Unit 11: Fitting Lines to Data

SUMMARY OF VIDEO

Scatterplots are a great way to visualize the relationship between two quantitative variables. For example, the scatterplot of temperatures and coral reef growth in Figure 11.1 shows that as temperatures go up, new coral growth goes down.

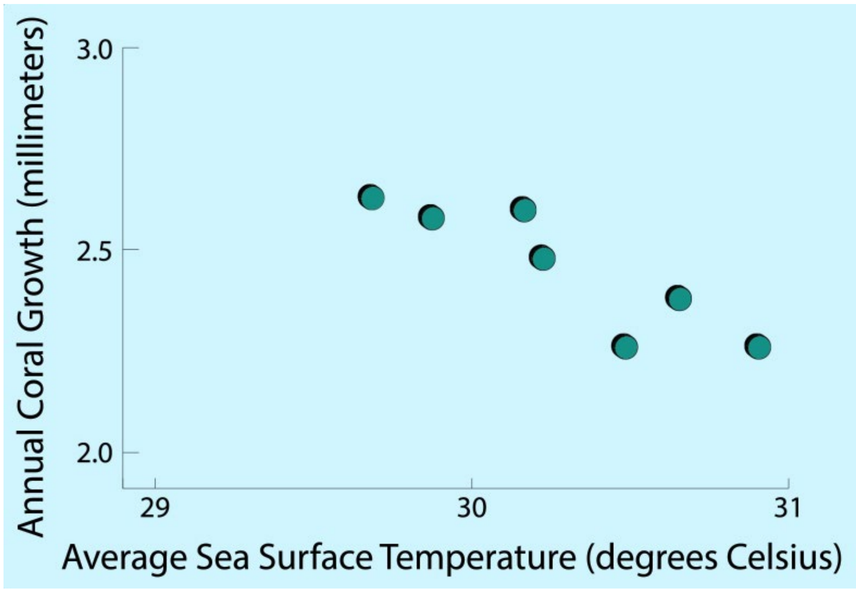

Figure 11.1. Scatterplot of coral growth versus ocean temperature.

Another example is the scatterplot in Figure 11.2 of height versus femur bone length in humans, which exhibits a positive linear relationship. In this case, the femur length is the explanatory variable and hence, is on the horizontal axis. The response variable is height, which is on the vertical axis. Since the dots in the scatterplot appear to form a linear pattern, a line has been added to the scatterplot. Statisticians call the line that describes how the response variable changes with the explanatory variable a regression line.

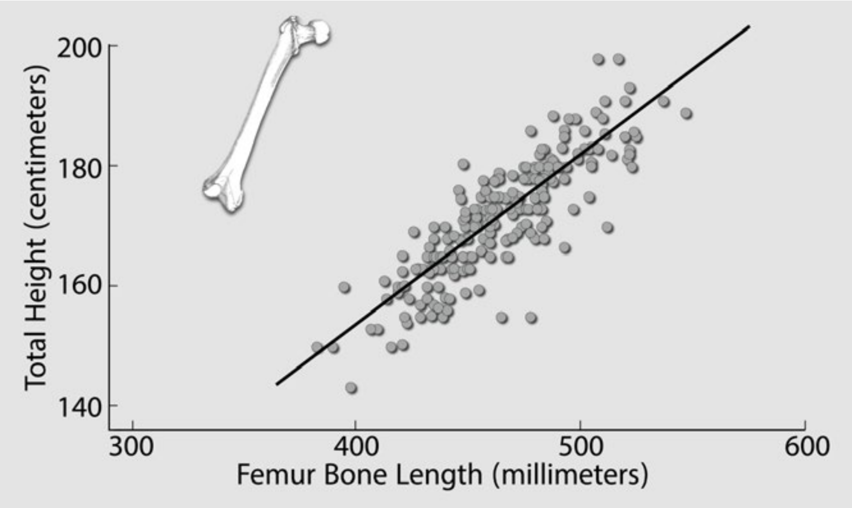

Figure 11.2. Scatterplot of height versus femur bone length.

Any line can be described by an equation of the form $y = a + bx$, where *a* is the *y*-intercept and *b* is the slope of the line. Recall the *y*-intercept is the *y*-value corresponding to *x* = 0 and the slope measures how much *y* changes when *x* increases by one unit. Sizing up the data points, we can just eyeball where the line should fall, but there is a statistical technique to figure out how best to fit a line to data. Once we have that line we can use it to make predictions. That's how a forensic scientist can estimate an unidentified crime victim's height just by using a femur bone measurement from incomplete skeletal remains.

The Colorado Climate Center uses regression lines for less sinister scenarios – to forecast the state's seasonal water supply. Farmers, city planners, and businesses, all need to know how much water is going to be available each year so they can plan accordingly. Climatologist Nolan Doesken introduces an important question for Colorado: How can we predict the water supply we're going to have as far ahead of time as possible? To answer this question, researchers have developed a model based on two types of data: the amount of winter snowpack in the high elevations and the resulting volume of water that flows out of the mountains throughout the summer. During the winter, Colorado's Natural Resources Conservation Service heads into the Rockies to collect data on the snowpack. Later, when the snowpack starts to melt, data related to the volume of water runoff are collected. Figure 11.3 shows a scatterplot of the data collected over many years.

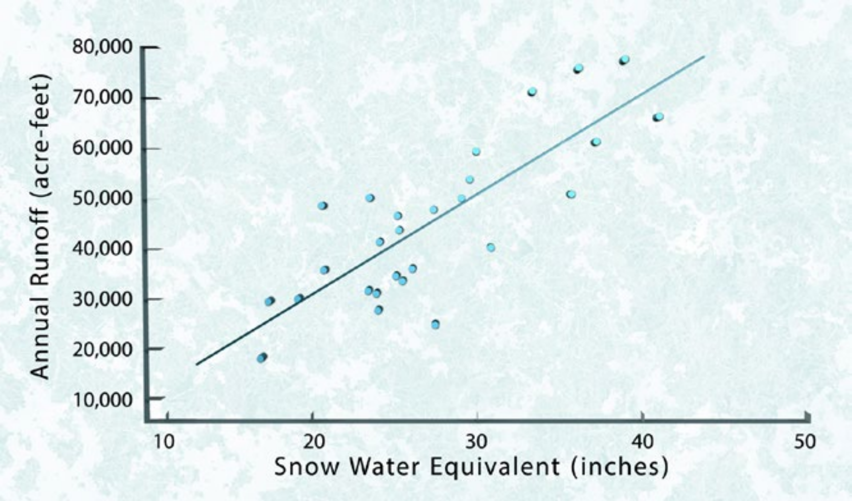

Figure 11.3. Scatterplot of annual water runoff versus snow water equivalent.

From the scatterplot in Figure 11.3, there appears to be a pretty strong positive linear relationship between the two variables and we have drawn a line to summarize that relationship. Of course, in the real world, all the points don't fall exactly on a line. So, we need a technique to determine the regression line that in some way minimizes the vertical distances of our data points from the line.

To get a better idea of how to fit a line to data, we zoom in on three of the data points from our Colorado water data. (See Figure 11.4.)

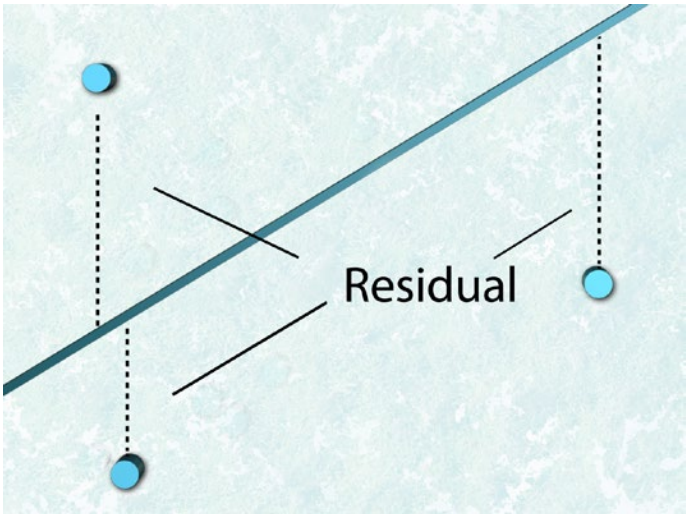

Figure 11.4. Zooming in on three data points and visualizing residuals.

The vertical distance of a data point from the line, together with a positive sign if the point lies above the line and a negative sign if the point lies below the line, is called a residual. As we shift the line trying to find the one that best fits these three points (see Figure 11.5.), some residuals get smaller while others get larger. So, we try to find the sweet spot where we've got them as a group as small as possible.

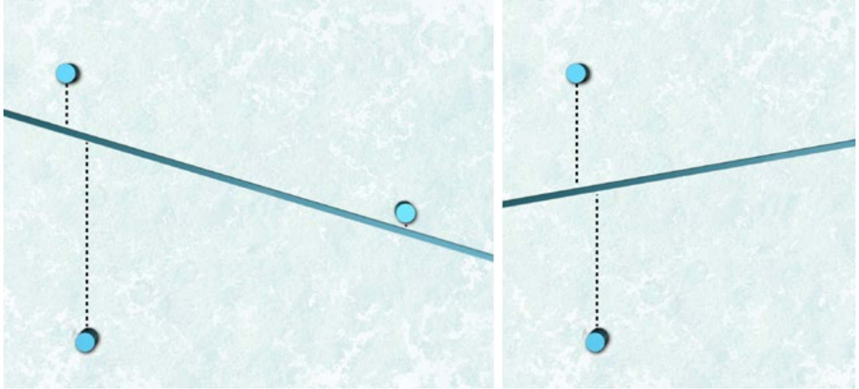

Figure 11.5. Residuals corresponding to different lines.

Statisticians use a method called least squares to find the "best-fitting" line. Since some of the residuals are positive and some are negative, we square them, which makes them all positive. If you add up the squared residuals, the bigger the sum, the more the line misses the points. So, we want to make that sum as small as possible. Software can compute the equation of the least-squares regression line – the line with the smallest sum of the squared residuals.

Back to the Colorado water data. Figure 11.6 shows both the graph of the least-squares regression line and its equation.

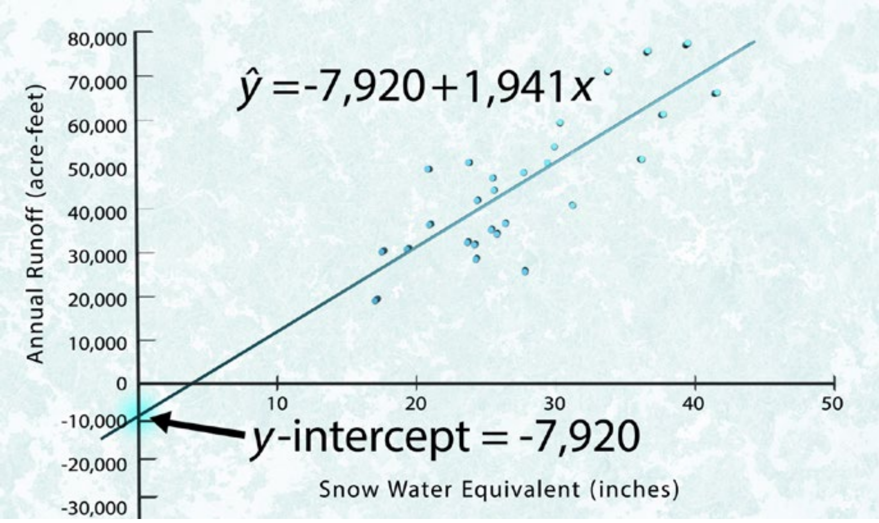

Figure 11.6. Graph of the least-squares regression line.

Note that in Figure 11.6 we have used the notation $\hat{y}$ (pronounced y-hat) to emphasize that we are talking about the predicted value of *y*, not a measured value from our data set. The slope of 1,941 predicts that for every one-inch increase in snowpack, the runoff increases by

1,941 acre-feet. The *y*-intercept is at -7,920. This value seems to say that if the snowpack was 0, the runoff would be -7,920 acre feet. Obviously that doesn't make sense, and it is a good reminder that you can't extrapolate from the regression line too far outside the range of the observed data. Keeping that limitation in mind, though, the regression line can be very useful for Colorado water users.

Now, we are ready to use the least-squares line to make a prediction. If you know that this winter the Rockies saw 30 inches of snowpack, you can look at the line in Figure 11.7 to predict how much water is going to flow into the system in the spring.

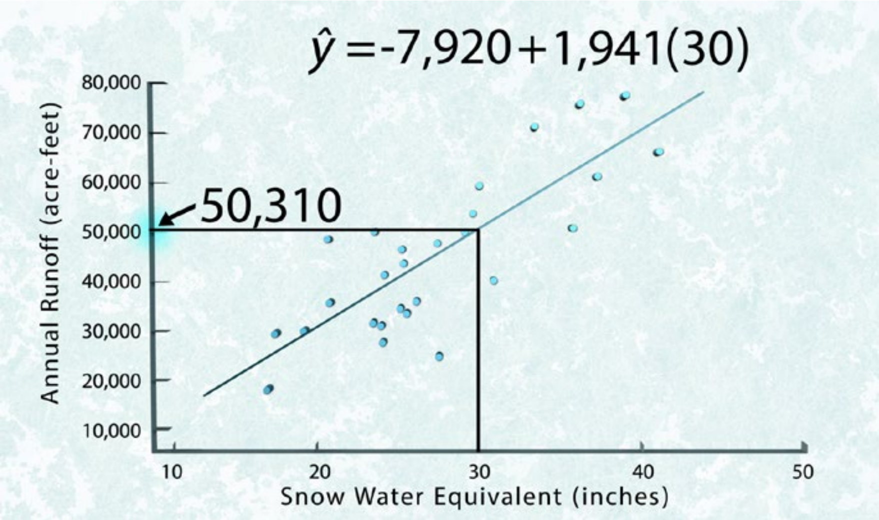

Figure 11.7. Using the least-squares equation to make a prediction.

The regression line works well to predict the Colorado water supply because the relationship between snowpack and runoff is linear. If the relationship you are trying to describe has a curved pattern, a *straight* regression line won't be a good fit. One way to assess how well a regression line fits the data is to make a residual plot, a plot of the residuals versus the explanatory variable. If the dots in the residual plot appear randomly scattered with no strong pattern, then the regression line has nicely captured the pattern in the data and a linear model is a good choice to describe it. For example, take a look at the residual plot for the Colorado data in Figure 11.8. The dots appear randomly scattered, with roughly equal numbers of dots above and below the horizontal axis.

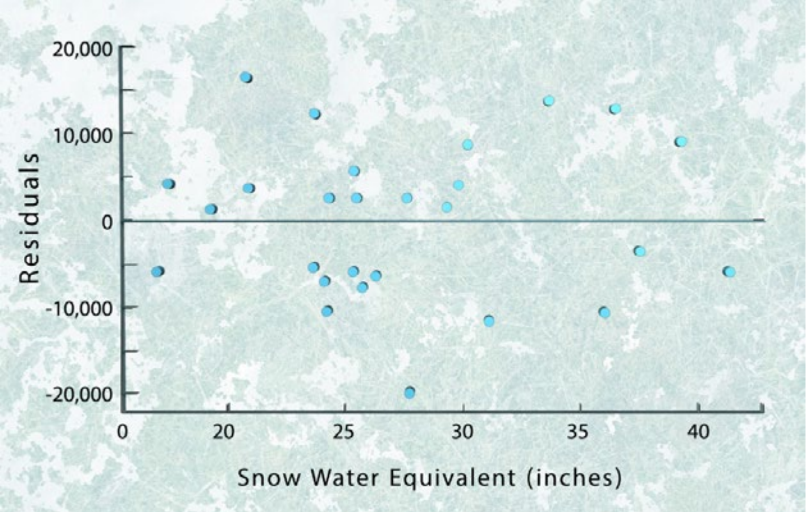

Figure 11.8. Residual plot for Colorado data.

What if your scatterplot had a curved pattern such as the one in Figure 11.9(a), made by data on alligator weight versus alligator length? If you fit a least-squares line to these data, then the residual plot will be curved as shown in Figure 11.9(b). You wouldn't want to use the equation of a line to make predictions about alligators! Instead you need to find an equation that describes the curved shape of the data.

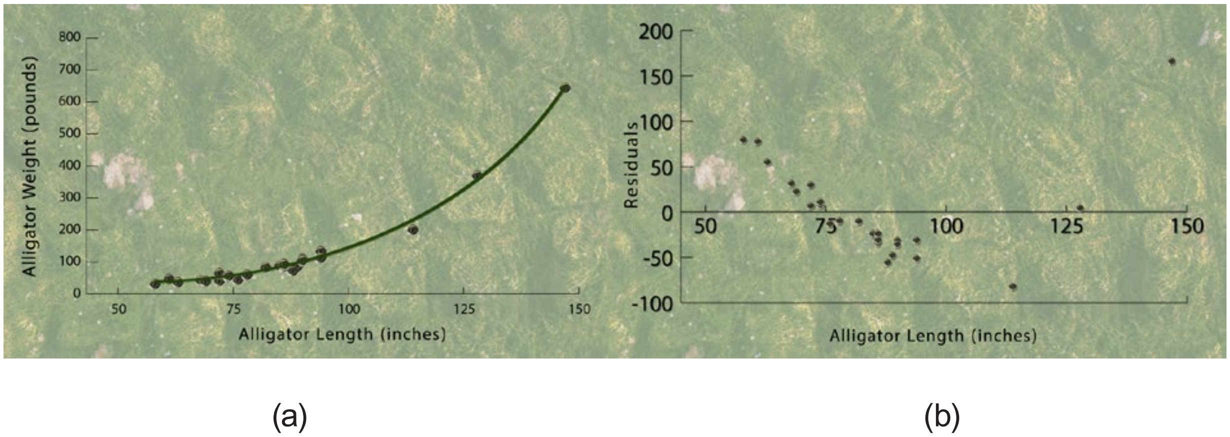

Figure 11.9. (a) Curved relationship between alligator weight and alligator length.

(b) Residual plot for fitting a line to alligator data.

STUDENT LEARNING OBJECTIVES

A. Be able to predict the value of the response variable corresponding to a given value of the explanatory variable from a graph of a line.

B. Understand the least-squares criterion for determining the line of “best fit” to data for the purpose of predicting *y* from *x*.

C. Know how to calculate the equation of the least-squares line for a small set of bivariate data (say n < 10). Know how to use software (such as Minitab or Excel) or a graphing calculator to calculate the equation of the least-squares line.

D. Understand the effect on the least-squares line of influential points that are extreme in the *x*-direction.

E. Know how to check the adequacy of a linear model by (1) viewing a scatterplot of the data to see if the pattern is linear and by (2) making a residual plot.

F. Know the difference between extrapolation and interpolation. Know when to beware of extrapolation.

CONTENT OVERVIEW

In this unit, we discuss a common method for fitting a line to data that show a linear pattern but whose points do not fall exactly on a line. We begin with a simple example that involves four data points: (2,1), (3,5), (4,3), and (5,2). Figure 11.10 shows a plot of these points and a line that might be a good fit because it goes through two of the four points.

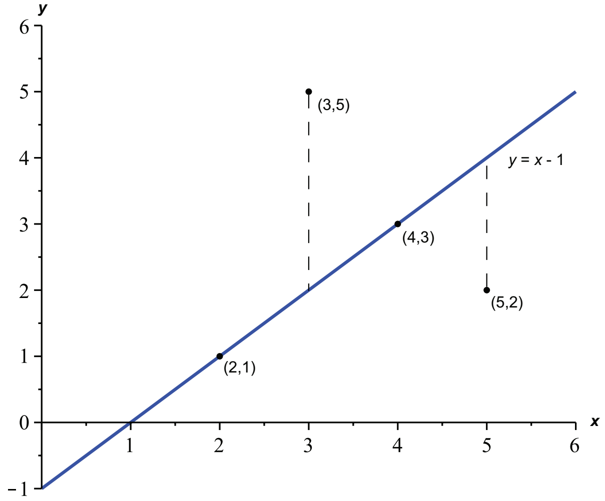

Figure 11.10. Fitting a line to four data points.

The line in Figure 11.10 can be described by the equation $y = x - 1$. But is this the best line to describe the pattern in these data? To measure how far off the line is from a point, we calculate the residual (also called the residual error) associated with each point:

residual = actual *y*-value from data – predicted *y*-value from the line

For example, the residual corresponding to (2,1) is 0. When *x* = 2, the *y*-value from the data point, *y* = 1, is the same as the predicted *y*-value from the equation, $\hat{y} = 2 - 1 = 1$. The residual corresponding to the point (4,3) is also zero, because this data point lies on the line. However, the residuals corresponding to (3,5) and (5,2) are 3 and -2, respectively. The residuals are represented graphically in Figure 11.10 by the dashed vertical lines. Notice that when a data point lies above the line, its residual is positive and when a data point lies below the line its residual is negative.

The least-squares method is the most common means of fitting a "best" line to data on an explanatory variable, *x*, and a response variable, *y*. This method fits a line that has the smallest sum of squares of residual (errors), or SSE for short. We express the formula for the least-squares regression line as $y = a + bx$, where

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}$$

$$a = \overline{y} - b\overline{x}$$

The formulas above are a bit complicated. So, to calculate the equation of the least-squares line for the data in Figure 11.10, we have set up Table 11.1. In the first two columns, we list the data values for *x* and *y*. Then, we calculate the mean for the *x*'s and *y*'s: $\overline{x} = 3.5$ and $\overline{y} = 2.75$. In columns three and four we subtract $\overline{x}$ and $\overline{y}$ from the *x*'s and *y*'s, respectively. In columns five and six, we form the products needed for the numerator and denominator of the formula for *b*. Finally, the sum of the entries in columns five and six gives us the numerator and denominator of *b*, respectively.

x	y	$x - 3.5$	$y - 2.75$	$(x - 3.5)(y - 2.5)$	$(x - 3.5)^2$
2	1	-1.5	-1.75	2.625	2.25
3	5	-0.5	2.25	-1.125	0.25
4	3	0.5	0.25	0.125	0.25
5	2	1.5	-0.75	-1.125	2.25
			Sum=	0.5	5

Table 11.1. Calculations for slope of least-squares line.

$$b = 0.5 / 5 = 0.1$$

$$a = 2.75 - (0.1)(3.5) = 2.4$$

The equation of the least-squares line can be expressed as $y = 0.1x + 2.4$. Figure 11.11 shows the graph of the least-squares line added to Figure 11.10. Unlike our first line, the least-squares line does not contain any of the data points. However, the SSE for the least-squares line, the sum of the squared lengths of the vertical line segments, will be less than the SSE for the line *y* = *x* – 1. (For confirmation, work through question 5 in the Review Questions.)

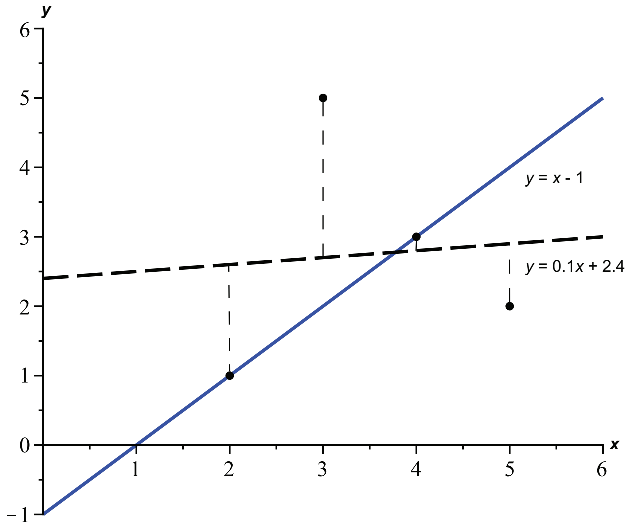

Figure 11.11. Comparing graphs of the least-squares line (dashed) and y = x – 1.

Although you now know how to compute the equation for the least-squares line, it is much faster use technology for the calculations particularly for larger data sets. Statistical software, spreadsheet software, and graphing calculators all have built-in linear regression capabilities that will report the equation of the least-squares line.

Regression assumes that we want to predict the response variable *y* given values of the explanatory variable *x*. The distinction between an explanatory and a response variable is essential; reversing the roles of the two variables produces a different regression line. (See question 6 in the Unit Activity.) In Figure 11.3, the response variable, *y*, was the annual water runoff (in acre-feet) in Colorado and the explanatory variable, *x*, was the snowpack (inches). We can use the equation of the least-squares line, $\hat{y} = -7{,}920 + 1{,}941x$ to make predictions. For example, suppose the snowpack one year turned out to be *x* = 40 inches. Then the predicted water flow for the next spring would be $\hat{y} = -7{,}920 + 1{,}941(40) = 69{,}720$ acre-feet.

Of course, it only makes sense to use the equation of the regression line to make predictions if the line adequately describes the pattern of the data. One way to check is with a residual plot, a plot of the residuals versus the explanatory variable. Figure 11.8 shows a residual plot for the Colorado runoff-snowpack data. A good residual plot looks like a bunch of dots randomly thrown onto the graph – there should be no strong pattern, and roughly half the points should lie above the horizontal axis and half below. The residual plot in Figure 11.8 meets both of these criteria. If, on the other hand, the residual plot shows a strongly curved pattern then the least-squares line fails to adequately describe the pattern in the data. Using the least-squares equation to make predictions in this case, could lead to some very inaccurate predictions. What is needed is a search for another model, one that would better describe the curved pattern of the data.

We conclude with a two warnings. The first is that the least-squares regression line can be strongly influenced by one or more extreme points. Points that lie far from the other data in the *x*-direction (as might be the case if data appear in clusters) are particularly dangerous. In this case, a residual plot will most likely exhibit some strong patterns indicating that the model is not adequate to describe the pattern in the data. The second warning is to beware of extrapolation – the use of a fitted line for prediction outside the range of *x*-values in the data. For example, return to the problem of predicting Colorado water runoff. Suppose in one very dry year the snowpack measured 2 inches, much lower than anything observed in the data. Using the least-squares equation to predict the spring water runoff gives:

y = -7920 + 1941(2) = -4038 acre feet.

Clearly, the volume of water runoff can't be negative. Contrast this situation with the previous prediction of 69,720 acre feet for a snowpack of 40 inches. That prediction made sense and is an example of **interpolation**, because a 40-inch snowpack lies within the range of observed values for snowpack.

KEY TERMS

A **residual** (or **residual error**) is the vertical deviation of a data point from the regression model:

residual = observed *y* – predicted *y*.

The **least-squares regression line** is the line that makes the sum of the **squares of the residual errors**, **SSE**, as small as possible. The equation of the least-squares line is $\hat{y} = a + bx$, where the *y*-intercept, *a*, and slope, *b*, are calculated from *n* data values on an explanatory variable *x* and a response value *y*. To calculate the values of *a* and *b* without the use of software, use the following formulas:

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}$$

$$a = \overline{y} - b\overline{x}$$

Because the formula for computing the least-squares regression line will fit a line to any data set on two variables *x* and *y*, it is important to assess the **adequacy** of the least-squares model for describing the pattern in the data. One way to judge the adequacy of a linear model is to make a **residual plot** – a plot of the residuals versus the explanatory variable. If the dots in the residual plot appear randomly scattered with roughly half above and half below the *x*-axis, then the linear model is adequate to describe the pattern in the data. Otherwise, look for a new model.

An observation is **influential** if removing it would greatly change the position of the regression line. Points that are separated in the *x*-direction from the other observations, such as in a cluster pattern, are often influential.

Interpolation is the use of the regression equation to predict *y* for values of *x* that lie inside the range of values in the data used to fit the line. **Extrapolation** is the use of the regression equation to predict *y* for values of *x* outside the range of values in the data used to fit the line. Models don't hold forever, hence extrapolation can be risky.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. How is the snowpack during wintertime in the Colorado Mountains measured?

2. What is a residual?

3. How does the least-squares method decide which line best fits the points in a scatterplot?

4. How can a particular year's data on the snowpack be used to predict the amount of water running downstream in the spring?

5. The video showed two examples of residual plots. What does a residual plot tell you if the dots in the plot appear to be randomly scattered? What if the dots appear to form a strong curved pattern instead?

UNIT ACTIVITY:
THE RELATIONSHIP BETWEEN FOREARM LENGTH AND FOOT LENGTH

In determining normal proportions in human bodies, doctors look at the relationships between the lengths of various body parts. Artists are also interested in these relationships. Knowing such relationships helps them draw human figures that appear appropriately proportioned. On a less serious note, this activity was inspired by a piece of trivia from Julia Roberts' character in the 1990 movie Pretty Woman, in which she states – your forearm length is the same as your foot length. To check the validity of this assertion, we collected data on forearm and foot lengths of 26 college students enrolled in an introductory statistics course. These data appear in Table 11.2.

Forearm Length (inches), *x*	Foot Length (inches), *y*	Forearm Length (inches), *x*	Foot Length (inches), *y*
10.00	9.50	8.75	9.00
9.00	9.00	9.00	10.50
10.00	9.50	8.50	11.00
10.00	10.00	10.25	11.50
11.50	12.50	10.25	11.25
9.00	11.50	8.50	9.00
8.50	9.00	9.25	10.50
6.75	9.25	10.50	10.50
10.00	10.00	8.25	8.50
8.25	8.25	9.00	10.00
8.25	9.50	7.00	8.75
9.00	9.50	9.50	8.75
8.00	9.50	9.75	10.00

Table 11.2: Data on forearm and foot length.

Since students objected to removing their shoes in order to measure their feet, forearm length is the explanatory variable and foot length is the response variable. That way, after determining a relationship between these two variables, we can use it to predict students' foot lengths from their forearm lengths (and they won't have to take off their shoes!).

1. a. Make a scatterplot of foot length, *y*, versus forearm length, *x*.

b. Based on your scatterplot, does it appear that students with longer forearms tend to have bigger feet? Explain how you can tell from the scatterplot.

2. Use technology to determine the equation of the least-squares regression line. Superimpose a graph of this line on your scatterplot. Does the line appear to do a good job of summarizing the pattern of the points in the scatterplot?

3. Use the equation of the least-squares line to predict the foot length of a person with a 10.5 inch forearm.

4. Find the residual corresponding to the first data point (10, 9.50). Show your calculations.

5. Before using the least-squares line to make predictions (as we did in question 3), we first should have checked on the adequacy of the linear model. Figure 11.12 shows a residual plot. Based on the residual plot, is the least-squares line adequate to describe the pattern in these data? Explain.

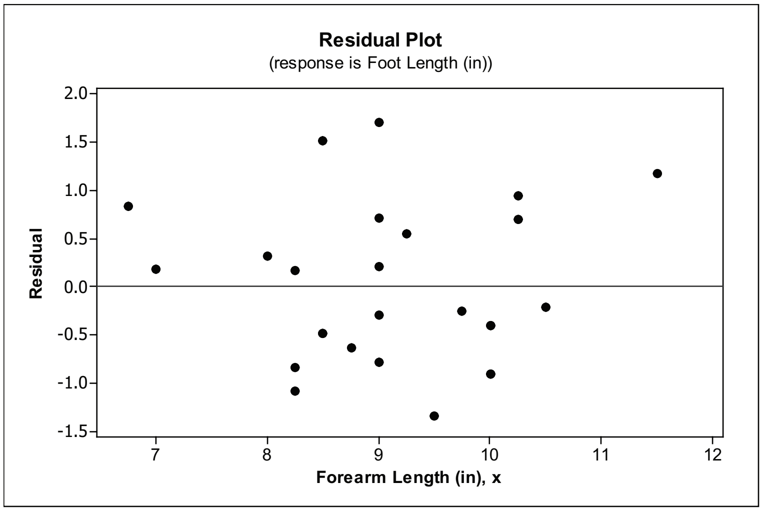

Figure 11.12. Residual plot.

6. Danny got his explanatory and response variables mixed up. He fit a least-squares line to forearm length, x, versus foot length, y, and got the following equation:

$x = 2.865 + 0.6332y$

Sarah, his partner, told him not to worry but to solve for y and he would get the correct equation for the least-squares line. Follow Sarah's advice and solve for y in terms of x. Do your results agree with the equation you got for question 2? What does this tell you about Sarah's strategy for fixing her partner's mistake?

7. Linda fit a line to the data by finding the equation of a line that contains (8,9.5) and (10.5,10.5). The equation for her line is $y = 0.4x + 6.3$. What can you say about the sum of squared residuals, SSE, for her line compared to the SSE for the least-squares line? Explain.

Extension to question 7: Find the SSE for the least-squares line and for Linda's line. Which line had the smaller SSE?

EXERCISES

The data on the average sea surface temperature and coral reef growth shown in Figure 11.1 appear below.

Temperature (°C), x	29.7	29.9	30.2	30.2	30.5	30.7	30.9
Coral Growth (mm), y	2.63	2.58	2.60	2.48	2.26	2.38	2.26

Table 11.3. Temperature and coral growth data.

1. a. Make a scatterplot of coral reef growth versus the average sea surface temperature. Describe the pattern. Are there any outliers (data points that appear to deviate from the overall pattern)?

b. Computer software gives the equation of the least-squares regression line as

$$\hat{y} = 12.3 - 0.325x$$

Add a graph of this line to your scatterplot in (a).

2. Return to the coral reef data in Table 11.3. Next, you will verify the equation in exercise 1(b) using the formulas given in the Content Overview.

a. Determine $\bar{x}$ and $\bar{y}$.

b. Make a copy of Table 11.4 and complete the entries. Record the sum of the last two columns in the shaded cells.

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
29.7	2.63				
29.9	2.58				
30.2	2.6				
30.2	2.48				
30.5	2.26				
30.7	2.38				
30.9	2.26				
			Sum =		

Table 11.4. Table used to compute b.

c. Using the information from (b), show the calculations for the slope, *b*, and *y*-intercept, *a*, of the least-squares line.

3. a. Figure 11.13 shows a residual plot corresponding to the least-squares regression line determined in exercise 2. Based on this plot is the least-squares line adequate to describe the pattern in the data? Explain.

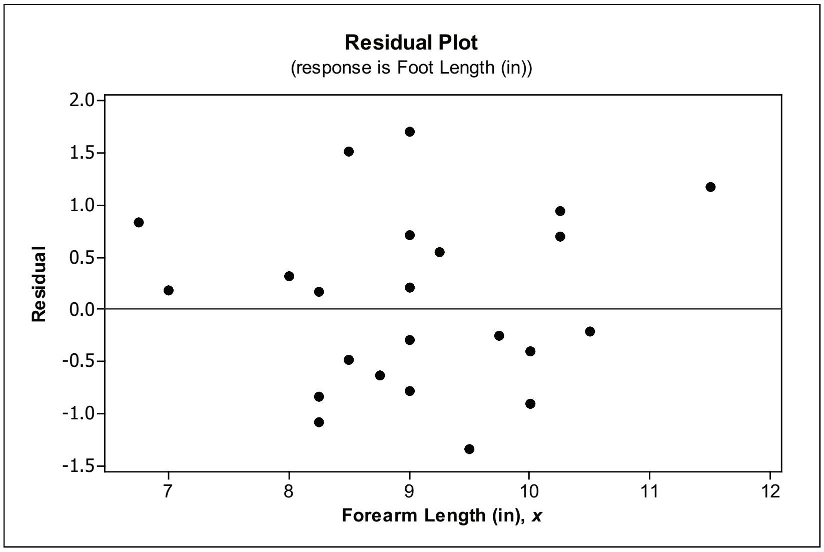

Figure 11.13. Residual plot.

b. Regardless of your answer to (a), assume that the least-squares line is adequate to describe the pattern in the data from Table 11.2. Use the equation given in 1(b) to predict the coral growth if the average sea surface temperature rises to 40°.

4. Satellites are one of the many tools used for predicting flash floods, heavy rainfall, and large amounts of snow. Geostationary Operational Environmental Satellites (GOES) collect data on cloud top brightness temperatures (measured in degrees Kelvin (°K)). It turns out that colder cloud temperatures are associated with higher and thicker clouds, which could be associated with heavier precipitation. Data consisting of cloud top temperature measured by a GOES satellite and rainfall rate measured by ground radar appear in Table 11.5. Because ground radar can be limited by location and obstructions, having an alternative for predicting the rainfall rates can be useful.

Temperature	Radar Rain Rate	Temperature	Radar Rain Rate
(°K)	(mm/h)	(°K)	(mm/h)
195	150	203	44
196	150	204	39
197	150	205	39
198	118	206	35
199	109	207	38
200	95	208	31
201	63	209	20
202	66	210	24

Table 11.5. Sixteen data pairs of (temperature, rain rate) data.

a. You first encountered these data in Unit 10, exercise 3, where you were asked to make a scatterplot of these data. Since we want to use temperature to predict rain rate, temperature is the explanatory variable. Fit a least-squares line to rain rate versus temperature data. Sketch a scatterplot of the data and the line. Report the equation of your line.

b. Use your equation from (a) to predict the rain rate when the cloud temperature is 220 °K. Does your answer make sense? Explain.

c. Make a residual plot. Do you think a straight-line model is adequate to describe the pattern in these data? Explain why or why not.

REVIEW QUESTIONS

1. A random sample of femur bone lengths (mm) and heights (cm) from 20 males appears in Table 11.6. These data are from the Forensic Anthropology Data Bank at the University of Tennessee.

Femur Length (mm)	Height (cm)
447	168
444	168
470	175
459	170
482	178
520	191
464	175
470	172
482	182
462	178
522	193
461	171
422	160
520	185
476	180
508	183
477	173
504	175
547	189
508	198

Table 11.6. Data on femur bone length and height.

a. We would like to predict the height of a male given the length of his femur bone. Which variable is the explanatory variable and which is the response variable? Explain.

b. Enter the data into columns of computer software (or calculator lists). Make a scatterplot of the data. Describe the pattern in your scatterplot. Does the pattern appear linear or nonlinear? Is the association between the variables positive or negative?

c. Fit a least-squares line to the data and report its equation. Overlay a graph of the line on your scatterplot from (b).

d. Do there appear to be any outliers? If so, identify the point(s). Do you think these are mistakes or real data values? Explain.

2. Return to your work from question 1.

a. Interpret the slope and *y*-intercept in the context of these data. Do these quantities make sense in the given context?

b. Two femur bones presumed to be from two men are measured and their lengths differ by 5 mm. Use the least-squares regression equation to predict the difference in heights between the two men.

c. Predict the height of a male whose femur length is 475 mm. Is this a reasonable height for a man? (Convert your answer to feet and inches. Recall there are 2.54 centimeters per inch.)

d. The femur length of a boy measures 250 mm. Predict the height of the child. Explain why this prediction might not be trustworthy.

3. Table 11.7 contains data on mercury concentration in tissue samples from 20 largemouth bass taken from Lake Natoma in California. Only fish of legal/edible size were used in this study.

Total Length (mm)	Mercury Concentration (μg/g wet wt.)
341	0.515
353	0.268
387	0.450
375	0.516
389	0.342
395	0.495
407	0.604
415	0.695
425	0.577
446	0.692
490	0.807
315	0.320
360	0.332
385	0.584
390	0.580

410	0.722
425	0.550
480	0.923
448	0.653
460	0.755

Table 11.7. Fish length and mercury concentration in fish tissue samples.

a. We want to be able to predict mercury concentration from fish length. Which variable is the explanatory variable and which is the response variable?

b. Fit a least-squares line to the data from Table 11.7. Report its equation. (Round the slope and *y*-intercept to four decimals.) Also, show a scatterplot of the data and the least-squares line.

c. Make a residual plot. Based on your plot, is the least-squares model adequate to describe the overall pattern in the data? Explain.

d. Interpret the slope of the least-squares line the context of this problem. Does the interpretation of slope make sense in the given context? Explain why or why not.

e. Interpret the *y*-intercept of the least-squares line the context of this problem. Does the interpretation of the *y*-intercept make sense in the given context? Explain why or why not.

4. Return to the data in exercise 3, Table 11.7. Use your answer to 3(b) to make the following predictions:

a. Predict the mercury concentration in a largemouth bass that is 430 mm in length. Is this prediction an example of interpolation or extrapolation? Explain.

b. Predict the mercury concentration in a largemouth bass that is 90 mm in length, which is below the legal/edible size. Is this an example of interpolation or extrapolation? Explain.

5. In the Content Overview, we fit two lines to the data in Table 11.8. Graphs of these lines appear in Figure 11.11.

x	y
2	1
3	5
4	3
5	2

Table 11.8. Data from Figure 11.10.

a. The equation of the first line was $y = x - 1$. Calculate the sum of the squares of the residuals, SSE, for this line.

b. The equation of the least-squares line is $y = 0.1x + 2.4$. Calculate the SSE for the least-squares line.

c. Which line had the smaller SSE? Why should we not be surprised by this result?