Unit 11: Fitting Lines to Data

PREREQUISITES

Students must be familiar with material from Unit 10, Scatterplots. They must have some mathematical background on linear functions and know that the graph of a linear function is a line. Students will need to be prepared for a change in notation from what they are used to seeing in mathematics textbooks, $y = mx + b$. In this unit, a generic linear model is written as $y = a + bx$, which is consistent with the notation used in many introductory statistics textbooks. In addition, students should be familiar with the meaning of summation notation.

ACTIVITY DESCRIPTION

The unit activity provides forearm length and foot length data collected from 26 college students enrolled in an introductory statistics course. Students will need to use technology (statistical software, spreadsheet software, or graphing calculators) for computing the equation of the least-squares regression line.

MATERIALS

Access to technology with regression capabilities; graph paper (optional).

For question 2 students are asked to make a scatterplot of the data and to graph the least-squares line. They can use technology and then make a rough sketch of the results. If you want them to make a scatterplot by hand and then graph the least-squares line, they will need graph paper.

Question 7 asks students to compare the SSE for the least-squares line to the SSE of another line. From theory they should know that the least-squares line has the smaller SSE. However, as an extension to question 7, students are asked to calculate the 26 residuals associated with each of the two lines and then to calculate the SSEs. Calculator lists or Excel work really well for this computation.

Fitting a line to data that are related to the students themselves is often more interesting than working with data provided from the outside. Consider a second extension to this activity. In

addition to the data provided in the activity, substitute data collected from your class and have students repeat the activity. Alternatively, students could gather data on other variables, such as height and armspan, or height and forearm length. If you want an example of a weaker relationship, have students collect data on height and how high they can jump. To gather the jump data, tape a yard stick on a wall (or use chalk to mark off a scale in inches). As a student jumps, an observer can record his or her height. Another alternative is to gather data on height and stride length. Forensic scientists might be able to determine the stride length of a criminal from foot prints and then use that information to predict the person's height.

THE VIDEO SOLUTIONS

1. Researchers push long tubes that have scales along the side into the snowpack.

2. It is the difference between the actual *y*-value and the *y*-value predicted by the least-squares line.

3. It finds the line for which the sum of the squares of the residuals is smallest.

4. Substitute the value for snowpack into the equation of the least-squares line to get the predicted value of water runoff.

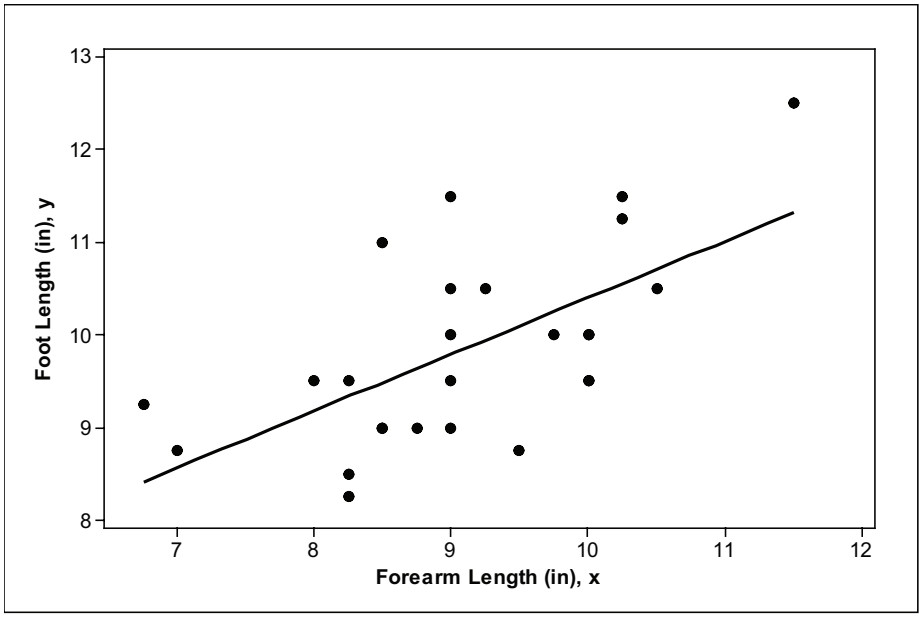
5. If the dots appear randomly scattered with no strong pattern, then the regression line is adequate to describe the pattern in the data. If the dots in a residual plot appear to have a strong curved pattern, then the linear model is not adequate to describe the pattern in the data and you need to look for a new model.

UNIT ACTIVITY SOLUTIONS

1. a. See solution to question 2.

b. Yes. The pattern of the dots in the scatterplot go from the lower left to the upper right.

2. The equation of the least-squares line is $y = 4.291 + 0.6112x$. Yes, the line provides a reasonable summary of the forearm-foot length data.



3. $\hat{y} = 4.291 + 0.6112(10.5) \approx 10.7$ inches.

4. First, calculate the predicted y-value: $\hat{y} = 4.291 + 0.6112(10) = 10.403$.

Residual = 9.50 – 10.403 = -0.903 inch.

5. The dots in the residual plot appear randomly scattered with no strong pattern. Therefore, the least-squares line is adequate to describe the pattern in the data.

6. $y = -4.525 + 1.579x$; It didn't even come out close to being the same. Clearly Sarah's strategy for fixing Danny's error was faulty.

7. The SSE for her line will be larger. The least-squares line is the line with the smallest SSE of all possible lines.

Extension to question 7: As expected, the least-squares line has the smaller SSE.

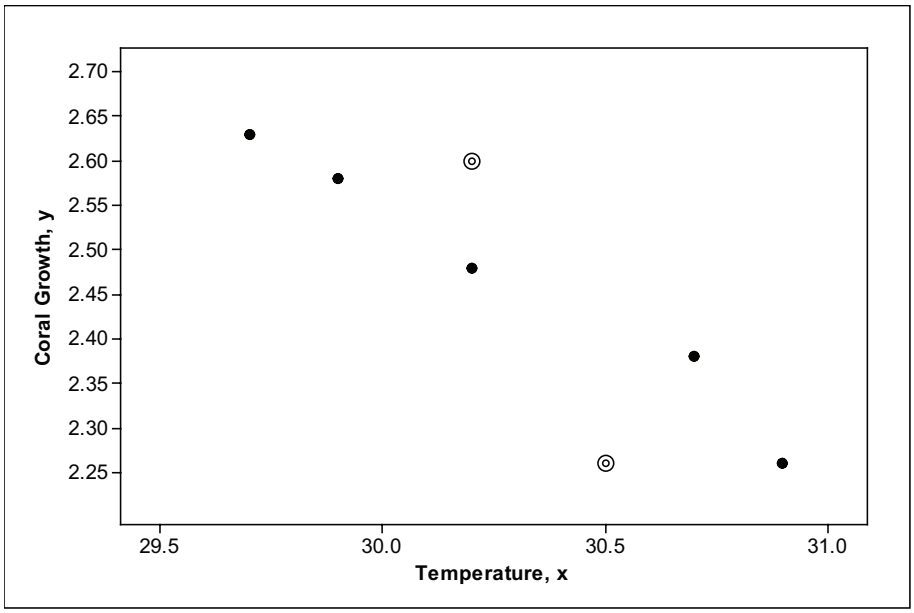
SSE for least-squares line = 17.025.

SSE for Linda's line = 18.503. (See spreadsheet table below for calculation of the residuals for Linda's line.)

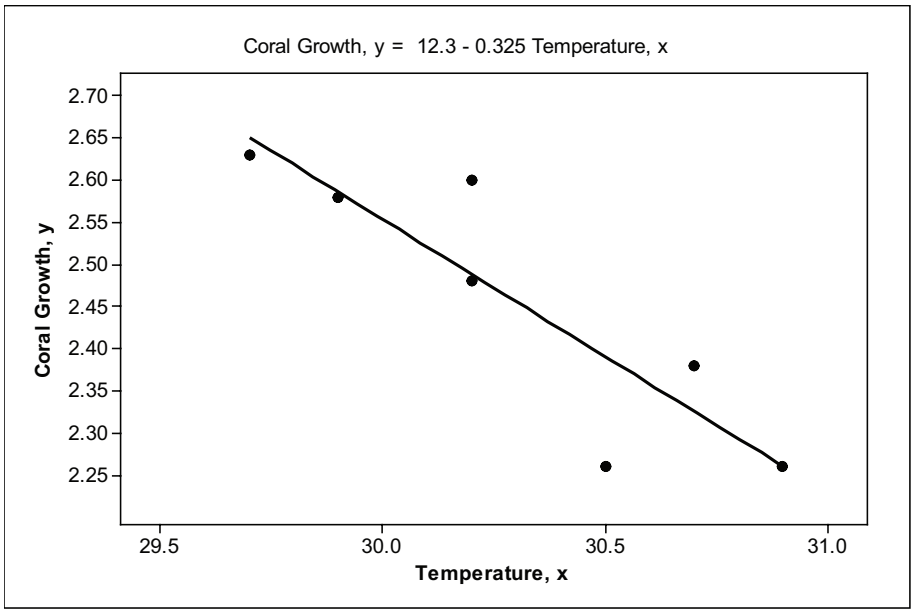
Forearm Length	Foot Length	Predicted	Residual	Residual^2
10.00	9.50	10.30	-0.80	0.640
9.00	9.00	9.90	-0.90	0.810
10.00	9.50	10.30	-0.80	0.640
10.00	10.00	10.30	-0.30	0.090
11.50	12.50	10.90	1.60	2.560
9.00	11.50	9.90	1.60	2.560
8.50	9.00	9.70	-0.70	0.490
6.75	9.25	9.00	0.25	0.063
10.00	10.00	10.30	-0.30	0.090
8.25	8.25	9.60	-1.35	1.823
8.25	9.50	9.60	-0.10	0.010
9.00	9.50	9.90	-0.40	0.160
8.00	9.50	9.50	0.00	0.000
8.75	9.00	9.80	-0.80	0.640
9.00	10.50	9.90	0.60	0.360
8.50	11.00	9.70	1.30	1.690
10.25	11.50	10.40	1.10	1.210
10.25	11.25	10.40	0.85	0.722
8.50	9.00	9.70	-0.70	0.490
9.25	10.50	10.00	0.50	0.250
10.50	10.50	10.50	0.00	0.000
8.25	8.50	9.60	-1.10	1.210
9.00	10.00	9.90	0.10	0.010
7.00	8.75	9.10	-0.35	0.123
9.50	8.75	10.10	-1.35	1.823
9.75	10.00	10.20	-0.20	0.040
			SSE =	18.503

EXERCISE SOLUTIONS

1. a. Sample answer: There do not appear to be any real outliers. However, the two circled data points depart somewhat from what would otherwise be a strong linear pattern.



b.



2. a. $\bar{x} = 30.3$; $\bar{y} = 2.46$

b.

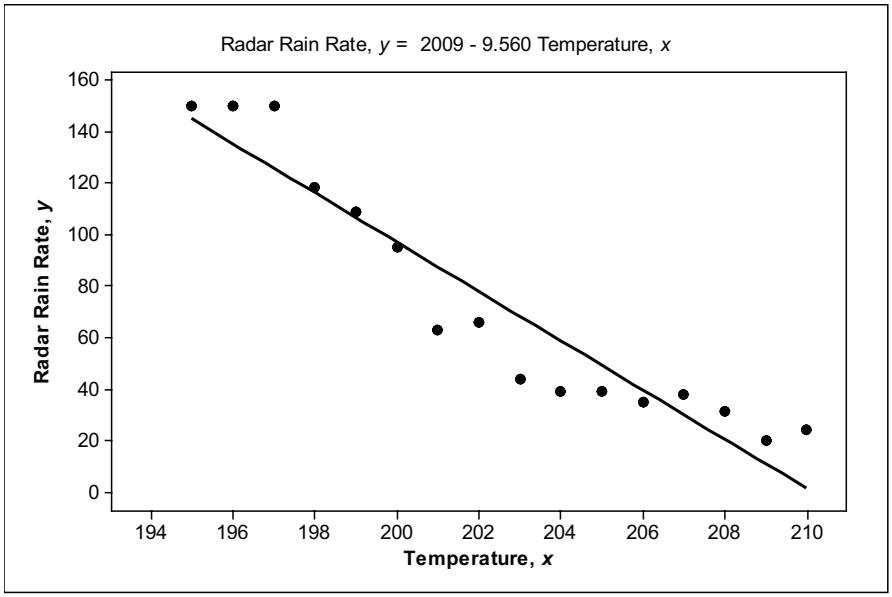
x	y	$(x-30.3)$	$(y-2.46)$	$(x-30.3)(y-2.46)$	$(x-30.3)^2$
29.7	2.63	-0.6	0.17	-0.102	0.36
29.9	2.58	-0.4	0.12	-0.048	0.16
30.2	2.6	-0.1	0.14	-0.014	0.01
30.2	2.48	-0.1	0.02	-0.002	0.01
30.5	2.26	0.2	-0.2	-0.04	0.04
30.7	2.38	0.4	-0.08	-0.032	0.16
30.9	2.26	0.6	-0.2	-0.12	0.36
			Sum =	-0.358	1.1

c. The slope b = -0.358/1.10 ≈ -0.325; the y-intercept a = 2.46 – (-0.325)(30.3) ≈ 12.3

3. a. Sample answer: The dots appear randomly scattered (although it is hard to tell if there is a strong pattern with so few points). Four dots are below the x-axis and 3 are on or above the x-axis. So, it appears that the line has taken out all the pattern in the data leaving only random noise. We can conclude that the least-squares regression line is adequate to describe the pattern in the data.

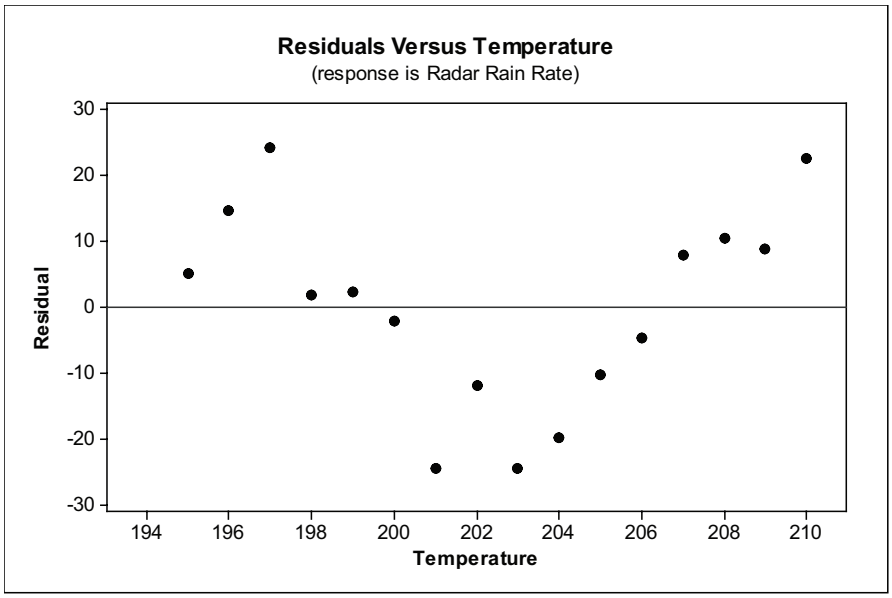
b. $\hat{y} = 12.3 - 0.325(40) \approx -0.7$

4. a. y = 2,009 - 9.56x, where x is temperature and y is radar rain rate.



b. 2,009 – 9.56(220) = -94.2 mm/h. No, you can't have a negative rain rate. This is an example of what can happen when you extrapolate beyond the observed data values.

c. There is a strong V-shaped pattern to the dots in the residual plot. A straight-line model is not adequate to describe the pattern in these data.

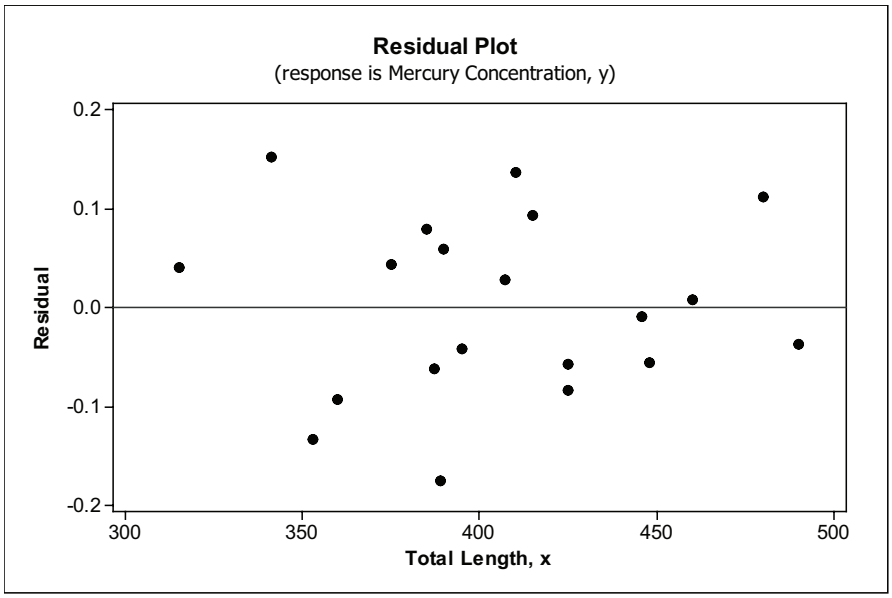

Residuals Versus Temperature
(response is Radar Rain Rate)
Residual
Temperature
30
20
10
0
-10
-20
-30
194
196
198
200
202
204
206
208
210

REVIEW QUESTIONS SOLUTIONS

1. a. Femur bone length is the explanatory variable since we wish to use it to explain a person's height. Therefore, height is the response variable.

b. The data appear to form a linear pattern. Men with longer femurs tend to be taller than men with shorter femurs. Thus, the association is positive.

c. The equation of the least-squares line is $y = 51.01 + 0.2637x$. The scatterplot with a graph of the least-squares line appears below.



d. Sample answer: There appears to be one outlier – data point (508, 198). This male is taller than we would expect given the pattern in the data. (See graph in (c). This point has been plotted with an open double circle.) The man is 198 cm tall or around 6' 6" tall. There are men that are this tall, even though it is quite tall for a man. So, it doesn't appear to be an error.

2. a. The slope is 0.2637; for each one mm increase in femur bone length we expect about a 0.26 cm increase in height. This makes sense in context.

The *y*-intercept is 51.01; this is the predicted height in centimeters of a person whose femur length is 0 mm. A femur length of 0 mm is far outside the range of observed femur lengths – the person would be missing his/her thigh. This does not make sense in context.

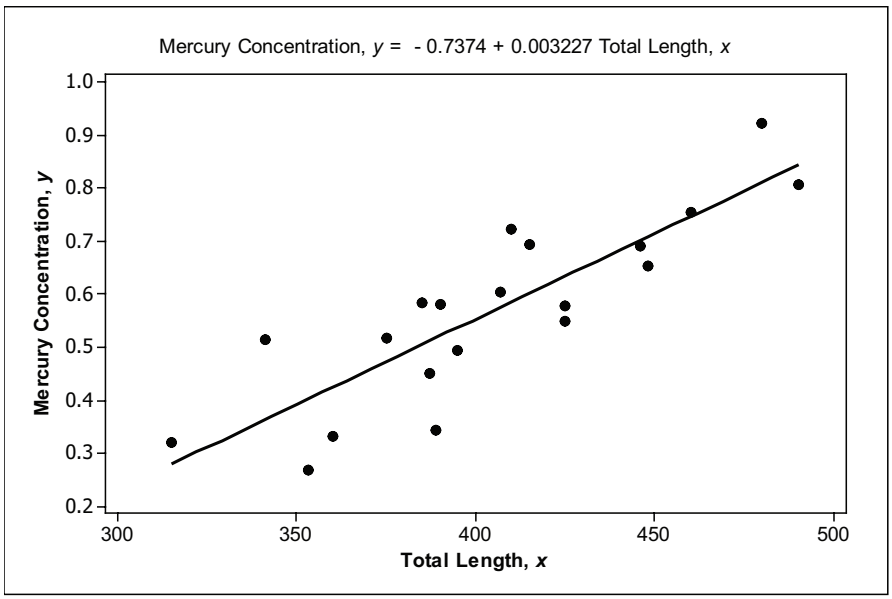
b. The slope gives the predicted change in height (cm) for each one millimeter increase in femur length. Hence, we would predict a (5)(0.264) or around 1.3 cm difference in height in response to a 5 mm difference in femur length.

c. $\hat{y} = 51.0 + 0.264(475) \approx 176.4 \text{ cm}$; or a little less than 5' 9½" tall. This is a reasonable height for a man.

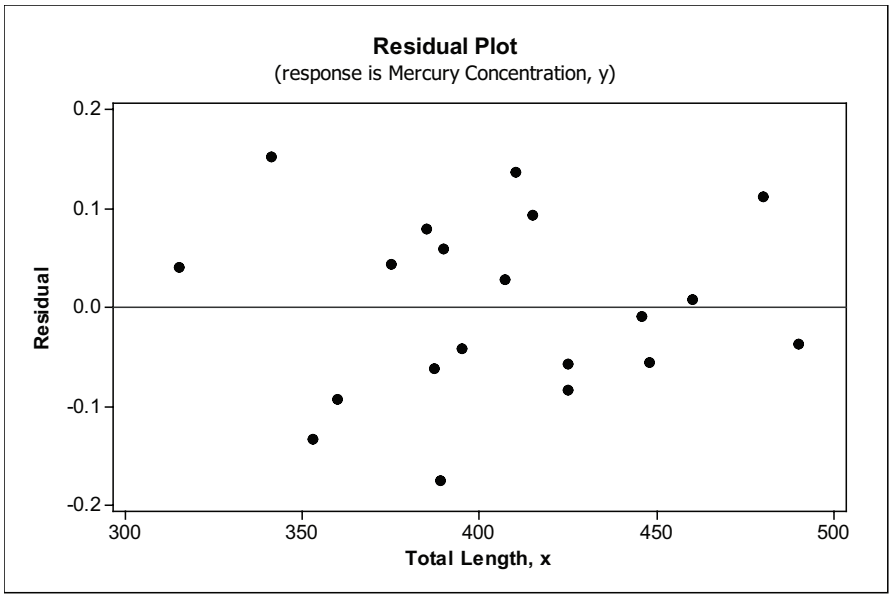
d. $\hat{y} = 51.0 + 0.264(250) \approx 117 \text{ cm}$; or around 3' 10" tall. This is an example of extrapolation. The smallest femur length in the data is 422 mm. All of the data is for adult males. We have no idea if the relationship between height and femur length is the same for children as it is for adults.

3. a. Total length is the explanatory variable, *x*, and mercury concentration is the response variable, *y*.

b. $y = 0.0032x - 0.7374$, where *x* is fish length and *y* is mercury concentration.



c. The least-squares line is adequate to describe the overall pattern in the data. The dots in the residual plot appear to be randomly scattered. Also, there is a good split between dots above the *x*-axis and below the *x*-axis.



d. For each additional 1 mm in fish length, we expect mercury concentration to increase by 0.0032 μg/g. This makes sense in the given context.

e. The *y*-intercept of the least squares line indicates that (0, - 0.7374) lies on the least-squares line. This means that a fish that is 0 mm in length will have a mercury concentration level of -0.7374 μg/g. It does not make sense for a fish to have zero length or a negative level of mercury concentration.

4. a. We used the equation of the least-squares line with constants rounded to four decimals: $y = 0.0032x - 0.7374$.

The prediction of mercury concentration is 0.0032(430) - 0.7374 = 0.6386 μg/g.

This prediction is an example of interpolation since x = 430 mm is between 315 mm and 490 mm, the range of the fish lengths in the observed data.

b. The prediction is 0.0032(90) - 0.7374 = -0.4494 μg/g.

This is an example of extrapolation. The length of the fish is far below the length of the smallest fish represented in the data. Furthermore, the sample of fish in the observed data were all of legal/edible size and this fish is not of legal/edible size.

5. a. SSE = $(0)^2 + (3)^2 + (0)^2 + (-2)^2 = 13$

b. SSE = $(-1.6)^2 + (2.3)^2 + (0.2)^2 + (-0.9)^2 = 8.7$

c. The least-squares line is the line that has the smallest SSE of all lines.