

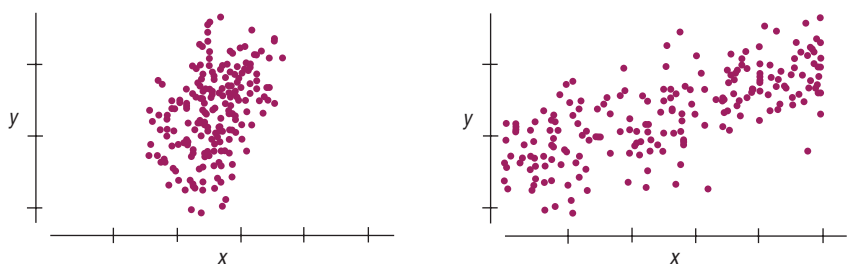


FIGURE 27.9
Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from the underlying population?

A plot like the one on the right has a broader range of x-values, so it gives a more stable base for the slope. We'd expect the slopes of samples from situations like that to vary less from sample to sample. A large standard deviation of x, s_x, provides a more stable regression.

- **Sample size.** Here we go again. What about these two?

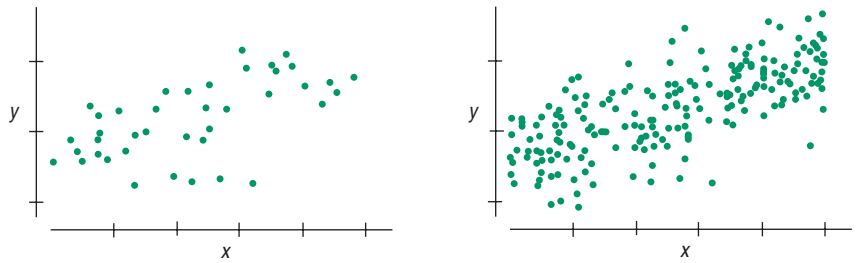


FIGURE 27.10
Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from the underlying population?

It shouldn't be a surprise that having a larger sample size, n, gives more consistent estimates from sample to sample.

Standard Error for the Slope

Three aspects of the scatterplot, then, affect the standard error of the regression slope:

- Spread around the line: s_e
- Spread of x values: s_x
- Sample size: n

These are in fact the *only* things that affect the standard error of the slope. Although you'll probably never have to calculate it by hand, the formula for the standard error is

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}.$$

The error standard deviation, s_e, is in the *numerator*, since spread around the line *increases* the slope's standard error. The denominator has both a sample size term $\sqrt{n-1}$ and s_x, because increasing either of these *decreases* the slope's standard error.

We know the b_1's vary from sample to sample. As you'd expect, their sampling distribution model is centered at β_1, the slope of the idealized regression line. Now we can estimate its standard deviation with $SE(b_1)$. What about its shape? Here the Central Limit Theorem and "Wild Bill" Gosset come to the rescue again. When we standardize the slopes by subtracting the model mean and dividing by their standard error, we get a Student's t-model, this time with $n - 2$ degrees of freedom:

$$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}.$$

A S ***Activity:* Regression Slope Standard Error.** See how $SE(b_1)$ is constructed and where the values used in the formula are found in the regression output table.

A S ***Simulation:* x-Variance and Slope Variance.** You don't have to just imagine how the variability of the slope depends on the spread of the x's.

NOTATION ALERT:

Don't confuse the standard deviation of the residuals, s_e, with the standard error of the slope, $SE(b_1)$. The first measures the scatter around the line, and the second tells us how reliably we can estimate the slope.

A SAMPLING DISTRIBUTION FOR REGRESSION SLOPES

When the conditions are met, the standardized estimated regression slope,

$$t = \frac{b_1 - \beta_1}{SE(b_1)},$$

follows a Student's t-model with $n - 2$ degrees of freedom. We estimate the standard error with

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\,s_x}, \text{ where } s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}},$$

n is the number of data values, and s_x is the ordinary standard deviation of the x-values.

FOR EXAMPLE Finding standard errors

Recap: Recent terrestrial impact craters seem to show a relationship between age and size that is linear when re-expressed using logarithms (see Chapter 10).

Here are summary statistics and regression output.

Variable	Count	Mean	StdDev
LogAge	39	−0.656310	1.57682
LogDiam	39	0.012600	1.04104

Dependent variable is: LogDiam
R-squared = 63.6%
s = 0.6362 with 39 − 2 = 37 degrees of freedom

Variable	Coefficient	Se(coeff)	t-ratio	P-value
Intercept	0.358262	0.1106	3.24	0.0025
LogAge	0.526674	0.0655	8.05	⩽ 0.0001

Questions: How are the standard error of the slope and the t-ratio for the slope calculated? (And aren't you glad the software does this for you?)

$$SE(b_1) = \frac{s_e}{\sqrt{n-1} \times s_x} = \frac{0.6362}{\sqrt{39-1} \times 1.57682} = 0.0655$$

Assuming no linear association ($\beta_1 = 0$), $t_{37} = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{0.526674 - 0}{0.0655} = 8.05$

What About the Intercept?

The same reasoning applies for the intercept. We could write

$$\frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$$

and use it to construct confidence intervals and test hypotheses, but often the value of the intercept isn't something we care about. The intercept usually isn't interesting. Most hypothesis tests and confidence intervals for regression are about the slope.

Regression Inference

TI-*nspire*

Regression Inference. How big must a slope be in order to be considered statistically significant? See for yourself by exploring the natural sample-to-sample variability in slopes.

Now that we have the standard error of the slope and its sampling distribution, we can test a hypothesis about it and make confidence intervals. The usual null hypothesis about the slope is that it's equal to 0. Why? Well, a slope of zero would say that y doesn't tend to change linearly when x changes—in other words, that there is no linear association between the two variables. If the slope were zero, there wouldn't be much left of our regression equation.

So a null hypothesis of a zero slope questions the entire claim of a linear relationship between the two variables—and often that's just what we want to know. In fact, every software package or calculator that does regression simply assumes that you want to test the null hypothesis that the slope is really zero.

What if the Slope Were 0?

If $b_1 = 0$, our prediction is $\hat{y} = b_0 + 0x$. The equation collapses to just $\hat{y} = b_0$. Now x is nowhere in sight, so y doesn't depend on x at all.

And b_0 would turn out to be $\bar{y}$. Why? We know that $b_0 = \bar{y} - b_1\bar{x}$, but when $b_1 = 0$, that becomes simply $b_0 = \bar{y}$. It turns out, then, that when the slope is 0, the equation is just $\hat{y} = \bar{y}$; at every value of x, we always predict the mean value for y.

To test H_0: $\beta_1 = 0$, we find

$$t_{n-2} = \frac{b_1 - 0}{SE(b_1)}.$$

This is just like every t-test we've seen: a difference between the statistic and its hypothesized value, divided by its standard error.

For our body fat data, the computer found the slope (1.7), its standard error (0.0743), and the ratio of the two: $\frac{1.7 - 0}{0.0743} = 22.9$ (see p. 656). Nearly 23 standard errors from the hypothesized value certainly seems big. The P-value (<0.0001) confirms that a t-ratio this large would be very unlikely to occur if the true slope were zero.

Maybe the standard null hypothesis isn't all that interesting here. Did you have any doubts that *%Body Fat* is related to *Waist* size? A more sensible use of these same values might be to make a confidence interval for the slope instead.

We can build a confidence interval in the usual way, as an estimate plus or minus a margin of error. As always, the margin of error is just the product of the standard error and a critical value. Here the critical value comes from the t-distribution with $n - 2$ degrees of freedom, so a 95% **confidence interval for β** is

$$b_1 \pm t^*_{n-2} \times SE(b_1).$$

For the body fat data, $t^*_{248} = 1.970$, so that comes to $1.7 \pm 1.97 \times 0.074$, or an interval from 1.55 to 1.85 *%Body Fat* per inch of *Waist* size.

FOR EXAMPLE — Interpreting a regression model

Recap: On a log scale, there seems to be a linear relationship between the diameter and the age of recent terrestrial impact craters. We have regression output from statistics software:

Dependent variable is: LogDiam
R-squared = 63.6%
s = 0.6362 with 39 − 2 = 37 degrees of freedom

Variable	Coefficient	Se(coeff)	t-ratio	P-value
Intercept	0.358262	0.1106	3.24	0.0025
LogAge	0.526674	0.0655	8.05	≤0.0001

Questions: What's the regression model, and what can it tell us?

For terrestrial impact craters younger than 35 million years, the logarithm of *Diameter* grows linearly with the logarithm of *Age*: $\widehat{\log Diam} = 0.358 + 0.527 \log Age$. The P-value for each coefficient's t-statistic is very small, so I'm quite confident that neither coefficient is zero. Based on my model, I conclude that, on average, the older a crater is, the larger it tends to be. This model accounts for 63.6% of the variation in *logDiam*.

Although it is possible that impacts (and their craters) are getting smaller, it is more likely that i'm seeing the effects of age on craters. Small craters are probably more likely to erode or become buried or otherwise be difficult to find as they age. Larger craters may survive the huge expanses of geologic time more successfully.

JUST CHECKING

Researchers in Food Science studied how big people's mouths tend to be. They measured mouth volume by pouring water into the mouths of subjects who lay on their backs. Unless this is your idea of a good time, it would be helpful to have a model to estimate mouth volume more simply. Fortunately, mouth volume is related to height. (Mouth volume is measured in cubic centimeters and height in meters.)

The data were checked and deemed suitable for regression. Take a look at the computer output.

1. What does the *t*-ratio of 3.27 tell us about this relationship? How does the P-value help our understanding?
2. Would you say that measuring a person's height could reliably be used as a substitute for the wetter method of determining how big a person's mouth is? What numbers in the output helped you reach that conclusion?
3. What does the value of s_e add to this discussion?

Summary of Mouth Volume

Mean	60.2704
StdDev	16.8777

Dependent variable is: Mouth Volume
R-squared = 15.3%
s = 15.66 with 61 − 2 = 59 degrees of freedom

Variable	Coefficient	SE(coeff)	t-ratio	P-value
Intercept	−44.7113	32.16	−1.39	0.1697
Height	61.3787	18.77	3.27	0.0018

Another Example

A S ***Activity:*** **A Hypothesis Test for the Regression Slope.** View an animated discussion of testing the standard null hypothesis for slope.

Every spring, Nenana, Alaska, hosts a contest in which participants try to guess the exact minute that a wooden tripod placed on the frozen Tanana River will fall through the breaking ice. The contest started in 1917 as a diversion for railroad engineers, with a jackpot of $800 for the closest guess. It has grown into an event in which hundreds of thousands of entrants enter their guesses on the Internet[3] and vie for as much as $300,000.

Because so much money and interest depends on the time of breakup, it has been recorded to the nearest minute with great accuracy ever since 1917. And because a standard measure of breakup has been used throughout this time, the data are consistent. An article in *Science*[4] used the data to investigate global warming—whether greenhouse gasses and other human actions have been making the planet warmer. Others might just want to make a good prediction of next year's breakup time.

Of course, we can't use regression to tell the *causes* of any change. But we can estimate the *rate* of change (if any) and use it to make better predictions.

Here are some of the data:

WHO	Years
WHAT	Year, day, and hour of ice breakup
UNITS	*x* is in years since 1900. *y* is in days after midnight Dec. 31.
WHEN	1917–present
WHERE	Nenana, Alaska
WHY	Wagering, but proposed to look at global warming

Year (since 1900)	Breakup Date (days after Jan. 1)	Year (since 1900)	Breakup Date (days after Jan. 1)
17	119.4792	30	127.7938
18	130.3979	31	129.3910
19	122.6063	32	121.4271
20	131.4479	33	127.8125
21	130.2792	34	119.5882
22	131.5556	35	134.5639
23	128.0833	36	120.5403
24	131.6319	37	131.8361
25	126.7722	38	125.8431
26	115.6688	39	118.5597
27	131.2375	40	110.6437
28	126.6840	41	122.0764
29	124.6535	⋮	⋮

[3] http://www.nenanaakiceclassic.com
[4] "Climate Change in Nontraditional Data Sets." *Science* 294 [26 October 2001]: 811.

STEP-BY-STEP EXAMPLE A Regression Slope *t*-Test

The slope of the regression gives the change in Nenana ice breakup date per year.

Questions: Is there sufficient evidence to claim that ice breakup times are changing? If so, how rapid is the change?

THINK

Plan State what you want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the slope.

Identify the variables and review the W's.

I wonder whether the date of ice breakup in Nenana has changed over time. The slope of that change might indicate climate change. I have the date of ice breakup annually since 1917, recorded as the number of days and fractions of a day until the ice breakup.

Hypotheses Write your null and alternative hypotheses.

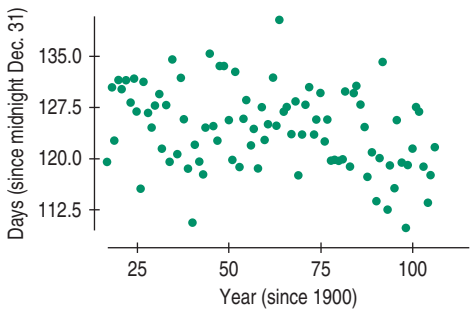
H_0: There is no change in the date of ice breakup: $\beta_1 = 0$

H_A: Yes, there is: $\beta_1 \neq 0$

Model Think about the assumptions and check the conditions.

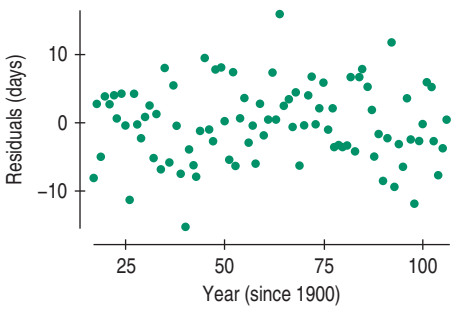
✓ **Straight Enough Condition:** I have quantitative data with no obvious bend in the scatterplot.

Make pictures. Because the scatterplot seems straight enough, we can find and plot the residuals.



✓ **Independence Assumption:** These data are a time series, which raises my suspicions that they may not be independent. To check, here's a plot of the residuals against time, the x-variable of the regression:

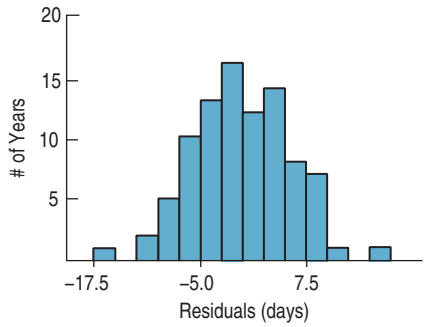
Usually, we check for suggestions that the Independence Assumption fails by plotting the residuals against the predicted values. Patterns and clusters in that plot raise our suspicions. But when the data are measured over time, it is always a good idea to plot residuals against time to look for trends and oscillations.



I see a hint that the data oscillate up and down, which suggests some failure of independence, but not so strongly that I can't

proceed with the analysis. These data are not a random sample, so I'm reluctant to extend my conclusions beyond this river and these years.

- ✓ **Does the Plot Thicken? Condition:** The residuals plot shows no obvious trends in the spread.
- ✓ **Nearly Normal Condition, Outlier Condition:** A histogram of the residuals is unimodal and symmetric.



State the sampling distribution model.

Under these conditions, the sampling distribution of the regression slope can be modeled by a Student's t-model with $(n - 2) = 89$ degrees of freedom.

Choose your method.

I'll do a **regression slope t-test.**

SHOW

Mechanics The regression equation can be found from the formulas in Chapter 8, but regressions are almost always found from a computer program or calculator.

The P-values given in the regression output table are from the Student's t-distribution on $(n - 2) = 89$ degrees of freedom. They are appropriate for two-sided alternatives.

Here's the computer output for this regression:

Dependent variable is: Breakup Date

R-squared = 11.3%
s = 5.673 with 91 − 2 = 89 degrees of freedom

Variable	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	128.950	1.525	84.6	<0.0001
Year Since 1900	−0.07606	0.0226	−3.36	0.0012

The estimated regression equation is

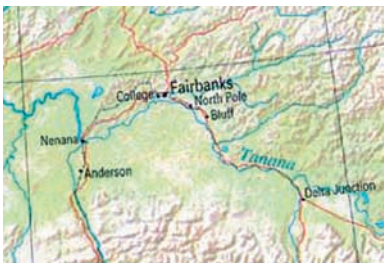
$$\widehat{\text{Date}} = 128.95 - 0.076\ \text{YearSince1900}.$$

TELL

Conclusion Link the P-value to your decision and state your conclusion in the proper context.

The P-value of 0.0012 means that the association we see in the data is unlikely to have occurred by chance. I reject the null hypothesis, and conclude that there is strong evidence that, on average, the ice breakup is occurring earlier each year. But the oscillation pattern in the residuals raises concerns.

SHOW MORE **Create a confidence interval for the true slope**	A 95% confidence interval for β_1 is $b_1 \pm t^*_{89} \times SE(b_1)$ $-0.076 \pm (1.987)(0.0226)$ or $(-0.12, -0.03)$ days per year.
TELL MORE **Interpret the interval** Simply rejecting the standard null hypothesis doesn't guarantee that the size of the effect is large enough to be important. Whether we want to know the breakup time to the nearest minute or are interested in global warming, a change measured in hours each year is big enough to be interesting.	I am 95% confident that the ice has been breaking up, on average, between 0.03 days (about 40 minutes) and 0.12 days (about 3 hours) earlier each year since 1900.



But is it global warming? So the ice is breaking up earlier. Temperatures are higher. Must be global warming, right?

Maybe.

An article challenging the original analysis of the Nenana data proposed a possible confounding variable. It noted that the city of Fairbanks is upstream from Nenana and suggested that the growth of Fairbanks could have warmed the river. So maybe it's not global warming.

Or maybe global warming is a lurking variable, leading more people to move to a now balmier Fairbanks and also leading to generally earlier ice breakup in Nenana.

Or maybe there's some other variable or combination of variables at work. We can't set up an experiment, so we may never really know.

Only one thing is for sure. When you try to explain an association by claiming cause and effect, you're bound to be on thin ice.[5]

TI Tips

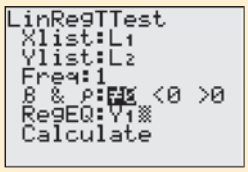
Doing regression inference

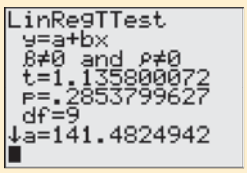
°F	Min
44	142.7
46	142.1
47	143.4
50	143.6
51	144.0
52	143.4
54	142.4
55	143.1
57	143.7
60	143.4
65	143.4

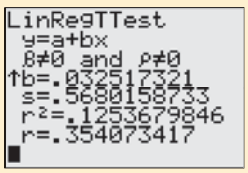
The TI will easily do almost everything you need for inference for regression: scatterplots, residual plots, histograms of residuals, and *t*-tests and confidence intervals for the slope of the regression line. OK, it won't tell you *SE(b)*, but it will give you enough information to easily figure it out for yourself. Not bad.

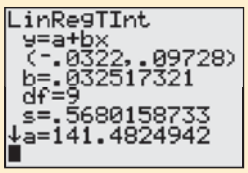
As an example we'll use data from *Chance* magazine (Vol. 12, No. 4, 1999), giving times and temperatures for 11 of the top performances in women's marathons during the 1990s. Let's examine the influence of temperature on the performance of elite runners in marathons.

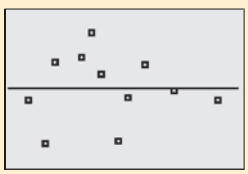
[5] How *do* scientists sort out such messy situations? Even though they can't conduct an experiment, they *can* look for replications elsewhere. A number of studies of ice on other bodies of water have also shown earlier ice breakup times in recent years. That suggests they need an explanation that's more comprehensive than just Fairbanks and Nenana.

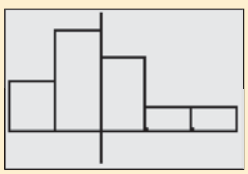












Test a Hypothesis About the Association

- Enter the temperatures (nearest degree Fahrenheit) in L1 and the runners' times (nearest tenth of a minute) in L2.
- Check the scatterplot. It's not obviously nonlinear, so go ahead.
- Under STAT TESTS choose LinRegTTest.
- Specify the two data lists (with Freq: 1).
- Choose the two-tailed option. (We are interested in whether higher temperatures enhance or interfere with a runner's performance.)
- Tell it to store the regression equation in Y1 (VARS, Y-VARS, Function . . . remember?), then Calculate.

The TI creates so much information you have to scroll down to see it all! Look what's there.

- The calculated value of t and the P-value.
- The coefficients of the regression equation, a and b.
- The value of s, our sample estimate of the common standard deviation of errors around the true line.
- The values of r^2 and r.

Wait, where's *SE(b)*? It's not there. No problem—if you need it, you can figure it out. Remember that the *t*-value is *b* divided by *SE(b)*. So *SE(b)* must be *b* divided by *t*. Here *SE(b)* = 0.0325 ÷ 1.1358 = 0.0286.

Create a Confidence Interval for the Slope

- Back to STAT TEST; this time you want LinRegTInt.
- The specifications for the data lists and the regression equation remain what you entered for the hypothesis test.
- Choose a confidence level, say 95%, and Calculate.

Checking Conditions

Beware!!! Before you try to interpret any of this, you must check the conditions to see if inference for regression is allowed.

- We already looked at the scatterplot; it was reasonably linear.
- To create the residuals plot, set up another scatterplot with RESID (from LIST NAMES) as your Ylist. OK, it looks fairly random.
- The residuals plot may show a slight hint of diminishing scatter, but with so few data values it's not very clear.
- The histogram of the residuals is unimodal and roughly symmetric.

What Does It All Mean?

Because the conditions check out okay, we can try to summarize what we have learned. With a P-value over 28%, it's quite possible that any perceived relationship could be just sampling error. The confidence interval suggests the slope could be positive or negative, so it's possible that as temperatures increase, women marathoners may run faster—or slower. Based on these 11 races there appears to be little evidence of a linear association between temperature and women's performances in the marathon.

*Standard Errors for Predicted Values

Once we have a useful regression, how can we indulge our natural desire to predict, without being irresponsible? We know how to compute predicted values of *y* for any value of *x*. We first did that in Chapter 8. This predicted value would be our best estimate, but it's still just an informed guess.

Now, however, we have standard errors. We can use those to construct a confidence interval for the predictions and to report our uncertainty honestly.

From our model of *%Body Fat* and *Waist* size, we might want to use *Waist* size to get a reasonable estimate of *%Body Fat*. A confidence interval can tell us how precise that prediction will be. The precision depends on the question we ask, however, and there are two questions: Do we want to know the mean *%Body Fat* for *all* men with a *Waist* size of, say, 38 inches? Or do we want to estimate the *%Body Fat* for a particular man with a 38-inch *Waist* without making him climb onto the X-ray table?

What's the difference between the two questions? The predicted *%Body Fat* is the same, but one question leads to an answer much more precise than the other. We can predict the *mean %Body Fat* for *all* men whose *Waist* size is 38 inches with a lot more precision than we can predict the *%Body Fat* of a *particular individual* whose *Waist* size happens to be 38 inches. Both are interesting questions.

We start with the same prediction in both cases. We are predicting the value for a new individual, one that was not part of the original data set. To emphasize this, we'll call his x-value "x sub new" and write it x_ν.[6] Here, x_ν is 38 inches. The regression equation predicts *%Body Fat* as $\hat{y}_\nu = b_0 + b_1 x_\nu$.

Now that we have the predicted value, we construct both intervals around this same number. Both intervals take the form

$$\hat{y}_\nu \pm t^*_{n-2} \times SE.$$

Even the t^* value is the same for both. It's the critical value (from Table T or technology) for $n - 2$ degrees of freedom and the specified confidence level. The intervals differ because they have different standard errors. Our choice of ruler depends on which interval we want.

The standard errors for prediction depend on the same kinds of things as the coefficients' standard errors. If there is more spread around the line, we'll be less certain when we try to predict the response. Of course, if we're less certain of the slope, we'll be less certain of our prediction. If we have more data, our estimate will be more precise. And there's one more piece: If we're farther from the center of our data, our prediction will be less precise. This last factor is new but makes intuitive sense: It's a lot easier to predict a data point near the middle of the data set than far from the center.

Each of these factors contributes uncertainty—that is, variability—to the estimate. Because the factors are independent of each other, we can add their variances to find the total variability. The resulting formula for the standard error of the predicted *mean* value explicitly takes into account each of the factors:

$$SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}.$$

For the Nenana Ice Classic, someone who planned to place a bet would want to predict this year's breakup time. By contrast, scientists studying global warming are likely to be more interested in the mean breakup time. Unfortunately if you want to gamble, the variability is greater for predicting for a single year.

Individual values vary more than means, so the standard error for a single predicted value has to be larger than the standard error for the mean. In fact, the standard error of a single predicted value has an *extra* source of variability: the variation of individuals around the predicted mean. That appears as the extra variance term, s_e^2, at the end under the square root:

$$SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}.$$

[6] Yes, this is a bilingual pun. The Greek letter ν is called "nu." Don't blame me; my co-author suggested this.

Keep in mind this distinction between the two kinds of confidence intervals: The narrower interval is a **confidence interval for the predicted mean value** at x_y, and the wider interval is a **prediction interval for an individual** with that x-value.

FOR EXAMPLE *Finding confidence intervals for predicted values

Let's use our analysis to create confidence intervals for predictions about %*Body Fat*. From the data and the regression output we know:

$$n = 250 \quad \bar{x} = 36.3 \quad s_e = 4.713 \quad SE(b_1) = 0.074$$

Question 1: What's a 95% confidence interval for the mean %*Body Fat* for all men with 38-inch waists?

For $x_y = 38$ the regression model predicts

$$\hat{y}_y = -42.7 + 1.7(38) = 21.9\%.$$

The standard error is

$$SE(\hat{\mu}_y) = \sqrt{0.074^2(38 - 36.3)^2 + \frac{4.713^2}{250}} = 0.32\%.$$

With $250 - 2 = 248$ df, for 95% confidence $t^* = 1.97$.

Putting it all together, the 95% *confidence interval* is: $21.9\% \pm 1.97(0.32)$

$$21.9\% \pm 0.63\%, \text{ or } (21.27, 22.53)$$

I'm 95% confidence that the mean body fat level for all men with 38-inch waists is between 21.3% and 22.5% body fat.

Question 2: What's a 95% prediction interval for the %*Body Fat* of an individual man with a 38-inch waist?

The standard error is

$$SE(\hat{y}_y) = \sqrt{0.074^2(38 - 36.3)^2 + \frac{4.713^2}{250} + 4.713^2} = 4.72\%.$$

The *prediction interval* is: $21.9\% \pm 1.97(4.72)$

$$21.9\% \pm 9.3\%, \text{ or } (12.6, 31.2)$$

I'm 95% confident that a randomly selected man with a 38-inch waist will have between 12.6% and 31.2% body fat.

Notice how much wider this interval is than the first one. As we've known since Chapter 18, the mean is such less variable than a randomly selected individual value.

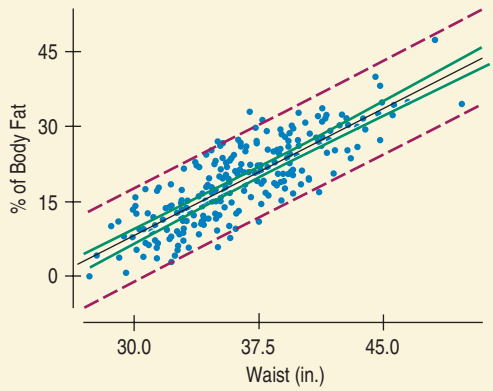


FIGURE 27.11

A scatterplot of %Body Fat *vs.* Waist *size with a least squares regression line. The solid green lines near the regression line show the extent of the 95% confidence intervals for* mean %*Body Fat at each* Waist *size. The dashed red lines show the prediction intervals. Most of the points are contained within the prediction intervals, but not within the confidence intervals.*

*MATH BOX

So where do those messy formulas for standard errors of predicted values come from? They're based on many of the ideas we've studied so far. Start with regression, add random variables, then throw in the Pythagorean Theorem, the Central Limit Theorem, and a dose of algebra. Mix well. . . .

We begin our quest with an equation of the regression line. Usually we write the line in the form $\hat{y} = b_0 + b_1 x$. Mathematicians call that the "slope-intercept" form; in your algebra class you wrote it as $y = mx + b$. In that algebra class you also learned another way to write equations of lines. When you know that a line with slope m passes through the point (x_1, y_1), the "point-slope" form of its equation is $y - y_1 = m(x - x_1)$.

We know the regression line passes through the mean-mean point $(\bar{x}, \bar{y})$ with slope b_1, so we can write its equation in point-slope form as $\hat{y} - \bar{y} = b_1(x - \bar{x})$. Solving for $\hat{y}$ yields $\hat{y} = b_1(x - \bar{x}) + \bar{y}$. This equation predicts the mean y-value for a specific x_v:

$$\hat{\mu}_y = b_1(x_v - \bar{x}) + \bar{y}.$$

To create a confidence interval for the mean value we need to measure the variability in this prediction:

$$Var(\hat{\mu}_y) = Var(b_1(x_v - \bar{x}) + \bar{y}).$$

We now call on the Pythagorean Theorem of Statistics once more: the slope, b_1, and mean, $\bar{y}$, should be independent, so their variances add:

$$Var(\hat{\mu}_y) = Var(b_1(x_v - \bar{x})) + Var(\bar{y}).$$

The horizontal distance from our specific x-value to the mean, $x_v - \bar{x}$, is a constant:

$$Var(\hat{\mu}_y) = (Var(b_1))(x_v - \bar{x})^2 + Var(\bar{y}).$$

Let's write that equation in terms of standard deviations:

$$SD(\hat{\mu}_y) = \sqrt{(SD^2(b_1))(x_v - \bar{x})^2 + SD^2(\bar{y})}.$$

Because we'll need to estimate these standard deviations using samples statistics, we're really dealing with standard errors:

$$SE(\hat{\mu}_y) = \sqrt{(SE^2(b_1))(x_v - \bar{x})^2 + SE^2(\bar{y})}.$$

The Central Limit Theorem tells us that the standard deviation of $\bar{y}$ is $\frac{\sigma}{\sqrt{n}}$. Here we'll estimate σ using s_e, which describes the variability in how far the line we drew through our sample mean may lie above or below the true mean:

$$\begin{aligned} SE(\hat{\mu}_y) &= \sqrt{(SE^2(b_1))(x_v - \bar{x})^2 + \left(\frac{s_e}{\sqrt{n}}\right)^2} \\ &= \sqrt{(SE^2(b_1))(x_v - \bar{x})^2 + \frac{s_e^2}{n}}. \end{aligned}$$

And there it is—the standard error we need to create a confidence interval for a predicted mean value.

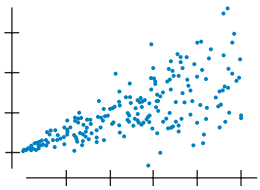
When we try to predict an individual value of y, we must also worry about how far the true point may lie above or below the regression line. We represent that uncertainty by adding another term, e, to the original equation:

$$y = b_1(x_v - \bar{x}) + \bar{y} + e.$$

To make a long story short (and the equation a wee bit longer), that additional term simply adds one more standard error to the sum of the variances:

$$SE(\hat{y}) = \sqrt{(SE^2(b_1))(x_v - \bar{x}) + \frac{s_e^2}{n} + s_e^2}.$$

WHAT CAN GO WRONG?



In this chapter we've added inference to the regression explorations that we did in Chapters 8 and 9. Everything covered in those chapters that could go wrong with regression can still go wrong. It's probably a good time to review Chapter 9. Take your time; we'll wait.

With inference, we've put numbers on our estimates and predictions, but these numbers are only as good as the model. Here are the main things to watch out for:

- **Don't fit a linear regression to data that aren't straight.** This is the most fundamental assumption. If the relationship between x and y isn't approximately linear, there's no sense in fitting a straight line to it.
- **Watch out for the plot thickening.** The common part of confidence and prediction intervals is the estimate of the error standard deviation, the spread around the line. If it changes with x, the estimate won't make sense. Imagine making a prediction interval for these data.

  When x is small, we can predict y precisely, but as x gets larger, it's much harder to pin y down. Unfortunately, if the spread changes, the single value of s_e won't pick that up. The prediction interval will use the average spread around the line, with the result that we'll be too pessimistic about our precision for low x-values and too optimistic for high x-values. A re-expression of y is often a good fix for changing spread.
- **Make sure the errors are Normal.** When we make a prediction interval for an individual, the Central Limit Theorem can't come to our rescue. For us to believe the prediction interval, the errors must be from the Normal model. Check the histogram and Normal probability plot of the residuals to see if this assumption looks reasonable.
- **Watch out for extrapolation.** It's tempting to think that because we have prediction *intervals*, they'll take care of all our uncertainty so we don't have to worry about extrapolating. Wrong. The interval is only as good as the model. The uncertainty our intervals predict is correct only if our model is true. There's no way to adjust for wrong models. That's why it's always dangerous to predict for x-values that lie far from the center of the data.
- **Watch out for influential points and outliers.** We always have to be on the lookout for a few points that have undue influence on our estimated model—and regression is certainly no exception.
- **Watch out for one-tailed tests.** Because tests of hypotheses about regression coefficients are usually two-tailed, software packages report two-tailed P-values. If you are using software to conduct a one-tailed test about slope, you'll need to divide the reported P-value in half.

CONNECTIONS

Regression inference is connected to almost everything we've done so far. Scatterplots are essential for checking linearity and whether the plot thickens. Histograms and normal probability plots come into play to check the Nearly Normal condition. And we're still thinking about the same attributes of the data in these plots as we were back in the first part of the book.

Regression inference is also connected to just about every inference method we have seen for measured data. The assumption that the spread of data about the line is constant is essentially the same as the assumption of equal variances required for the pooled-t methods. Our use of all the residuals together to estimate their standard deviation is a form of pooling.

Inference for regression is closely related to inference for means, so your understanding of means transfers directly to your understanding of regression. Here's a table that displays the similarities:

	Means	Regression Slope
Parameter	μ	β_1
Statistic	$\bar{y}$	b_1
Population spread estimate	$s_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$	$s_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$
Standard error of the statistic	$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$	$SE(b_1) = \frac{s_e}{s_x\sqrt{n-1}}$
Test statistic	$\frac{\bar{y}-\mu_0}{SE(\bar{y})} \sim t_{n-1}$	$\frac{b_1-\beta_1}{SE(b_1)} \sim t_{n-2}$
Margin of error	$ME = t^*_{n-1} \times SE(\bar{y})$	$ME = t^*_{n-2} \times SE(b_1)$

WHAT HAVE WE LEARNED?

In Chapters 7, 8, and 9, we learned to examine the relationship between two quantitative variables in a scatterplot, to summarize its strength with correlation, and to fit linear relationships by least squares regression. And we saw that these methods are particularly powerful and effective for modeling, predicting, and understanding these relationships.

Now we have completed our study of inference methods by applying them to these regression models. We've found that the same methods we used for means—Student's t-models—work for regression in much the same way as they did for means. And we've seen that although this makes the mechanics familiar, there are new conditions to check and a need for care in describing the hypotheses we test and the confidence intervals we construct.

- We've learned that under certain assumptions, the sampling distribution for the slope of a regression line can be modeled by a Student's t-model with $n - 2$ degrees of freedom.
- We've learned to check four conditions to verify those assumptions before we proceed with inference. We've learned the importance of checking these conditions in order, and we've seen that most of the checks can be made by graphing the data and the residuals with the methods we learned in Chapters 4, 5, and 8.
- We've learned to use the appropriate t-model to test a hypothesis about the slope. If the slope of our regression line is significantly different from zero, we have strong evidence that there is an association between the two variables.
- We've also learned to create and interpret a confidence interval for the true slope.
- And we've been reminded yet again never to mistake the presence of an association for proof of causation.

Terms

Conditions for inference in regression (and checks for some of them)

- 651. **Straight Enough Condition** for linearity. (Check that the scatterplot of y against x has linear form and that the scatterplot of residuals against predicted values has no obvious pattern.)
- 652. **Independence Assumption.** (Think about the nature of the data. Check a residuals plot.)
- 652. **Does the Plot Thicken? Condition** for constant variance. (Check that the scatterplot shows consistent spread across the range of the x-variable, and that the residuals plot has constant variance, too. A common problem is increasing spread with increasing predicted values—the *plot thickens!*)

- 652. **Nearly Normal Condition** for Normality of the residuals. (Check a histogram of the residuals.)

Residual standard deviation

657. The spread of the data around the regression line is measured with the residual standard deviation, s_e:

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\sum e^2}{n - 2}}.$$

t-test for the regression slope

658, 662. When the assumptions are satisfied, we can perform a test for the slope coefficient. We usually test the null hypothesis that the true value of the slope is zero against the alternative that it is not. A zero slope would indicate a complete absence of linear relationship between y and x.

To test H_0: $\beta_1 = 0$, we find

$$t = \frac{b_1 - 0}{SE(b_1)}$$

where

$$SE(b_1) = \frac{s_e}{\sqrt{n - 1}\, s_x}, \quad s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}},$$

n is the number of cases, and s_x is the standard deviation of the x-values. We find the P-value from the Student's t-model with $n - 2$ degrees of freedom.

Confidence interval for the regression slope (β)

660. When the assumptions are satisfied, we can find a confidence interval for the slope parameter from $b_1 \pm t^*_{n-2} \times SE(b_1)$. The critical value, t^*_{n-2}, depends on the confidence level specified and on Student's t-model with $n - 2$ degrees of freedom.

Skills

THINK

- Understand that the "true" regression line does not fit the population data perfectly, but rather is an idealized summary of that data.
- Know how to examine your data and a scatterplot of y vs. x for violations of assumptions that would make inference for regression unwise or invalid.
- Know how to examine displays of the residuals from a regression to double-check that the conditions required for regression have been met. In particular, know how to judge linearity and constant variance from a scatterplot of residuals against predicted values. Know how to judge Normality from a histogram and Normal probability plot.
- Remember to be especially careful to check for failures of the Independence Assumption when working with data recorded over time. To search for patterns, examine scatterplots both of x against time and of the residuals against time.

SHOW

- Know how to test the standard hypothesis that the true regression slope is zero. Be able to state the null and alternative hypotheses. Know where to find the relevant numbers in standard computer regression output.
- Be able to find a confidence interval for the slope of a regression based on the values reported in a standard regression output table.

TELL

- Be able to summarize a regression in words. In particular, be able to state the meaning of the true regression slope, the standard error of the estimated slope, and the standard deviation of the errors.
- Be able to interpret the P-value of the t-statistic for the slope to test the standard null hypothesis.
- Be able to interpret a confidence interval for the slope of a regression.

REGRESSION ANALYSIS ON THE COMPUTER

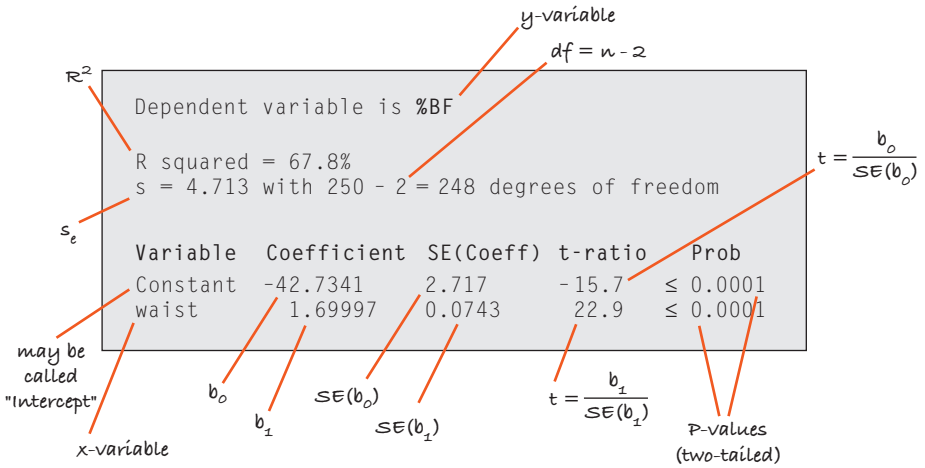
All statistics packages make a table of results for a regression. These tables differ slightly from one package to another, but all are essentially the same. We've seen two examples of such tables already.

All packages offer analyses of the residuals. With some, you must request plots of the residuals as you request the regression. Others let you find the regression first and then analyze the residuals afterward. Either way, your analysis is not complete if you don't check the residuals with a histogram or Normal probability plot and a scatterplot of the residuals against x or the predicted values.

You should, of course, always look at the scatterplot of your two variables before computing a regression.

Regressions are almost always found with a computer or calculator. The calculations are too long to do conveniently by hand for data sets of any reasonable size. No matter how the regression is computed, the results are usually presented in a table that has a standard form. Here's a portion of a typical regression results table, along with annotations showing where the numbers come from:

A S ***Activity:*** **Regression on the Computer.** How fast is the universe expanding? And how old is it? A prominent astronomer used regression to astound the scientific community. Read the story, analyze the data, and interactively learn about each of the numbers in a typical computer regression output table.



The regression table gives the coefficients (once you find them in the middle of all this other information), so we can see that the regression equation is

$$\widehat{\%BF} = -42.73 + 1.7\,Waist$$

and that the R^2 for the regression is 67.8%. (Is accounting for 68% of the variation in %Body Fat good enough to be useful? Is a prediction ME of more than 9% good enough? Health professionals might not be satisfied.)

The column of t-ratios gives the test statistics for the respective null hypotheses that the true values of the coefficients are zero. The corresponding P-values are also usually reported.

EXERCISES

T **1. Hurricane predictions.** In Chapter 7 we looked at data from the National Oceanic and Atmospheric Administration about their success in predicting hurricane tracks. Here is a scatterplot of the error (in nautical miles) for predicting hurricane locations 72 hours in the future vs. the year in which the prediction (and the hurricane) occurred: