12.6 Outliers

To determine if a point is an outlier, do one of the following:

1. Input the following equations into the TI 83, 83+,84, 84+:
   $y_1 = a + bx$
   $y_2 = a + bx + 2s$ where s is the standard deviation of the residuals
   $y_3 = a + bx - 2s$

   If any point is above y_2 or below y_3 then the point is considered to be an outlier.
2. Use the residuals and compare their absolute values to $2s$ where s is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to $2s$, then the corresponding point is an outlier.
3. Note: The calculator function LinRegTTest (STATS TESTS LinRegTTest) calculates s.

Formula Review

12.1 Linear Equations

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

12.4 Testing the Significance of the Correlation Coefficient

Least Squares Line or Line of Best Fit:

$\hat{y} = a + bx$

where

a = y-intercept

b = slope

Standard deviation of the residuals:

$s = \sqrt{\frac{SEE}{n-2}}$.

where

SSE = sum of squared errors

n = the number of data points

Practice

12.1 Linear Equations

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

1. What are the dependent and independent variables?

2. Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

3. Graph the equation from Exercise 12.2.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

4. Find the equation that expresses the total fee in terms of the number of days the payment is late.

5. Graph the equation from Exercise 12.4.

6. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

7. Which of the following equations are linear?
   a. $y = 6x + 8$
   b. $y + 7 = 3x$
   c. $y - x = 8x^2$
   d. $4y = 8$

8. Does the graph show a linear equation? Why or why not?

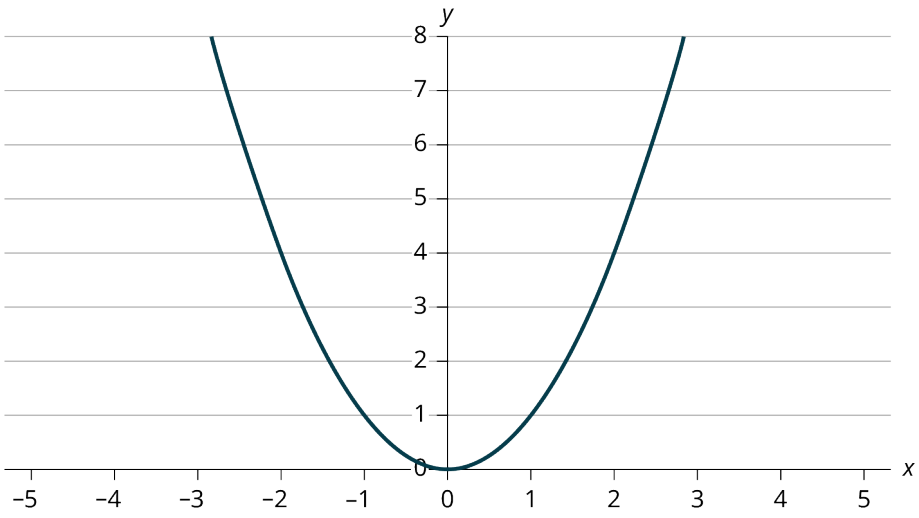

Figure 12.27

Table 12.13 contains real data for the first two decades of flu reporting.

Year	# flu cases diagnosed	# flu deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095

Table 12.13 Adults and Adolescents only, United States

1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.13 Adults and Adolescents only, United States

9. Use the columns "year" and "# flu cases diagnosed. Why is “year” the independent variable and “# flu cases diagnosed.” the dependent variable (instead of the reverse)?

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

10. What are the independent and dependent variables?

11. What is the y-intercept and what is the slope? Interpret them using complete sentences.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

12. What are the independent and dependent variables?

13. How many pounds of soil does the shoreline lose in a year?

14. What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

15. What are the slope and y-intercept? Interpret their meaning.

16. If you owned this stock, would you want a positive or negative slope? Why?

12.2 Scatter Plots

17. Does the scatter plot appear linear? Strong or weak? Positive or negative?

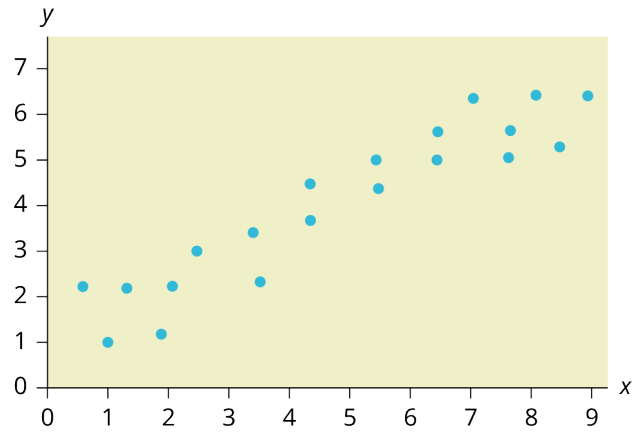

Figure 12.28

18. Does the scatter plot appear linear? Strong or weak? Positive or negative?

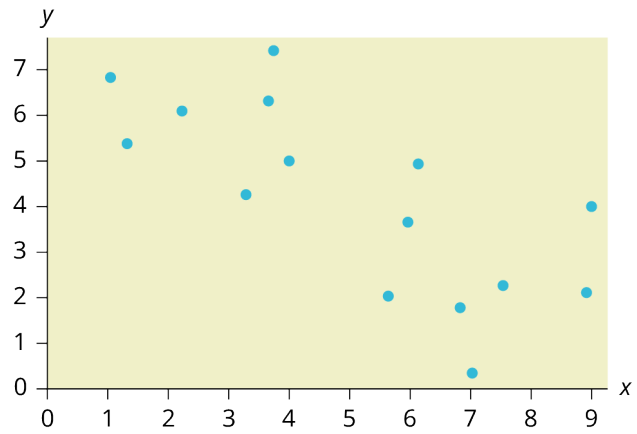

Figure 12.29

19. Does the scatter plot appear linear? Strong or weak? Positive or negative?

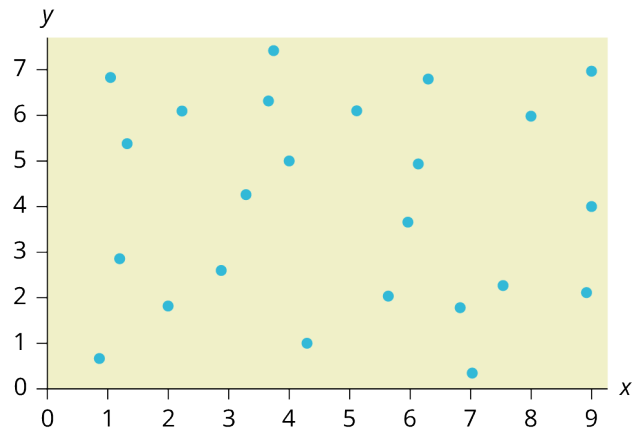

Figure 12.30

12.3 The Regression Equation

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Table 12.14

20. Draw a scatter plot of the data.

21. Use regression to find the equation for the line of best fit.

22. Draw the line of best fit on the scatter plot.

23. What is the slope of the line of best fit? What does it represent?

24. What is the y-intercept of the line of best fit? What does it represent?

25. What does an r value of zero mean?

26. When $n = 2$ and $r = 1$, are the data significant? Explain.

27. When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

12.4 Testing the Significance of the Correlation Coefficient

28. When testing the significance of the correlation coefficient, what is the null hypothesis?

29. When testing the significance of the correlation coefficient, what is the alternative hypothesis?

30. If the level of significance is 0.05 and the p-value is 0.04, what conclusion can you draw?

12.5 Prediction

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$\hat{y} = 101.32 + 2.48x$ where $\hat{y}$ is in thousands of dollars.

31. What would you predict the sales to be on day 60?

32. What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$\hat{y} = 1350 - 1.2x$ where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

33. How many acres will be left to mow after 20 hours of work?

34. How many acres will be left to mow after 100 hours of work?

35. How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

Table 12.15 contains real data for the first two decades of flu cases reporting.

Year	# flu cases diagnosed	# flu deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.15 Adults and Adolescents only, United States

36. Graph "year" versus "# flu cases diagnosed" (plot the scatter plot). Do not include pre-1981 data.

37. Perform linear regression. What is the linear equation? Round to the nearest whole number.

38. Find the correlation coefficient.
   a. r = _____

39. Solve.
   a. When $x = 1985$, $\hat{y}$ = _____
   b. When $x = 1990$, $\hat{y}$ =_____
   c. When $x = 1970$, $\hat{y}$ =_____ Why doesn't this answer make sense?

40. Does the line seem to fit the data? Why or why not?

41. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the U.S.?

42. Plot the two given points on the following graph. Then, connect the two points to form the regression line.

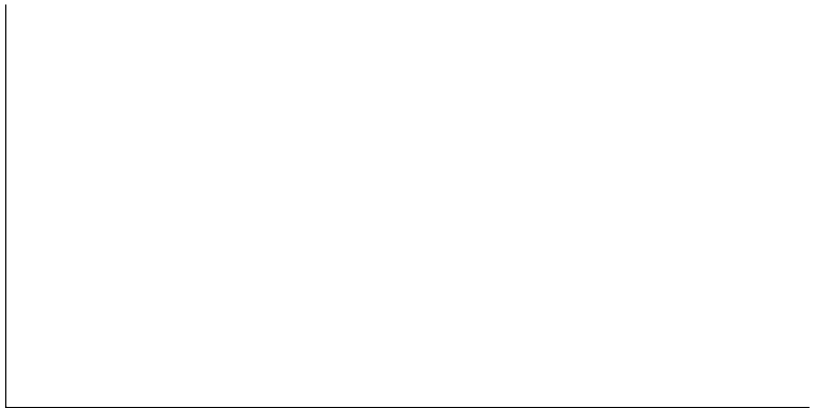

Figure 12.31

Obtain the graph on your calculator or computer.

43. Write the equation: $\hat{y}$= _____

44. Hand draw a smooth curve on the graph that shows the flow of the data.

45. Does the line seem to fit the data? Why or why not?

46. Do you think a linear fit is best? Why or why not?

47. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the U.S.?

48. Graph "year" vs. "# flu cases diagnosed." Do not include pre-1981. Label both axes with words. Scale both axes.

49. Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so? Write the linear equation, rounding to four decimal places:

50. Find the correlation coefficient.
   a. correlation = _____

12.6 Outliers

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

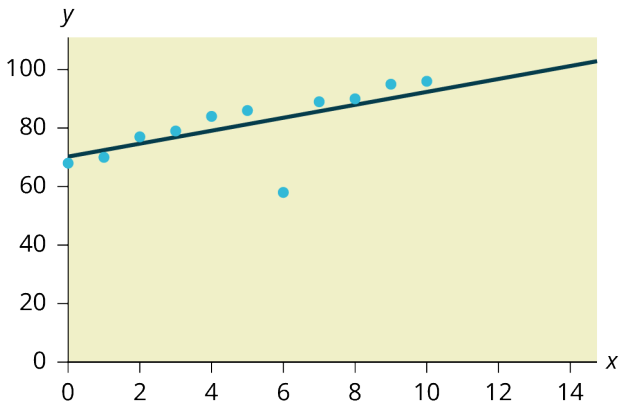

Figure 12.32

51. Do there appear to be any outliers?

52. A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

53. What effect did the potential outlier have on the line of best fit?

54. Are you more or less confident in the predictive ability of the new line of best fit?

55. The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?

56. The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

Homework

12.1 Linear Equations

57. For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

58. Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

Table 12.16

If a loan officer makes 95% of their goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

12.2 Scatter Plots

59. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. Table 12.17 shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1	1,700	8	4,000
2	1,700	9	11,000
4	2,300	10	9,500
5	2,900	11	9,700
6	3,000	12	9,900
7	3,500		

Table 12.17

60. The following table shows the number of faculty and number of students at several colleges. Construct a scatter plot of the data

College	Faculty	Students
College A	47	940
College B	58	1102
College C	26	533
College D	63	1244

Table 12.18

61. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	57,410
Harvey Mudd	135	62,817
CalTech	127	60,864
US Naval Academy	122	0
West Point	120	0

Table 12.19

School	Mid-Career Salary (in thousands)	Yearly Tuition
MIT	118	57,986
Lehigh University	118	59,930
NYU-Poly	117	58,168
Babson College	117	56,576
Stanford	114	56,169

Table 12.19

62. If the level of significance is 0.05 and the *p*-value is 0.06, what conclusion can you draw?

63. If there are 15 data points in a set of data, what is the number of degree of freedom?

12.3 The Regression Equation

64. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

65. Explain what it means when a correlation has an r^2 of 0.72.

66. Can a coefficient of determination be negative? Why or why not?

12.5 Prediction

67. Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

Table 12.20

a. For each age group, pick the midpoint of the interval for the *x* value. (For the 75+ group, use 80.)
b. Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 and 60.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least squares (best-fit) line? Interpret the slope.

68. Table 12.21 shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

Table 12.21

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in Table 12.21 that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

69. The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given in Table 12.22

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15

Table 12.22

Page number	Maximum value ($)
72	16
85	15
90	17

Table 12.22

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page ten and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
g. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

70. Table 12.23 gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1
2016	52.7

Table 12.23

a. Decide which variable should be the independent variable and which should be the dependent variable.

b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from part f different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

71.

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.24

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

12.6 Outliers

72. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

Table 12.25

a. Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data in Table 12.25, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

73. Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.
Percent return:74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38
Percent new:5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20
a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator’s regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

74. The following table shows data on average per capita coffee consumption and heart disease rate in a random sample of 10 countries.

Yearly coffee consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

Table 12.26

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of coffee consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

75. The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

Table 12.27

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

76. A researcher is investigating whether population impacts homicide rate. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are White men.

Population Size	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

Table 12.28

a. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context.
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation r
   iv. The coefficient of determination r2.
d. Do the data provide convincing evidence that there is a linear relationship between population size and homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

77.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	57,410
Harvey Mudd	135	62,817
CalTech	127	60,864
US Naval Academy	122	0

Table 12.29

School	Mid-Career Salary (in thousands)	Yearly Tuition
West Point	120	0
MIT	118	57,986
Lehigh University	118	59,930
NYU-Poly	117	58,168
Babson College	117	56,576
Stanford	114	56,169

Table 12.29

Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

Bringing It Together: Homework

78. The average number of people in a family that attended college for various years is given in Table 12.30.

Year	Number of Family Members Attending College
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

Table 12.30

a. Using "year" as the independent variable and "Number of Family Members Attending College" as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated number of family members attending college.
e. Based on the data in Table 12.30, is there a linear relationship between the year and the average number of family members attending college?
f. Using the least-squares line, estimate the number of family members attending college for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
g. Are there any outliers in the data?
h. What is the estimated average number of family members attending college for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least squares (best-fit) line? Interpret the slope.

79. The percent of women who are wage and salary workers who are paid hourly rates is given in Table 12.31 for a