Chapter 6
Exploring Data: Relationships

Chapter Outline

Chapter Summary

A useful device for examining the relationship between two variables is the scatterplot, consisting of points in the plane. These points represent pairs of values for the variables in question, one variable being plotted along the x-axis and the other along the y-axis. Often, a scatterplot suggests a linear relation between the two variables; i.e., the points look as though they may be scattered about a line in the plane. Such a line is called a regression line, which is often used to predict the value of y for a given value of x. There are several ways of producing regression lines, depending on the method used to measure the discrepancy between the line and the data. The most common of these lines is the *least-squares regression line*, which minimizes the sum of the squares of the vertical distances between the data and the line.

However, the least-squares regression line exists for any set of data, even one that does not follow a linear pattern. Hence, there are situations in which it is not appropriate to compute this line nor to use it in predicting values of y. One way to determine whether the relationship between x and y is linear is through *correlation*, denoted by r, which measures the direction and strength of the linear relationship between two variables. r is always a number between -1 and 1. Values of r close to 0 indicate a very weak linear relationship; but if $|r|$ is close to 1, then the linear relationship is strong, with the slope of the line determined by the sign of r.

Skill Objectives

1. Draw a scatterplot for a small data set consisting of pairs of numbers.
2. From a scatterplot, draw an estimated regression line.
3. Compute the correlation for a small data set consisting of pairs of numbers and understand the significance of the correlation between two variables.
4. Describe how the concept of distance is used in determining the least-squares regression line.
5. Compute the equation of the least-squares regression line for a small data set.
6. Understand correlation and regression describe relationships that need further interpretation because association does not imply causation and outliers have an effect on these relationships.

Teaching Tips

1. When discussing the rationale for the least-squares regression line, some students notice that the concept of distance used is not the ordinary geometric distance from a point to a line, but rather vertical distance. Demonstrating the difference between the complexity of the algebraic expressions for the two makes the choice obvious. In addition, commenting on the reason for squaring the distance provides a nice review of the subtraction of signed numbers.

2. As students attempt to draw a scatterplot of data given from observations, they have a good opportunity to review the concepts of independent and dependent variable and functional notation.

3. When trying to draw a regression line by hand, it should be observed that a rule of thumb is that about half of the data points should lie above the line and half should lie below.

4. Depending on which forms of technology you choose to implement in the discussion of these topics, you may choose to remind students to be very careful when they input data. If one piece of data is not correctly placed in the calculator or spreadsheet, all calculations and displays will be inaccurate.

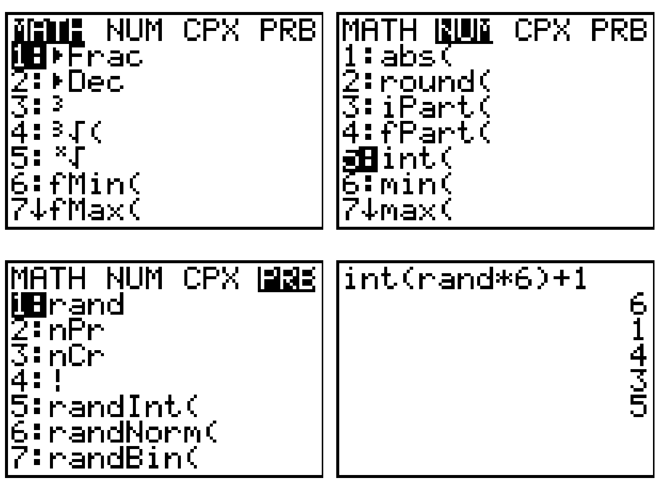
Research Paper

In the early 1800's, Karl Friedrich Gauss (1777–1855) introduced the procedure for obtaining least-squares estimates. This German born mathematician, who is also sometimes called the "prince of mathematics," has played a very important role in mathematics and statistics. Students can research the life of Gauss and his contributions to mathematics and statistics. Also, the life and contributions of the Scottish born George Undy Yule (1871–1951) can be suggested as a topic for a paper.

Spreadsheet Project

To do this project, go to http://www.whfreeman.com/fapp7e.

Spreadsheets are used in this project to analyze mean, standard deviation (Chapter 5) and the least –squares line. You will need two dice. If dice are not available, a calculator such as a TI-83 can simulate a toss of a die. The following screens will guide you to simulating the toss of a die.



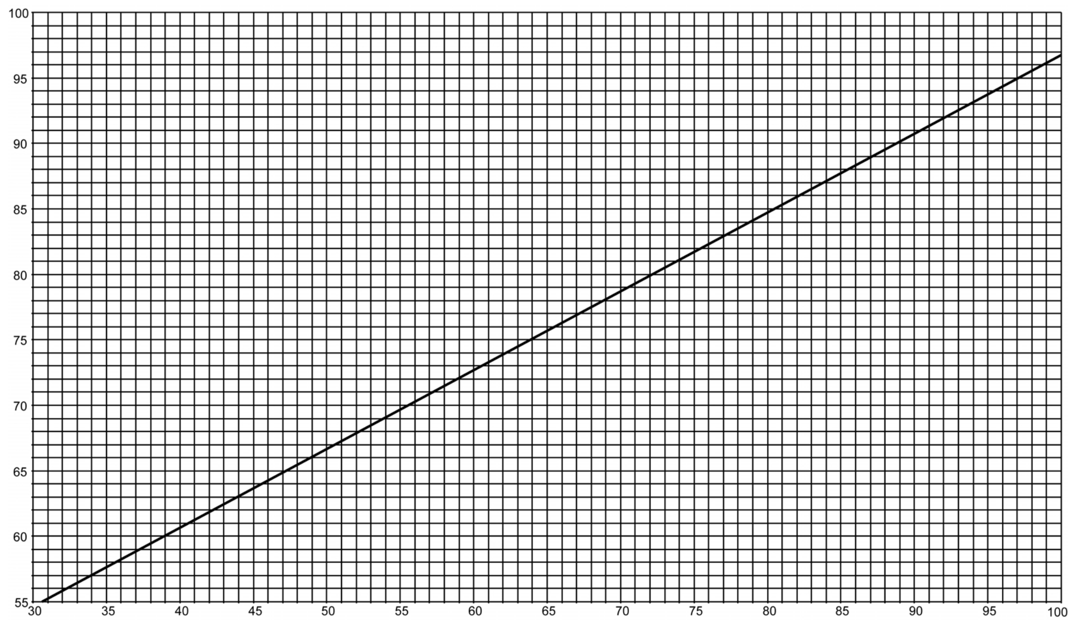
Collaborative Learning

Correlation Calculation

Professor Amanda Nunley gives a pop quiz before each exam in her algebra course. She has collected the following data regarding the average of the quizzes and the average of the exams for her ten students.

	Dean M.	John S.	Nadia A.	Denis P.	Scott H.	Dan D.	Joanne P.	Kevin L.	Adam A.	Phil P.
Quiz Average	52	86	72	35	90	92	85	62	54	77
Exam Average	63	91	83	60	89	95	87	65	77	80

1. Draw a scatterplot by treating the quiz average as the explanatory variable and the exam average as the response variable. The regression line has already been drawn in.



2. Determine how many units are between each point and the regression line. This distance should be determined by counting the number of units (points) above or below the line to each point. If a point is above the line, consider the distance has a sign of positive. If the point is below the line, consider the distance as a sign of negative.

3. Find the sum of all 10 “distances” you found in Part 2. What observation do you make about the sum? What observations can you make about the number of points above or below the line?

4. Determine the slope of the regression line. Give an interpretation of the meaning of the slope of the regression line for average quiz scores verses average exam scores

5. Determine the *y*-intercept of this line. Give an interpretation of the meaning of the *y*-intercept of the regression line for average quiz scores verses average exam scores.

6. Fill in the following table of values.

	x	y	x^2	y^2	xy
	52	63			
	86	91			
	72	83			
	35	60			
	90	89			
	92	95			
	85	87			
	62	65			
	54	77			
	77	80			
Sum	$\sum x =$	$\sum y =$	$\sum x^2 =$	$\sum y^2 =$	$\sum xy =$

7. Another way to calculate correlation is to use the following formula.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Using the calculations from Part 6 (where $n = 10$), determine the correlation.

8. Interpret the meaning of the correlation in terms of average quiz scores verses average exam scores

Solutions

Skills Check:

1. a	2. a	3. c	4. b	5. c	6. c	7. a	8. b	9. c	10. b
11. c	12. b	13. b	14. b	15. a	16. a	17. a	18. c	19. a	20. b

Exercises:

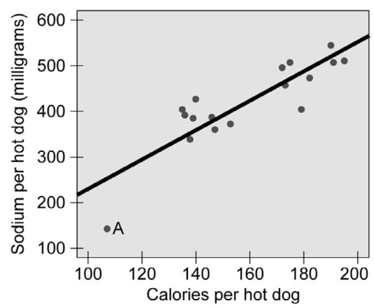
1. (a) It is more reasonable to explore study time as an explanatory variable and the exam grade as the response variable.

(b) It is more reasonable to explore the relationship only.

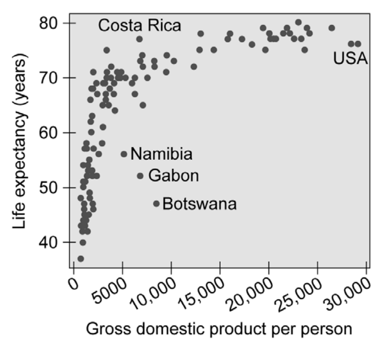
(c) It is more reasonable to explore rainfall as an explanatory variable and the corn yield as the response variable.

(d) It is more reasonable to explore the relationship only.

2. There is a moderately strong positive linear pattern, but the two clusters and the outlier marked A are at least as important. Brand A is much lower in both calories and sodium than any other brand. (In fact, it is the only brand made from veal and also weighs less than others.) The regression line has been drawn based on the data points.

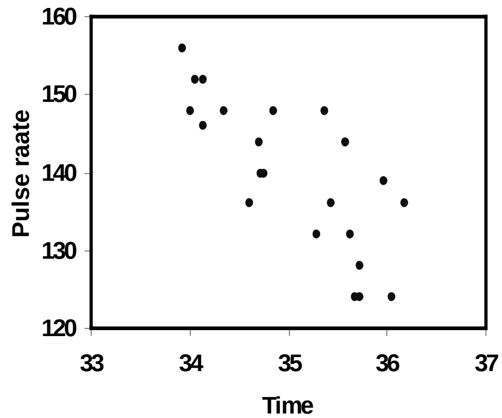


3. (a) Life expectancy increases with GDP in a curved pattern. The increase is very rapid at first, but levels off for GDP above roughly $5000 per person.

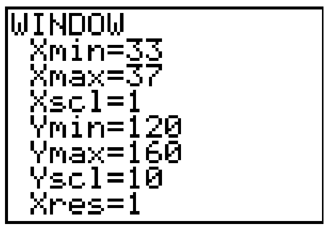


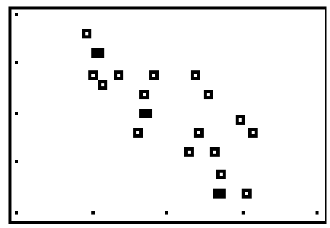
(b) Richer nations have better diets, clean water, and better health care, so we expect life expectancy to increase with wealth. But once food, clean water, and basic medical care are in place, greater wealth has only a small effect on lifespan.

4. (a) Here is the scatterplot, with time as the explanatory variable:



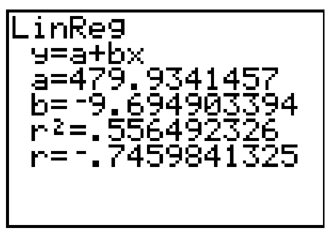
Using a TI-83, we get the following.



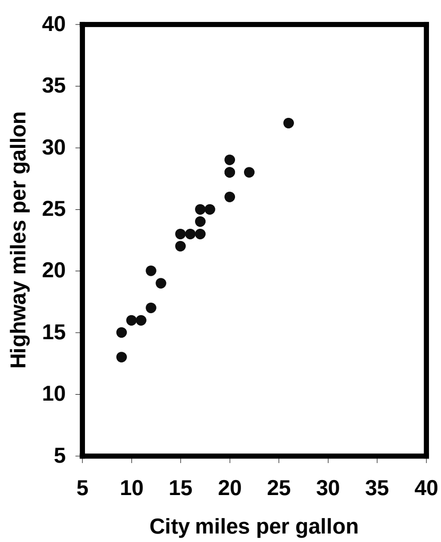


(b) Negative: We expect faster times (fewer minutes) to lead to higher pulse rates.

(c) Moderately strong negative linear relationship. (The correlation is $r = -0.746$.)

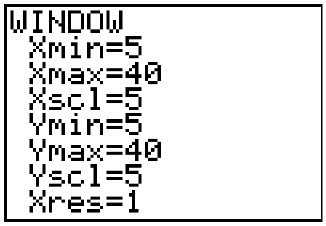


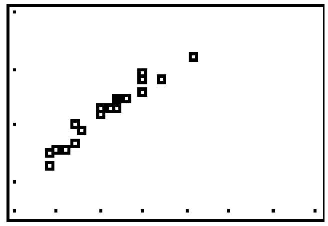
5. (a) The scatterplot is as follows.



Continued on next page

5. (a) continued
Using a TI-83, we get the following.

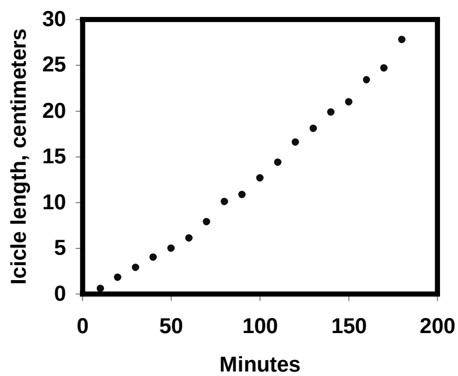




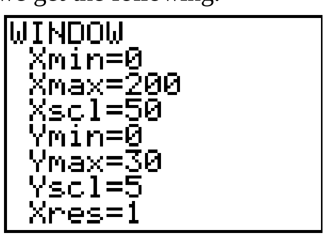
Purists should notice that because the variables measure similar quantities the plot is square with the same scales on both axes.

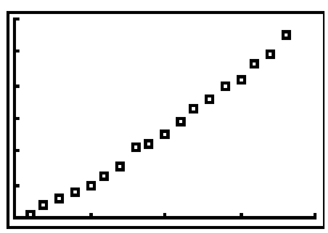
(b) There is a strong positive straight-line relationship.

6. The scatterplot, with time as the explanatory variable, is



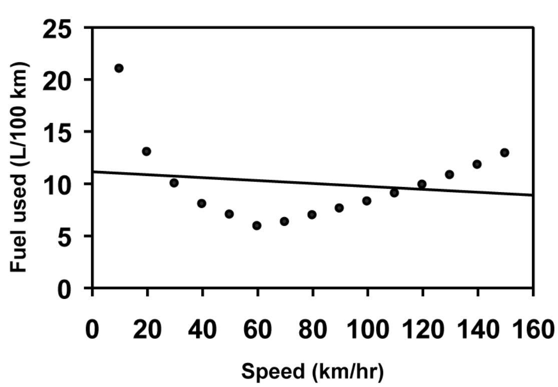
Using a TI-83, we get the following.





There is an extremely strong positive linear relationship.

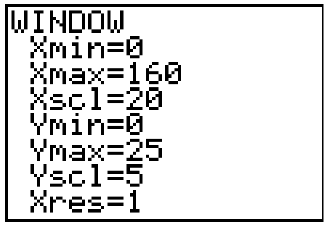
7. (a) Here is the scatterplot, with speed as the explanatory variable:

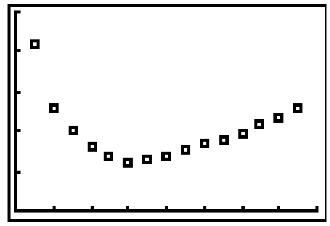


Continued on next page

7. (a) continued

Using a TI-83, we get the following.





(b) The relationship is curved; fuel usage first decreases as speed increases (higher gears cover more distance per motor revolution) then increases as speed is further increased (air resistance builds at higher speeds).

(c) There is no overall direction.

(d) The relationship is quite strong. There is little scatter about the overall curved pattern.

8. Answers will vary.

Positive: years of schooling and income (for adults).

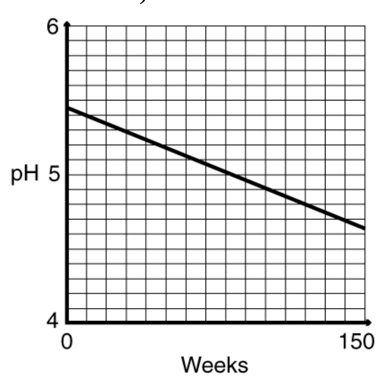
Negative: age and speed to walk or run a mile (for adults).

9. The estimated slope would be $\frac{506-386}{191-139}=\frac{120}{52}\approx 2.31$.

10. (a) Slope $=-9.695$. For each additional minute, pulse rate decreases by about 9.7 beats per minute on average.

(b) Since pulse $=479.9-(9.695\times 34.30)=479.9-332.5385=147.3615$, the prediction of approximately 147.4 beats per minute is low by $152-147.4=4.6$ beats per minute.

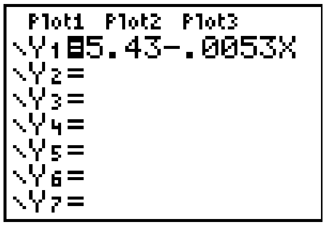
11. (a) Choose two values of weeks, preferably near 1 and 150. Find pH from the equation given, plot the two points (weeks horizontal) and draw the line through them.

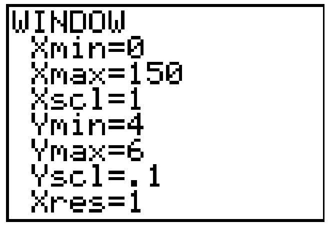


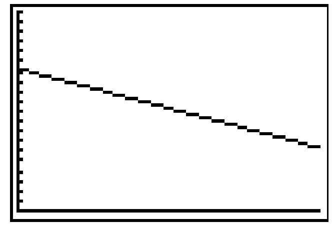
Continued on next page

11. (a) continued

Using a TI-83, we get the following.







(b) Week 1: predicted pH $= 5.43 - (0.0053 \times 1) = 5.43 - 0.0053 = 5.4247 \approx 5.42$

Week 150: predicted pH $= 5.43 - (0.0053 \times 150) = 5.43 - 0.795 = 4.635 \approx 4.64$

(c) The slope -0.0053 says that on average pH declined by 0.0053 per week during the study period.

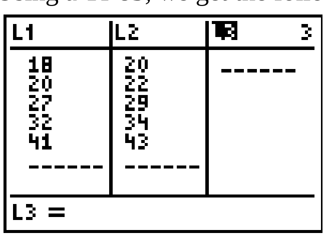
12. (a) Slope $= -2.02$. This indicates as the percent taking increases, predicted SAT score decreases.

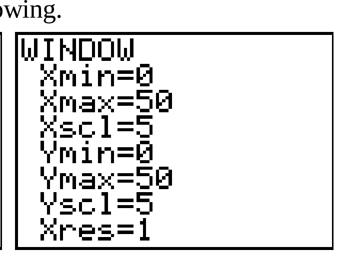
(b) Since predicted SAT score $= 1150 - (2.02 \times 82) = 1150 - 165.64 = 984.36$, the predicted score is 984.

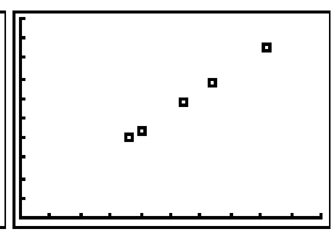
13. Some sample ages would be as follows.

Age of Woman	Age of Man
18	20
20	22
27	29
32	34
41	43

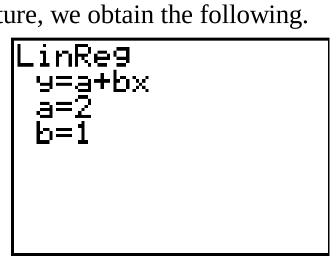
Using a TI-83, we get the following.







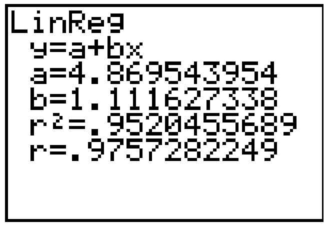
Using the linear regression feature, we obtain the following.



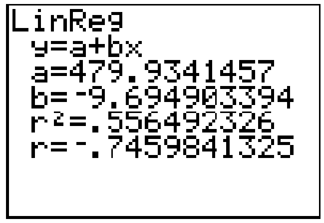
The line is $y = x + 2$, and the slope would be 1.

14. Since $b = 1.1$ kilograms per additional centimeter of height, we have $b = 1.1 \times 1000 = 1100$ grams per additional centimeter of height.

15. With a correlation of 0.9757, the indication is a very strong straight-line pattern.

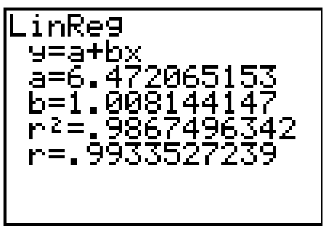


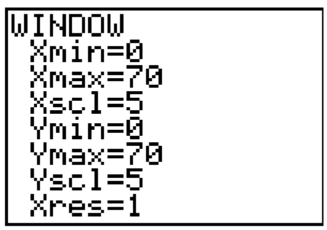
16. (a) With a correlation of –0.746, the indication is a moderately strong negative linear relationship.

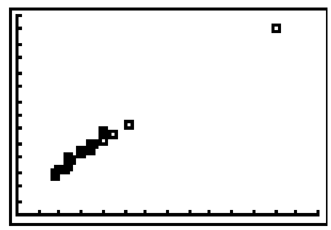


(b) The correlation, *r*, would not change.

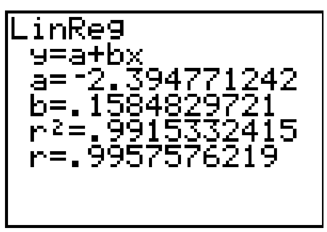
17. The correlation is 0.9934. The correlation is stronger when the Insight is added, because that point extends (strengthens) the straight-line pattern.



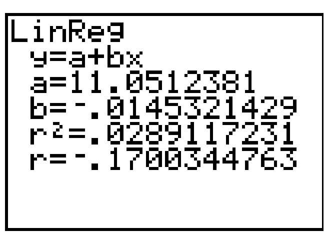




18. The correlation is 0.9958. This matches the close-to-perfect straight-line pattern of the scatterplot.



19. See the answer to Exercise 7 for the scatterplot. The correlation is –0.1700. Correlation measures the strength of only linear (straight-line) relationship. This relationship is strong but curved.



20. Correlation does not change when units of measurement change.

21. The correlation would be 1 because there is a perfect straight-line relationship, $y = x + 2$. This is verified by using the sample data from Exercise 13.

```
LinReg
 y=a+bx
 a=2
 b=1
 r²=1
 r=1
```

22. Think about which relationships are stronger and determine accordingly.

(a) r = 0.5, heredity partly determines height of son;

(b) r = 0.2, weaker relationship than (a) and (c);

(c) r = 0.8, height of same person.

23. (a) Negative: older cars will in general sell for lower prices.

(b) Negative: heavier cars will (other things being equal) get fewer miles per gallon.

(c) Positive: taller people are on average heavier than shorter people.

(d) Small: there is no reason to expect that height and IQ are related.

24. (a) Correlation does not make sense for gender, which does not have meaningful numerical values.

(b) r is always between -1 and 1, so r = 1.09 is impossible.

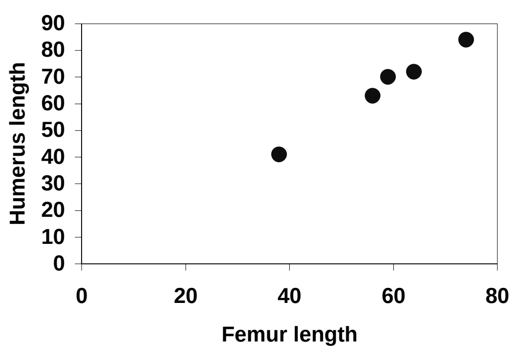
(c) r has no units.

25. Ask how similar the market sector of each fund is to large U.S. stocks and arrange in order.

(a) Dividend Growth, r = 0.98; Small Cap Stock, r = 0.81; Emerging Markets, r = 0.35.

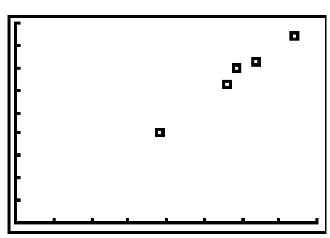
(b) No: it just says that they tend to move in the same direction, whether up or down.

26. (a) The plot shows a strong linear pattern, so we think all come from one species.



Using a TI-83, we get the following.

```
WINDOW
 Xmin=0
 Xmax=80
 Xscl=10
 Ymin=0
 Ymax=90
 Yscl=10
 Xres=1
```



Continued on next page

26. continued

(b) *Step 1*: We have for the femur $\bar{x} = 58.2,\ s_x = 13.20$ and for humerus $\bar{y} = 66,\ s_y = 15.89$.

```
1-Var Stats
 x̄=58.2
 Σx=291
 Σx²=17633
 Sx=13.19848476
 σx=11.80508365
↓n=5
```

```
1-Var Stats
 x̄=66
 Σx=330
 Σx²=22790
 Sx=15.89024858
 σx=14.2126704
↓n=5
```

Step 2: The terms in the sum for *r* are as follows.

$$\frac{38-58.2}{13.20}\times\frac{41-66}{15.89},\ \frac{56-58.2}{13.20}\times\frac{63-66}{15.89},\ \frac{59-58.2}{13.20}\times\frac{70-66}{15.89},$$
$$\frac{64-58.2}{13.20}\times\frac{72-66}{15.89}, \text{ and } \frac{74-58.2}{13.20}\times\frac{84-66}{15.89}$$

Step 3: Substituting into the formula for *r* we have the following.

$$r = \frac{1}{5-1}\left[\frac{38-58.2}{13.20}\times\frac{41-66}{15.89}+\frac{56-58.2}{13.20}\times\frac{63-66}{15.89}+\frac{59-58.2}{13.20}\times\frac{70-66}{15.89}\right.$$
$$\left.+\frac{64-58.2}{13.20}\times\frac{72-66}{15.89}+\frac{74-58.2}{13.20}\times\frac{84-66}{15.89}\right]$$
$$= \frac{1}{4}\left[\frac{-20.2}{13.20}\times\frac{-25}{15.89}+\frac{-2.2}{13.20}\times\frac{-3}{15.89}+\frac{0.8}{13.20}\times\frac{4}{15.89}\right.$$
$$\left.+\frac{5.8}{13.20}\times\frac{6}{15.89}+\frac{15.8}{13.20}\times\frac{18}{15.89}\right]$$
$$= \frac{1}{4}\left[\frac{505}{209.748}+\frac{6.6}{209.748}+\frac{3.2}{209.748}+\frac{34.8}{209.748}+\frac{284.4}{209.748}\right]$$
$$= \frac{1}{4}\left[\frac{834}{209.748}\right] = \frac{834}{838.992} \approx 0.994$$

(c) The two approximations, 0.994, match to three decimal places.

```
LinReg
 y=a+bx
 a=-3.659586682
 b=1.196900115
 r²=.9883313819
 r=.9941485714
```

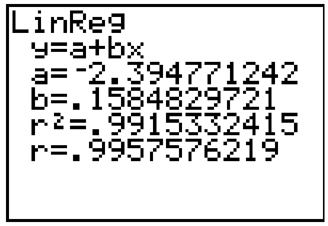
27. (a) Predicted MPG $= 4.87 + 1.11x$.

```
LinReg
 y=a+bx
 a=4.869543954
 b=1.111627338
 r²=.9520455689
 r=.9757282249
```

(b) Predicted MPG $= 4.87 + 1.11(18) = 4.87 + 19.98 = 24.85$ MPG.

(c) We assess accuracy from how closely the points in the plot follow a straight line. Looking at the plot in Exercise 5, we expect quite accurate predictions.

28. (a) Predicted length $= -2.395 + 0.158x$.

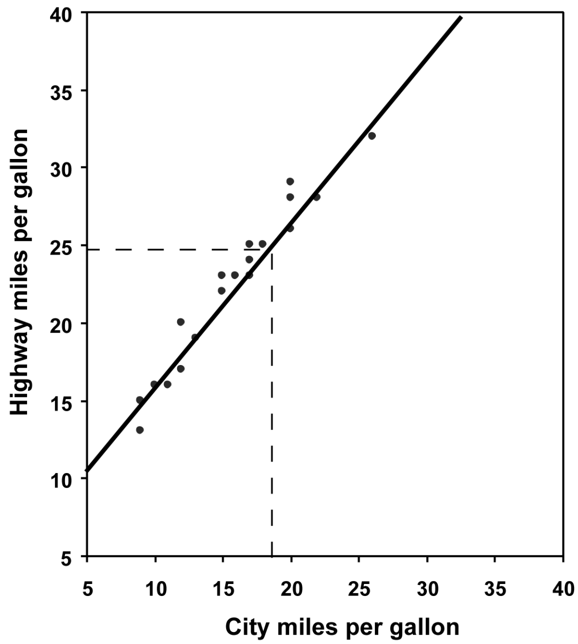


(b) Predicted length $= -2.395 + 0.158(75) = -2.395 + 11.85 = 9.455 \approx 9.46$ cm.

29. Choose two city MPG values, such as x = 10 and x = 30, and use the equation of the line to find each value of y. Plot the two points and draw the line between them. Here is the plot.

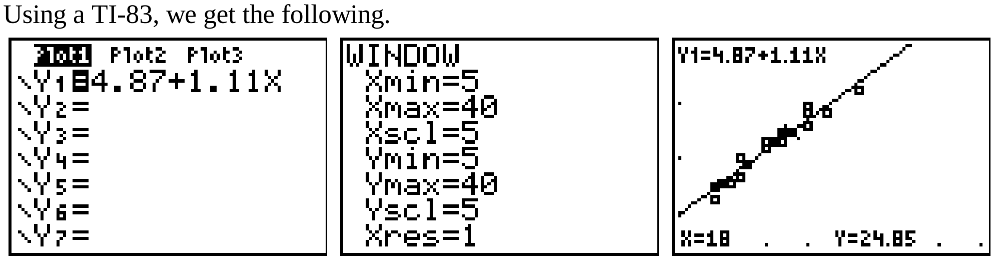
x = 10: Predicted MPG $= 4.87 + 1.11(10) = 4.87 + 11.1 = 15.97$ MPG.

x = 30: Predicted MPG $= 4.87 + 1.11(30) = 4.87 + 33.3 = 38.17$ MPG.



Predicted highway mileage of a car that gets 18 mpg in the city is approximately 25 mpg (24.85 mph) from exercise 27.

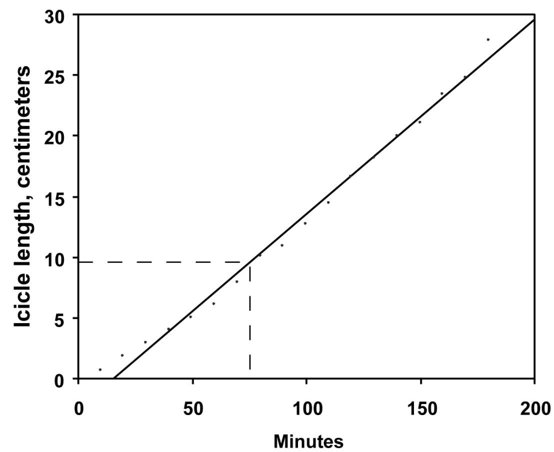
Using a TI-83, we get the following.



30. Choose two times, such as $x = 50$ and $x = 150$, use the equation of the line to find each value of y. Plot the two points and draw the line between them. Here is the plot.

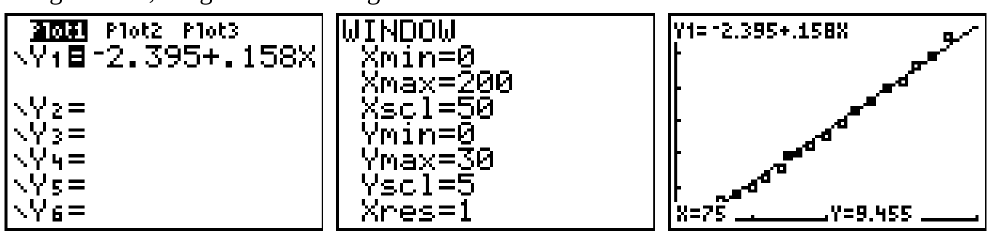
$x = 50$: Predicted length $= -2.395 + 0.158(50) = -2.395 + 7.9 = 5.505$ cm.

$x = 150$: Predicted length $= -2.395 + 0.158(150) = -2.395 + 23.7 = 21.305$ cm.



Predicted length of icicle is approximately 9.5 cm (9.46 cm) from exercise 28.

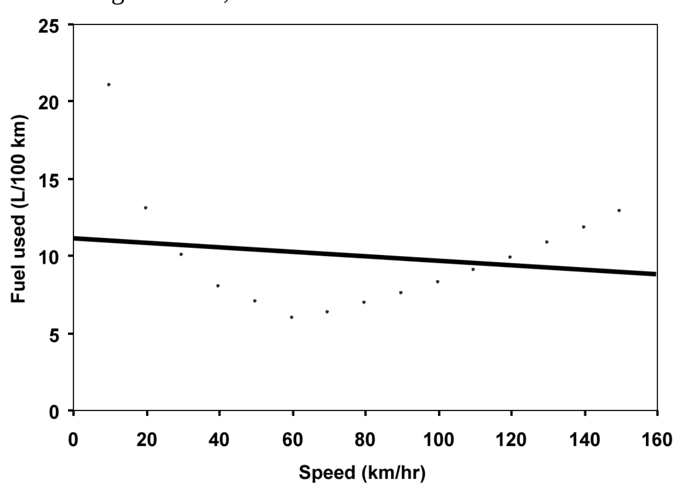
Using a TI-83, we get the following.



31. Since predicted fuel $= 11.058 - 0.0147 \times \text{speed}$, we have the following.

Speed = 10 kph: predicted fuel $= 11.058 - 0.0147 \times 10 = 11.058 - 0.147 = 10.911$ kpg
Speed = 70 kph: predicted fuel $= 11.058 - 0.0147 \times 70 = 11.058 - 1.029 = 10.029$ kpg
Speed = 150 kph: predicted fuel $= 11.058 - 0.0147 \times 150 = 11.058 - 2.205 = 8.853$ kpg

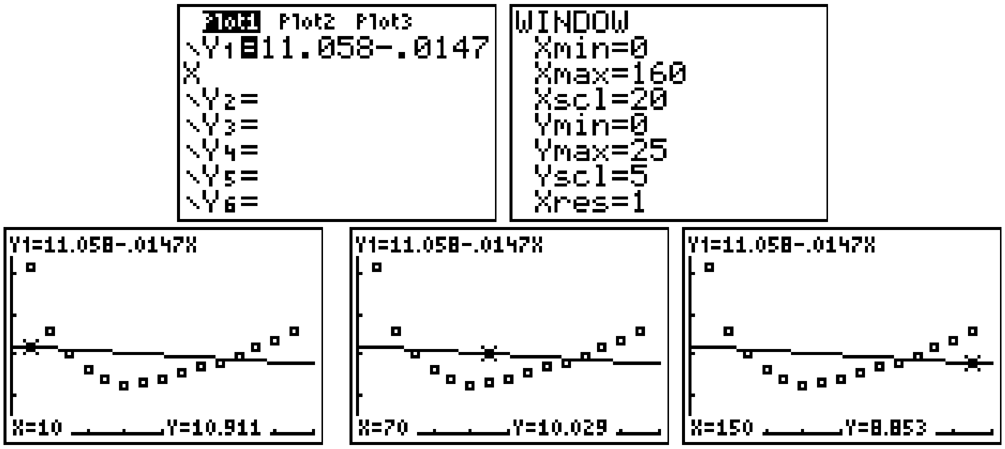
The predicted values from the equation given are approximately 10.9, 10.0, and 8.85, respectively. The observed values are 21.00, 6.30, and 12.88, respectively. The least-squares line gives the best straight-line fit, which is of little value here.



Continued on next page

31. continued

Using a TI-83, we get the following.



32. Slope b = 0.158 centimeter per additional minute of water flow. This is the same as $\frac{0.158}{2.54} \approx 0.062$ inch per minute.

33. The slope of the least-squares line is $b = r\frac{s_y}{s_x} = 0.5\left(\frac{2.7}{2.5}\right) = 0.54.$ The intercept is as follows.

$$a = \bar{y} - b\bar{x} = 68.5 - (0.54)(64.5) = 68.5 - 34.83 = 33.67$$

For x = 67, we predict $33.67 + (0.54)(67) = 33.67 + 36.18 = 69.85$ inches.

34. (a) Slope is $b = r\frac{s_y}{s_x} = 0.6\left(\frac{8}{30}\right) = 0.16.$ The intercept is as follows.

$$a = \bar{y} - b\bar{x} = 75 - (0.16)(280) = 75 - 44.8 = 30.2$$

(b) For x = 300, we predict $30.2 + (0.16)(300) = 30.2 + 48 = 78.2$, or 78 was Julie's exam score.

35. The predicted y for $x = \bar{x}$ is as follows.

$$\text{predicted } y = a + b\bar{x} = (\bar{y} - b\bar{x}) + \left(r\frac{s_y}{s_x}\right)\bar{x} = \bar{y} - \left(r\frac{s_y}{s_x}\right)\bar{x} + \left(r\frac{s_y}{s_x}\right)\bar{x} = \bar{y}$$

36. (a) For x = time and y = length, x = 95, s_x = 53.3854, $\bar{y}$ = 12.6611, s_y = 8.4967 and r = 0.9958. The regression line has slope $b = r\frac{s_y}{s_x} = 0.9958\left(\frac{8.4967}{53.3854}\right) \approx 0.1585$ and intercept $a = \bar{y} - b\bar{x} = 12.6611 - (0.1585)(95) = 12.6611 - 15.0575 = 2.3964.$ These results agree up to roundoff error with those in Exercise 28. Carrying fewer places in intermediate steps will increase the roundoff error.

(b) Reverse the roles of x and y. The slope is now $b = r\frac{s_y}{s_x} = 0.9958\left(\frac{53.3854}{8.4967}\right) \approx 6.2567,$ and the intercept is as follows.

$$a = \bar{y} - b\bar{x} = 95 - (6.2567)(12.6611) = 95 - 79.21670437 = 15.78329563 \approx 15.78$$

If length x = 15, we predict time to be as follows.

$$15.78 + (6.2567)(15) = 15.78 + 93.8505 = 109.6305 \approx 109.6 \text{ minutes}$$

Look at the plot in Exercise 6 to see that this is reasonable.

37. First compare the distributions for the two years. To make the boxplots, we need the five-number summary for each data set.

Putting the 2002 data in order, we have the following.

$$-50.5,\ -49.5,\ -47.8,\ -42,\ -37.8,\ -26.9,\ -23.4,\ -21.1,\ -18.9,\ -17.2,\ -17.1,$$
$$-12.8,\ -11.7,\ -11.5,\ -11.4,\ -9.6,\ -7.7,\ -6.7,\ -5.6,\ -2.3,\ -0.7,\ -0.7,\ 64.3$$

The minimum is -50.5 and the maximum is 64.3. The median is the $\frac{23+1}{2}=\frac{24}{2}=12^{\text{th}}$ piece of data, namely -12.8. Since there are 11 observations to the left of the median, Q_1 is the $\frac{11+1}{2}=\frac{12}{2}=6^{\text{th}}$ piece of data, namely -26.9. Since there are 11 observations to the right of the median, Q_3 is the $12+6=18^{\text{th}}$ piece of data, namely -6.7.

Thus, the five - number summary is $-50.5,\ -26.9,\ -12.8,\ -6.7,\ 64.3$.

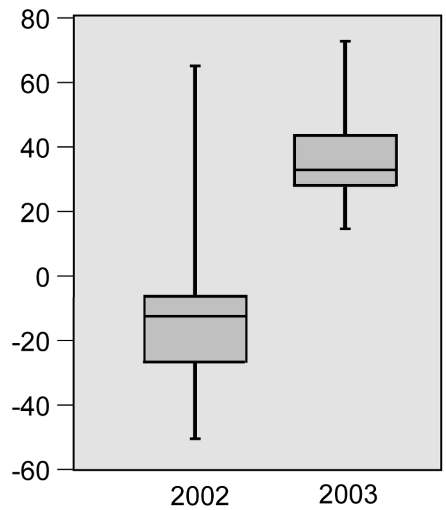
Putting the 2003 data in order, we have the following.

$$14.1, 19.1, 22.9, 23.9, 26.1, 27.5, 28.7, 29.5, 30.6, 31.1, 32.1,$$
$$32.3, 35.0, 36.5, 36.9, 36.9, 41.8, 43.9, 57.0, 59.4, 62.7, 68.1, 71.9$$

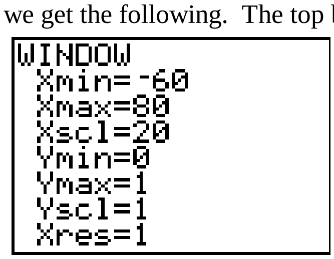
The minimum is 14.1 and the maximum is 71.9. The median is the $\frac{23+1}{2}=\frac{24}{2}=12^{\text{th}}$ piece of data, namely 32.3. Since there are 11 observations to the left of the median, Q_1 is the $\frac{11+1}{2}=\frac{12}{2}=6^{\text{th}}$ piece of data, namely 27.5. Since there are 11 observations to the right of the median, Q_3 is the $12+6=18^{\text{th}}$ piece of data, namely 43.9.

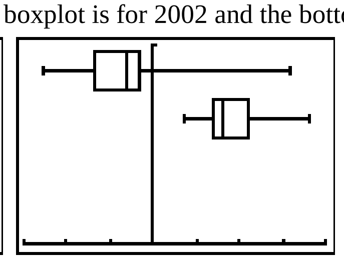
Thus, the five - number summary is 14.1, 27.5, 32.3, 43.9, 71.9.

Here are boxplots.



Using a TI-83, we get the following. The top boxplot is for 2002 and the bottom one is for 2003.



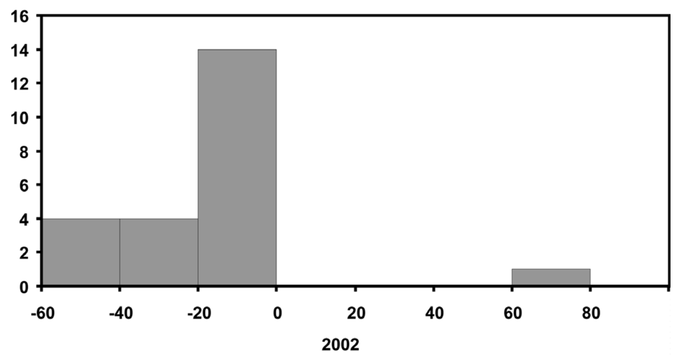


Continued on next page

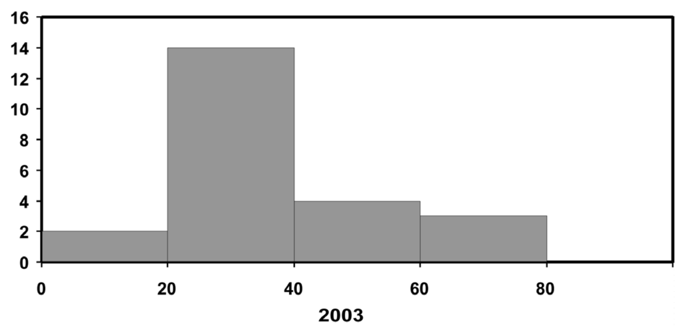
37. continued

Histograms (or stemplots) show that the 2003 returns are roughly single-peaked and symmetric, and that the 2002 returns are left-skewed with an extreme high outlier. Below are the histograms.

Class	Count
– 60 – (– 40.1)	4
– 40 – (– 20.1)	4
– 20 – (– 0.1)	14
0 – 19.9	0
20 – 39.9	0
40 – 59.9	0
60 – 79.9	1

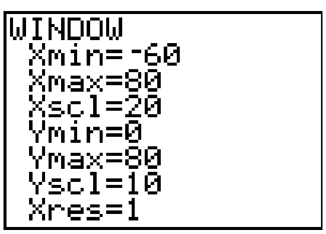


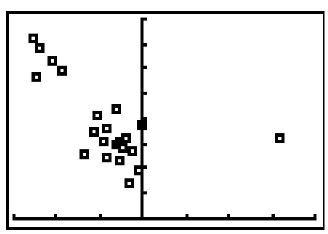
Class	Count
0 – 19.9	2
20 – 39.9	14
40 – 59.9	4
60 – 79.9	3

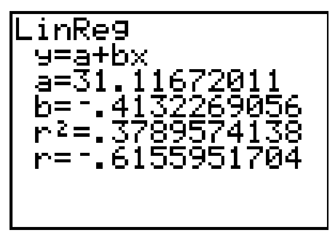


The median returns are – 12.8% in 2002 and 32.1% in 2003 (from the five - number summary). The correlation is $r = -0.616$; because of the influence of outliers on correlation, it is better to report the correlation without the outlier, $r = -0.838$.

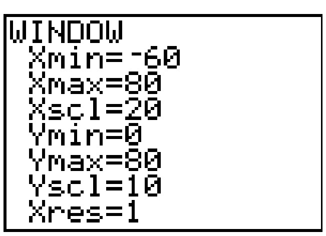
With Outlier

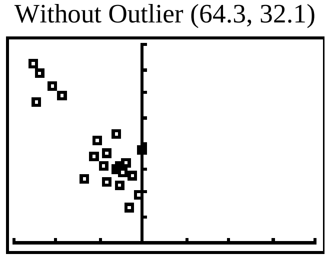


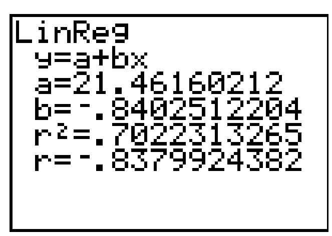




Without Outlier (64.3, 32.1)

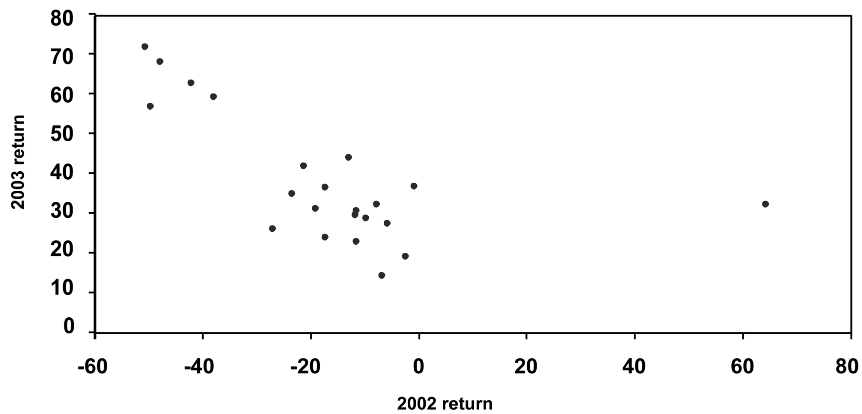






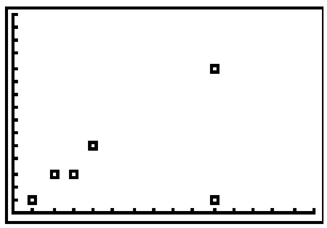
Continued on next page

37. continued
That is, the funds that went down most in 2002 tended to go up most in 2003. The scatterplot confirms this:



38. We have r = 0.481. The scatterplot shows that five points lie close to a line, but the point (10,1) lies far from the line at the lower right. This outlier reduces r.

```
WINDOW
 Xmin=0
 Xmax=15
 Xscl=1
 Ymin=0
 Ymax=15
 Yscl=1
 Xres=1
```



```
LinReg
 y=a+bx
 a=1.75
 b=.45
 r²=.2314285714
 r=.4810702354
```

39. (a) All four sets of data have r = 0.816 and regression line $y = 3.0 + 0.5x$ to a close approximation.

Set A

```
LinReg
 y=a+bx
 a=3.000090909
 b=.5000909091
 r²=.6665424595
 r=.8164205163
```

Set B

```
LinReg
 y=a+bx
 a=3.000909091
 b=.5
 r²=.6662420337
 r=.816236506
```

Set C

```
LinReg
 y=a+bx
 a=3.002454545
 b=.4997272727
 r²=.6663240411
 r=.8162867395
```

Set D

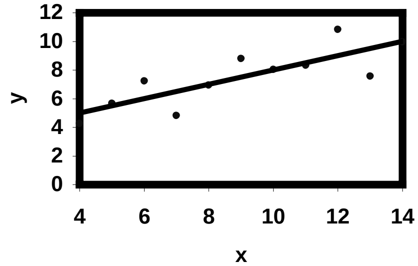
```
LinReg
 y=a+bx
 a=3.001727273
 b=.4999090909
 r²=.6667072569
 r=.8165214369
```

Continued on next page

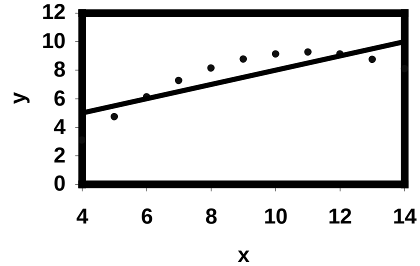
39. continued

(b) Here are the plots:

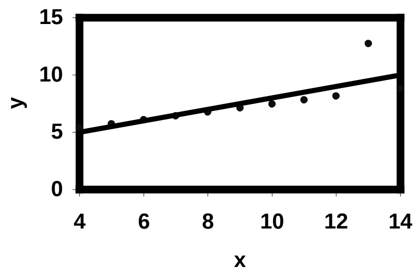
Data Set A



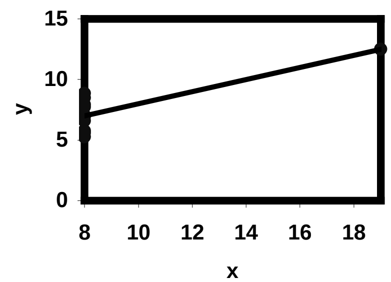
Data Set B



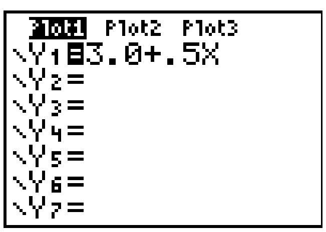
Data Set C

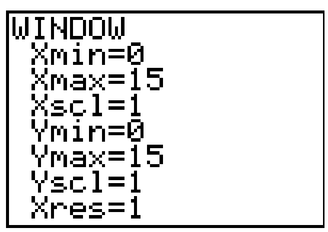


Data Set D

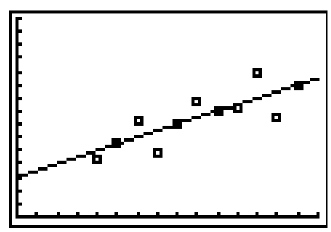


Using a TI-83, we get the following.

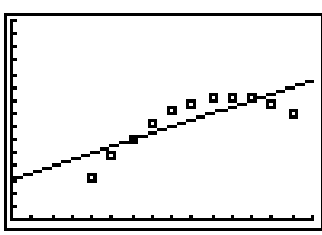




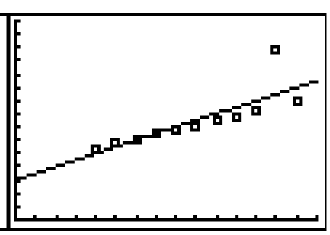
Set A



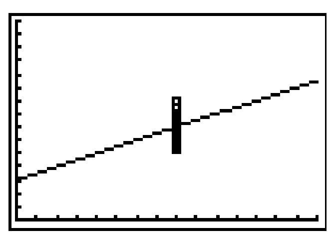
Set B



Set C



Set D



(c) Only A is a normal" regression setting in which the line is useful for prediction. Plot B is curved, C has an extreme outlier in *y*, and D has all but one *x* identical. The lesson: Plot your data before calculating.

In Exercises 40 – 44, answers will vary.

40. Children who watch lots of TV may lack parental supervision. Such children would study less and read less and take part in fewer outside activities.

41. Heavier people who are concerned about their weight may be more likely than lighter people to choose artificial sweeteners in place of sugar.

42. Companies with many employees tend to pay their CEOs more than smaller companies. If they lay off (say) 5% of their employees, the number laid off is greater than 5% for smaller companies. So CEO pay and numbers laid off both tend to go up with company size.

43. Higher income generally means better water and sewage utilities, better diet, and better medical care, which will produce better health. But better health means more children can go to school and more workers are able to work and can stay on the job, which raises national income. For example, AIDS is having a direct negative effect on the economies of African nations.

44. Suppose that SAT scores roughly measure some combination of ability plus knowledge. Students in general are learning more in school, so SAT scores in general go up. But grade inflation means that A students are weaker on the average than they once were, so SAT scores for A students go down.

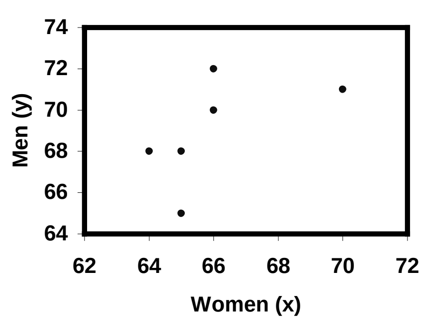
45. Explanatory: parents' income. Response: amount of college debt.
We expect a negative association: children of richer parents do not need to borrow as much to pay for college.

46. (a) IQ is supposed to measure "general problem-solving ability" and understanding written material is one kind of problem solving. The scatterplot does show a general lower-left to upper-right pattern, that is, a positive association.

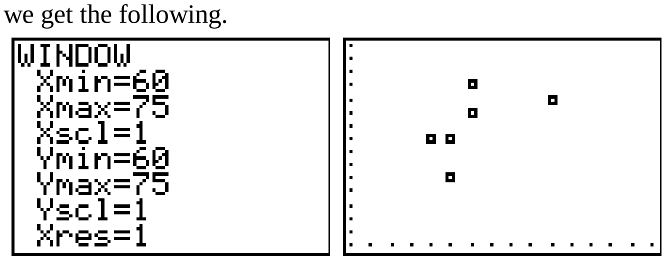
(b) There are four children with IQs roughly 108 to 125 who have much lower reading scores than other children with similar IQs.

(c) The lower-left to upper-right pattern is roughly a straight line, but the association is weak because there is a great deal of scatter above and below a line that would describe the overall pattern.

47. (a) There is a positive association, so r will be positive. The pattern is a bit irregular, so r won't be close to 1.



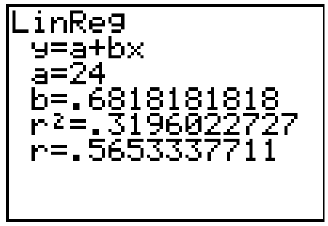
Using a TI-83, we get the following.



Continued on next page

47. continued

(b) $r = 0.5653$.



48. Correlation is not affected by changing the scale of one or both variables. Subtracting 3 inches from all male heights and changing from inches to centimeters are both changes of scale. So r is unchanged.

49. (a) The slope is $b = 0.68$. For each additional inch of women's height, the height of the next man dated goes up by 0.68 inch on average.

(b) The prediction is as follows.

predicted male height = 24 + 0.68 $\times$ female height

= 24+(0.68)(67) = 24 + 45.56 = 69.56 inches

Word Search Solution

```
E R V L E S A N G D Z F F Y M Q Z D E I F A E M H G
A E V E N F M T Z C H B E Q L N P X I R H N S R F R
E E O L P W H F E B Z H X O X A S M O F K N E F J F
E S M U W S N R R R T J G C E E R N M A A I H U H E
N I E D O L T F E F E T G J O E F R P S L L T A A X
N O I T A L E R R O C K A O T S R S F T D F E I M P
E N I L N O I S S E R G E R S E R A U Q S T S A E L
A S S T Y T C S S W N V R V T S Y O C L S M V P T A
O E X E A S K L E E X D U E O M R E B S P R D C C N
E O D I U I H H J K H P T I R E L L C I E I F A S A
T A S Q E E C T S O D T O E O I C Y R S G G V M C T
N O I T A I C O S S A E V I T I S O P Z I H R R X O
R O S N E G N O S D S A O I Y D L O S M E S R E E R
W B A D I P E L L S D N I P B N N W D T R Q N N S Y
G U E S H Y E A D I A N E O N S X G S R X Q I T S V
E X K R I N D T B A C E C F E H G I N V A H S I J A
L K S L Z R C E M T D X V V A T E I T E E B R A M R
C R E L L I I E S E E F A I O C M K T D M M T O S I
G T C E Q R Y T O L P R E T T A C S D E T T D W O A
N T E H N E C E P Y I U C G R A L T H P Y S W R P B
I P M C F A E O V A W T E I V U G K J T M F N W F L
Z E C R O S T M B C S C T P G D P E N F O R R R E E
Z R E V K E T L I H J O I Z E N T F N O O N I S V T
E E A E T E E Z S D A F N F L E X M V M A M F E E M
K W A S Z R U I I C T E R A A R X E O E M P M R L P
T C Z O M E N I L N O I S S E R G E R A I M N F Q R
```