

Mind on Statistics**Chapter 3****Section 3.1**

1. Which one of the following is *not* appropriate for studying the relationship between two quantitative variables?
 - A. Scatterplot
 - B. Bar chart
 - C. Correlation
 - D. Regression

KEY: B

2. Which one of the following is a variable that we usually put on the horizontal axis of a scatterplot?
 - A. *y* variable
 - B. explanatory variable
 - C. response variable
 - D. dependent variable

KEY: B

3. Which one of the following *cannot* be determined from a scatterplot?
 - A. a positive relationship
 - B. a negative relationship
 - C. a cause and effect relationship
 - D. a curvilinear relationship

KEY: C

4. A scatterplot is a
 - A. one-dimensional graph of randomly scattered data.
 - B. two-dimensional graph of a straight line.
 - C. two-dimensional graph of a curved line.
 - D. two-dimensional graph of data values.

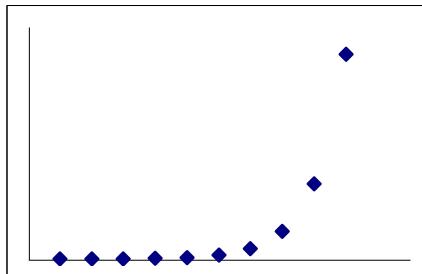
KEY: D

5. Two variables have a positive association when
 - A. the values of one variable tend to increase as the values of the other variable increase.
 - B. the values of one variable tend to decrease as the values of the other variable increase.
 - C. the values of one variable tend to increase regardless of how the values of the other variable change.
 - D. the values of both variables are always positive.

KEY: A

Chapter 3

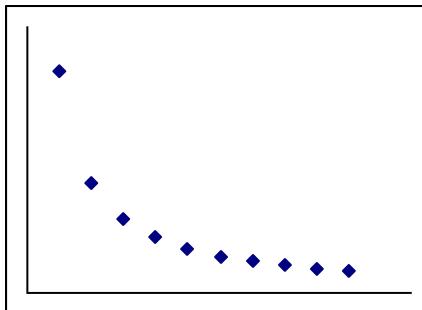
6. Describe the type of association shown in the scatterplot below:



- A. Positive linear association
- B. Negative linear association
- C. Positive curvilinear association
- D. Negative curvilinear association

KEY: C

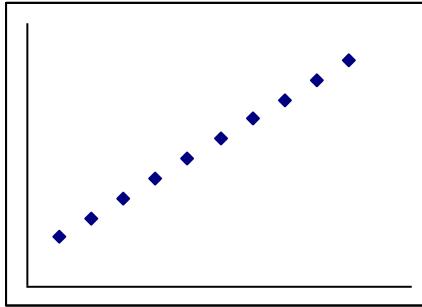
7. Describe the type of association shown in the scatterplot below:



- A. Positive linear association
- B. Negative linear association
- C. Positive curvilinear association
- D. Negative curvilinear association

KEY: D

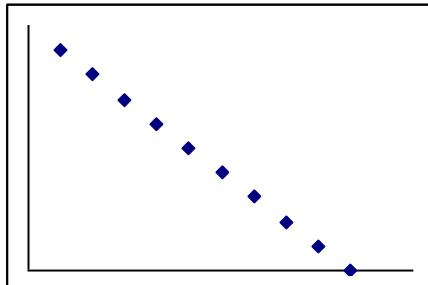
8. Describe the type of association shown in the scatterplot below:



- A. Positive linear association
- B. Negative linear association
- C. Positive curvilinear association
- D. Negative curvilinear association

KEY: A

9. Describe the type of association shown in the scatterplot below:



- A. Positive linear association
- B. Negative linear association
- C. Positive curvilinear association
- D. Negative curvilinear association

KEY: B

10. Which of the following sets of variables is most likely to have a negative association?

- A. the height of the son and the height of the father
- B. the age of the wife and the age of the husband
- C. the age of the mother and the number of children in the family
- D. the age of the mother and the ability to have children

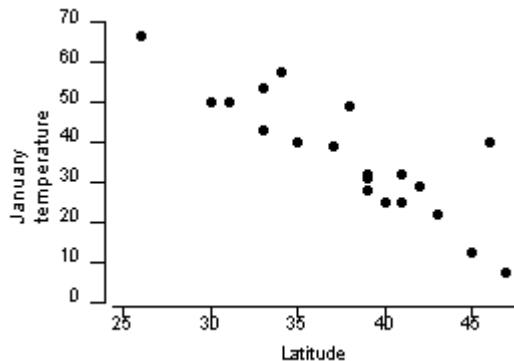
KEY: D

11. Which of the following sets of variables is most likely to have a negative association?

- A. the number of bedrooms and the number of bathrooms in a house
- B. the number of rooms in a house and the time it takes to vacuum the house
- C. the age of a house and the cleanliness of the carpets inside
- D. the size of a house and its selling price

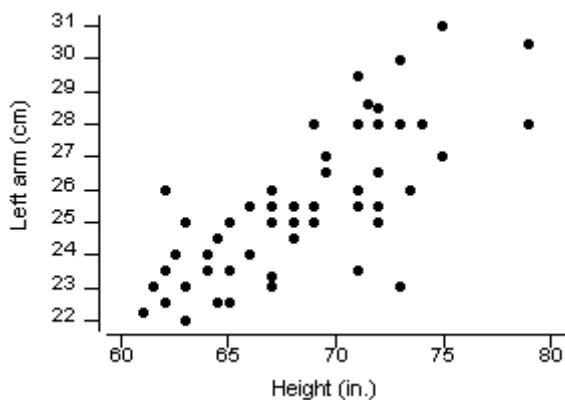
KEY: C

12. A scatterplot of geographic latitude (X) and average January temperature (Y) for 20 cities in the United States is given below. Is there a positive association or a negative association? Explain what such an association means in the context of this situation.



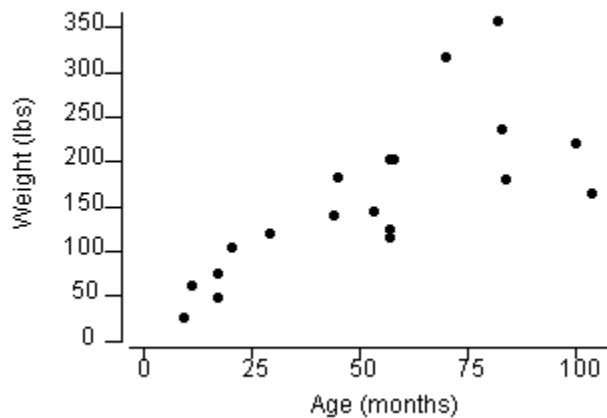
KEY: There is a negative association, which means that as the latitude increases, the value of the average January temperatures tends to decrease.

13. A scatterplot of $y = \text{height}$ and $x = \text{left forearm length (cm)}$ for 55 college students is given below. Is there a positive association or a negative association? Explain what this association means in the context of this situation.



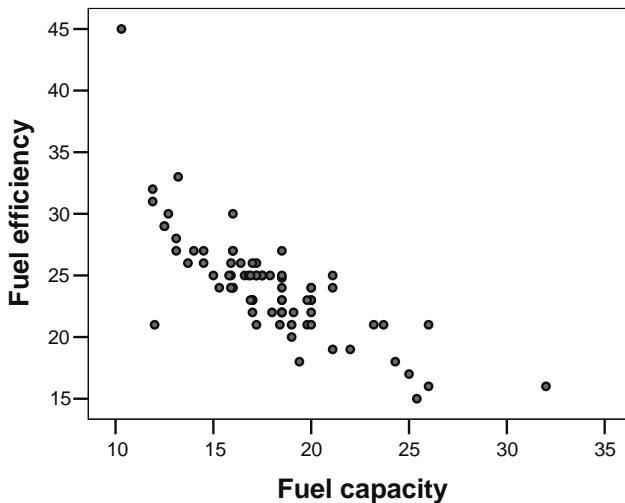
KEY: There is a positive association, which means that as height increases, left forearm length also tends to increase.

14. The following plot shows the association between weight (pounds) and age (months) for 19 female bears. Write an interpretation of this plot. Is the association negative or positive? Linear or curvilinear? Are there any outliers?



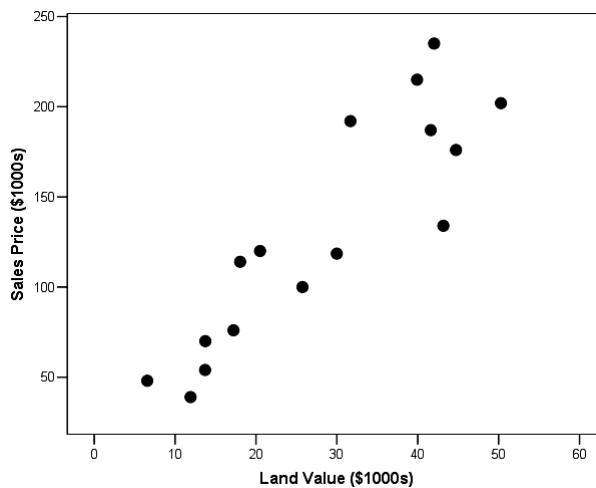
KEY: There is a positive association. One possible interpretation is that there is a curvilinear relationship that flattens out as age increases, and there may be two outliers (the two bears with the greatest weights).

15. A car salesman is curious if he can predict the fuel efficiency of a car (in MPG) if he knows the fuel capacity of the car (in gallons). He collects data on a variety of makes and models of cars. The scatterplot below shows the association between fuel capacity and fuel efficiency for a sample of 78 cars. Is the association negative or positive? Linear or curvilinear? Are there any outliers?



KEY: There is a negative association, which means that as the fuel capacity (gallons) increases, the value of the fuel efficiency (mpg) tends to decrease. There does seem to be a very slightly curvilinear relationship and a few unusual observations. One car has the smallest fuel capacity and a fuel efficiency that is a bit higher than expected. Another small fuel capacity car with a fuel efficiency that is a bit lower.

16. In an attempt to model the relationship between sales price and assessed land value, a realtor has taken a simple random sample of houses recently sold in the area and obtained the following data (both are expressed in thousands of dollars). The scatterplot below shows the association between fuel capacity and fuel efficiency for a sample of 78 cars. Is the association negative or positive? Linear or curvilinear? Are there any outliers?

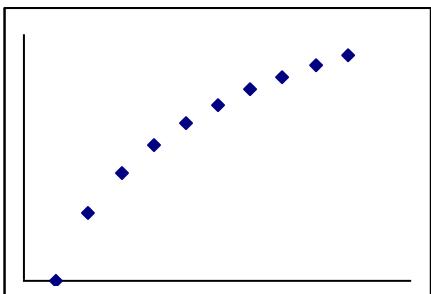


KEY: There is a positive association, which means that as the assessed land value of a home increases, the sales price of the home also tends to increase.

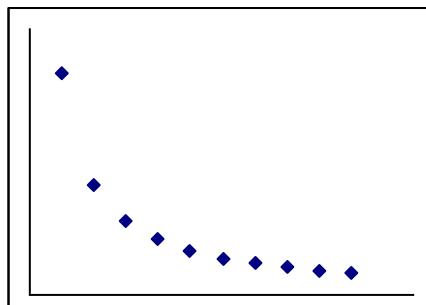
Section 3.2

17. Which graph shows a pattern that would be appropriately described by the equation $\hat{y} = b_0 + b_1x$?

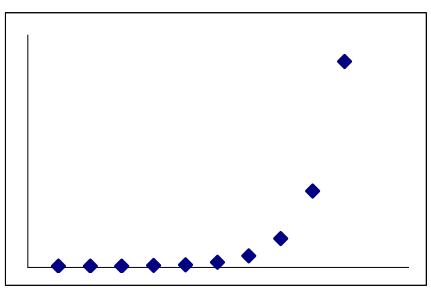
A.



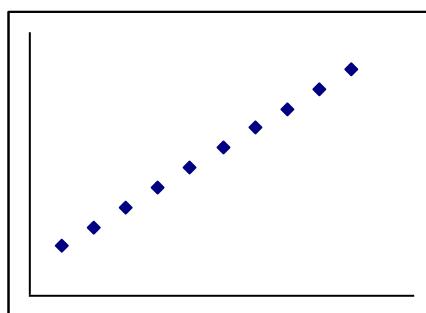
B.



C.



D.



KEY: D

18. Which of the following can *not* be answered from a regression equation?

- A. Predict the value of y at a particular value of x .
- B. Estimate the slope between y and x .
- C. Estimate whether the linear association is positive or negative.
- D. Estimate whether the association is linear or non-linear

KEY: D

19. A regression line is a straight line that describes how values of a quantitative explanatory variable (y) are related, on average, to values of a quantitative response variable (x).

- A. True
- B. False

KEY: B

20. The equation of a regression line is called the regression equation.

- A. True
- B. False

KEY: A

21. A regression line can be used to estimate the average value of y at any specified value of x .

- A. True
- B. False

KEY: A

22. A regression line can be used to predict the unknown value of x for an individual, given that individual's y value.
- True
 - False

KEY: B

Questions 23 to 27: The simple linear regression equation can be written as $\hat{y} = b_0 + b_1x$.

23. In the simple linear regression equation, the symbol \hat{y} represents the
- average or predicted response.
 - estimated intercept.
 - estimated slope.
 - explanatory variable.

KEY: A

24. In the simple linear regression equation, the term b_0 represents the
- estimated or predicted response.
 - estimated intercept.
 - estimated slope.
 - explanatory variable.

KEY: B

25. In the simple linear regression equation, the term b_1 represents the
- estimated or predicted response.
 - estimated intercept.
 - estimated slope.
 - explanatory variable.

KEY: C

26. In the simple linear regression equation, the symbol x represents the
- estimated or predicted response.
 - estimated intercept.
 - estimated slope.
 - explanatory variable.

KEY: D

27. In the simple linear regression equation, b_1 represents the slope of the regression line. Which of the following gives the best interpretation of the slope?
- It is an estimate of the average value of y at any specified value of x .
 - It indicates how much of a change there is for the predicted or average value of y when x increases by one unit.
 - It is the value of y , when $x = 0$.
 - It is the value of x , when $y = 0$.

KEY: B

Questions 28 to 30: A regression between foot length (response variable in cm) and height (explanatory variable in inches) for 33 students resulted in the following regression equation:

$$\hat{y} = 10.9 + 0.23x$$

28. One student in the sample was 73 inches tall with a foot length of 29 cm. What is the predicted foot length for this student?

- A. 17.57 cm
- B. 27.69 cm
- C. 29 cm
- D. 33 cm

KEY: B

29. One student in the sample was 73 inches tall with a foot length of 29 cm. What is the residual for this student?

- A. 29 cm
- B. 1.31 cm
- C. 0.00 cm
- D. -1.31 cm

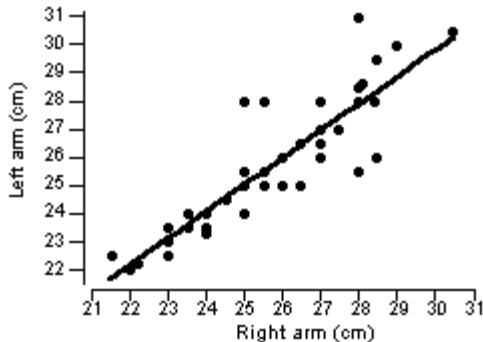
KEY: B

30. What is the estimated average foot length for students who are 70 inches tall?

- A. 27 cm
- B. 28 cm
- C. 29 cm
- D. 30 cm

KEY: A

31. The following scatterplot shows the relationship between the left and right forearm lengths (cm) for 55 college students along with the regression line, where y = left forearm length x = right forearm length. (Source: *physical* dataset on the CD.)



Which is the appropriate equation for the regression line in the plot?

- A. $\hat{y} = 1.22 + 0.95x$
- B. $\hat{y} = 1.22 - 0.95x$
- C. $\hat{x} = 1.22 + 0.95y$
- D. $\hat{x} = 1.22 - 0.95y$

KEY: A

Questions 32 to 34: A group of adults aged 20 to 80 were tested to see how far away they could first hear an ambulance coming towards them. An equation describing the relationship between distance (in feet) and age was found to be:

$$\text{Distance} = 600 - 3 \times \text{Age}.$$

32. Estimate the distance for an individual who is 20 years old.

- A. 11,940 feet
- B. 597 feet
- C. 540 feet
- D. 20 feet

KEY: C

33. How much does the estimated distance change when age is increased by 1?

- A. It goes down by 1 foot.
- B. It goes down by 3 feet.
- C. It goes up by 1 foot.
- D. It goes up by 3 feet.

KEY: B

34. Based on the equation, what is the direction of the association between distance and age?

- A. Positive
- B. Negative
- C. Zero
- D. Direction can't be determined from the equation.

KEY: B

35. For which one of these relationships could we use a regression analysis? Only one choice is correct.

- A. Relationship between weight and height.
- B. Relationship between political party membership and opinion about abortion.
- C. Relationship between gender and whether person has a tattoo.
- D. Relationship between eye color (blue, brown, etc.) and hair color (blond, etc.).

KEY: A

36. The regression line for a set of points is given by $\hat{y} = 12 - 6x$. What is the slope of the line?

- A. -12
- B. -6
- C. +6
- D. +12

KEY: B

37. A scatter plot and regression line can be used for all of the following *except*

- A. to determine if any (x, y) pairs are outliers.
- B. to predict y at a specific value of x .
- C. to estimate the average y at a specific value of x .
- D. to determine if a change in x causes a change in y .

KEY: D

Chapter 3

Questions 38 to 41: Past data has shown that the regression line relating the final exam score and the midterm exam score for students who take statistics from a certain professor is:

$$\text{final exam} = 50 + 0.5 \times \text{midterm}$$

38. What is the predicted final exam score for a student with a midterm score of 50?

- A. 50
- B. 50.5
- C. 75
- D. 100

KEY: C

39. Which of the following gives a correct interpretation of the intercept?

- A. A student who scored 0 on the midterm would be predicted to score 50 on the final exam.
- B. A student who scored 0 on the final exam would be predicted to score 50 on the midterm exam.
- C. A student who scored 2 points higher than another student on the midterm would be predicted to score 1 point higher than the other student on the final exam.
- D. None of the interpretations above are correct interpretations of the intercept.

KEY: A

40. Which of the following gives a correct interpretation of the slope?

- A. A student who scored 0 on the midterm would be predicted to score 50 on the final exam.
- B. A student who scored 0 on the final exam would be predicted to score 50 on the midterm exam.
- C. A student who scored 2 points higher than another student on the midterm would be predicted to score 1 point higher than the other student on the final exam.
- D. None of the interpretations above are correct interpretations of the slope.

KEY: C

41. Midterm exam scores could range from 0 to 100. Based on the equation, final exam scores are predicted to range from

- A. 0 to 100.
- B. 50 to 100.
- C. 50 to 75.
- D. 0 to 75.

KEY: B

42. A regression analysis done with Minitab for left foot length (y variable) and right foot length (x variable) for 55 college students gave the following output.

Predictor	Coef	StDev	T	P
Constant	2.516	1.116	2.25	0.028
RtFoot	0.89756	0.04317	20.79	0.000
S = 0.8332		R-Sq = 89.1% R-Sq(adj) = 88.9%		

The regression equation for left foot length (y variable) and right foot length (x variable) is

- A. $\hat{y} = 0.89756 + 2.516x$
- B. $\hat{y} = 1.116 + 0.04217x$
- C. $\hat{y} = 2.25 + 20.797x$
- D. $\hat{y} = 2.516 + 0.89756x$

KEY: D

43. In a certain study the relationship between (social) mobility of female foreign American residents and the number of years they have lived in the U.S. is examined. The researcher hopes to predict the mobility of these women based on their length of stay. Part of the SPSS output is shown below:

Coefficients(a)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant) years	-3.799 7.745	7.353 .904	.778	-.517 8.569	.608 .000

a. Dependent Variable: mobility

What is the correct regression equation for summarizing this linear relationship?

- A. $\hat{y} = -3.799 + 7.745 \text{ years}$
- B. $\hat{y} = 7.745 + 0.778 \cdot \text{years}$
- C. $\hat{y} = -3.799 + 0.778 \text{ mobility}$
- D. $\hat{y} = 7.745 + -3.799 \text{ mobility}$

KEY: A

Questions 44 to 47: A graduate student is doing a research project on stress among recent immigrants. She focused on parents of young children. All parents answered questions on a questionnaire (with the help of a translator if necessary). A stress score was calculated for all participants. SPSS was used to run a simple linear regression in which y = stress of the parent and x = age of the parent.

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
					.440 ^a
1 Regression	.257	1	.257	.610	.440 ^a
Residual	14.354	34	.422		
Total	14.611	35			

a. Predictors: (Constant), Age parent in years

b. Dependent Variable: mean stress

Coefficients^a

Model	Unstandardized Coefficients		t	Sig.
	B	Std. Error		
1 (Constant)	2.422	.482	5.025	.002
Age parent in years	-.011	.014	-.781	.440

a. Dependent Variable: mean stress

44. What is the predicted stress score for a 35 year old parent?

- A. 84.76
- B. 0.48
- C. 4.64
- D. 2.04

KEY: D

Chapter 3

45. What is an estimate for the average stress score for parents who are 28 years old?

- A. 2.114
- B. 67.805
- C. 15.918
- D. 27.692

KEY: A

46. Madeline is a 30-year old mother of young children who just emigrated from Germany. She scored a 1.9 on the stress test. What is the value of Madeline's residual?

- A. 0.192
- B. 42.649
- C. -14.982
- D. -0.192

KEY: D

47. A least squares line or least squares regression line has the property that the sum of squared differences between the observed values of y and the predicted values \hat{y} is smaller for that line than it is for any other line. What is the value of the SSE for this regression line?

- A. 0.257
- B. 14.354
- C. 14.611
- D. 0.422

KEY: B

48. A basketball coach of a youth team wishes to predict the number of points the players will score in their first season as a junior (y) based on the number of points they scored in their last season as youth players (x). The average number of points the team scored as youth players was $\bar{x} = 7.9$ and the average number of points they scored in their first year as junior players was $\bar{y} = 10.2$. The slope is $b_1 = 0.79$. What is the predicted number of goals for a player who scored 7 goals in his last season as a youth player?

- A. 7.78
- B. 9.49
- C. 11.46
- D. 15.58

KEY: B

Questions 49 to 51: A regression was done for 20 cities with latitude as the explanatory variable (x) and average January temperature as the response variable (y). The latitude is measured in degrees and average January temperature in degrees Fahrenheit. The latitudes ranged from 26 (Miami) to 47 (Duluth). The regression equation is

$$\hat{y} = 49.4 - 0.313x$$

49. The city of Pittsburgh, PA has latitude 40 degrees with average January temperature of 25 degrees Fahrenheit. What is the estimated average January temperature for Pittsburgh, based on the regression equation? What is the residual?

KEY: The estimated January temperature is 36.88 degrees Fahrenheit. The residual is -11.88 degrees.

50. The city of Miami, Florida has latitude 26 degrees with average January temperature of 67 degrees Fahrenheit. What is the estimated average January temperature for Miami, based on the regression equation? What is the residual?

KEY: The estimated January temperature is 41.3 degrees Fahrenheit. The residual is 25.7 degrees.

51. Mexico City has latitude 20 degrees. What is the problem with using the regression equation to estimate the average January temperature for Mexico City?

KEY: The regression equation is based on 20 cities, with the lowest latitude being 26 degrees. Estimating the temperature in Mexico City would be an extrapolation.

Questions 52 to 54: The next three questions are based on a regression equation for 55 college students with x = left forearm length (cm) and y = height. The forearm lengths ranged from 22 cm to 31 cm. The regression equation is

$$\hat{y} = 30.3 + 1.49x$$

52. One student's left forearm length was 27 cm, and his height was 75 inches. What is the estimated height for this student, based on the regression equation? What is the residual?

KEY: The estimated height is 70.53 inches. The residual is 4.47 inches.

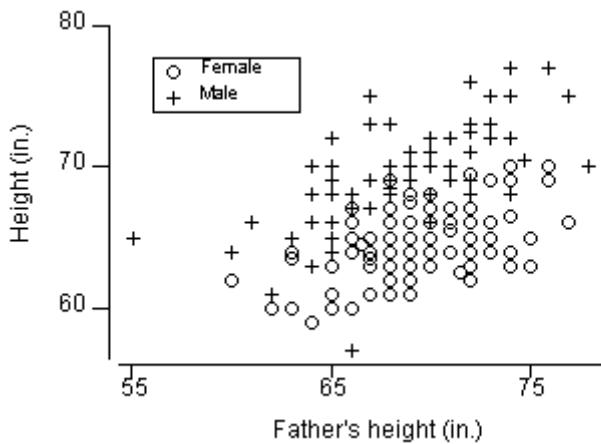
53. One student's left forearm length was 22 cm, and her height was 63 inches. What is the estimated height for this student, based on the regression equation? What is the residual?

KEY: The estimated height is 63.08 inches. The residual is -0.08 inches.

54. The instructor's left forearm length is 35 cm. Is there any problem with using the regression equation to estimate the instructor's height?

KEY: The instructor's left forearm length is larger than any of the students, so using the regression equation would be an extrapolation.

55. The scatterplot below shows students' heights (y) versus father's heights (x) for a sample of 173 college students. The symbol "+" represents a male student and the symbol "o" represents a female student.



A linear regression equation is determined for the relationship between female students' heights and their father's heights and, separately, one for male students and their fathers. How do the y -intercepts compare for these two regression equations?

KEY: The y -intercept for the female students will be lower than the y -intercept for the male students. Note that the slopes of the two equations will be approximately the same.

Questions 56 and 57: The following output is for a simple regression in which y = grade point average (GPA) and x = number of classes skipped in a typical week. The results were determined using self-reported data for a sample of $n = 1,673$ students at a large northeastern university.

The regression equation is GPA = 3.25 - 0.0614 Skip_Classes					
Predictor	Coef	SE Coef	T	P	
Constant	3.25415	0.01409	230.88	0.000	
Skip_Cla	-0.061411	0.006593	-9.32	0.000	
S = 0.4378	R-Sq = 4.9%	R-Sq(adj) = 4.9%			

56. What is the value of the slope of the sample regression line? Write a sentence that interprets this slope in the context of this situation.

KEY: The slope is -0.0614 , or about -0.06 . Average grade point average decreases about 0.06 per each additional class skipped per week.

57. What is the predicted grade point average for a student who skips 4 classes per typical week?

KEY: The predicted GPA = $3.25 - 0.0614(4) \approx 3.00$.

Questions 58 to 61: The 25 item “Parenting Hassles Scale” asks parents questions about certain situations in their family. They are asked to rate these situations on (1) how often they occur (frequency), and (2) how much they bother them (intensity). The score system for the frequency was 0-4 points (0 = never, 4 = all the time) and for the intensity 1-5 (1 = not much at all, 5 = very much). If an item received a 0 (zero) for frequency, it automatically received a 0 (zero) for intensity. The two variables, frequency and intensity, were formed by taking the total of the corresponding scores from the two rating scales across the various situations. In a scatterplot, these two variables showed an approximate linear relationship. They were then used to run a linear regression to predict intensity based on frequency.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3188.002	1	3188.002	12.382	.001 ^a
	Residual	13131.168	51	257.474		
	Total	16319.170	52			

a. Predictors: (Constant), Frequency

b. Dependent Variable: Intensity

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
	B	Std. Error			
1	(Constant)	36.457	6.160	5.918	.000
	Frequency	.335	.095	.442	.001

a. Dependent Variable: Intensity

58. What is the predicted intensity score for someone who scores a 45 on the frequency scale?

KEY: 51.532

59. What is an estimate for the average intensity score for parents who score 33 on the frequency scale?

KEY: 47.512

60. Dan filled out the “Parenting Hassles Scale”. He scored 25 on the frequency scale and 41 on the intensity scale. What is the value of Dan’s residual?

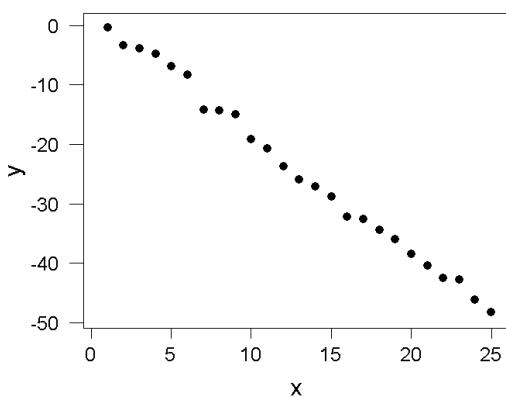
KEY: Dan’s predicted value is 44.832. So his residual is $41 - 44.832 = -3.832$.

61. What is the value of the Sum of Squared Errors for this regression line?

KEY: SSE = 13131.168

Section 3.3

62. A scatterplot of X and Y is shown below. Which value of the correlation coefficient, r , best describes the relationship?



- A. $r = 0.80$
- B. $r = -0.95$
- C. $r = -1.00$
- D. $r = -2$

KEY: B

63. You wish to describe the relationship between exam grades and the amount of time students watch the Discovery Channel. The correlation turns out to be $r = +0.30$. What does this mean?

- A. The more a student watches the Discovery Channel, the higher his or her exam grades tend to be.
- B. The more a student watches the Discovery Channel, the lower his or her exam grades tend to be.
- C. In order to increase your exam grades, it is recommended that you spend more time watching the Discovery Channel.
- D. 30% of the variation in exam grades is explained by the linear relationship with time spent watching the Discovery Channel.

KEY: A

64. A regression equation for left palm length (y variable) and right palm length (x variable) for 55 college students gave an error sum of squares (SSE) of 10.7 and a total sum of squares (SSTO) of 85.2. What proportion of the variation in y is explained by x ?

- A. 11.2%.
- B. 12.6%.
- C. 87.4%.
- D. 88.8%.

KEY: C

Chapter 3

65. A group of adults aged 20 to 80 were tested to see how far away they could first hear an ambulance coming towards them. An equation describing the relationship between distance (in feet) and age was found to be:

$$\text{Distance} = 600 - 3 \times \text{Age}$$

Based on the equation, what is the strength of the relationship between distance and age?

- A. There is a strong relationship.
- B. There is a weak relationship.
- C. There is no relationship.
- D. Strength can't be determined from the equation.

KEY: D

66. Which of the following is a possible value of r^2 , and indicates the strongest linear relationship between two quantitative variables?

- A. -90%
- B. 0%
- C. 80%
- D. 120%

KEY: C

67. The correlation between two variables is given by $r = 0.0$. What does this mean?

- A. The best straight line through the data is horizontal.
- B. There is a perfect positive relationship between the two variables.
- C. There is a perfect negative relationship between the two variables.
- D. All of the points must fall exactly on a horizontal straight line.

KEY: A

68. If the correlation coefficient between two quantitative variables is positive, what conclusion can we draw?

- A. High values on the first variable are associated with high values on the second variable.
- B. Low values on the first variable are associated with low values on the second variable.
- C. High values on the first variable are associated with low values on the second variable.
- D. Both A and B are correct.

KEY: D

69. Which of the following correlation values indicates the strongest linear relationship between two quantitative variables?

- A. $r = -0.65$
- B. $r = -0.30$
- C. $r = 0.00$
- D. $r = 0.50$

KEY: A

70. The value of a correlation is reported by a researcher to be $r = -0.5$. Which of the following statements is correct?

- A. The x -variable explains 50% of the variability in the y -variable.
- B. The x -variable explains -50% of the variability in the y -variable.
- C. The x -variable explains 25% of the variability in the y -variable.
- D. The x -variable explains -25% of the variability in the y -variable.

KEY: C

71. Which one of the following statements involving correlation is possible and reasonable?

- A. The correlation between hair color and eye color is 0.80.
- B. The correlation between the height of a father and the height of his first son is 0.6
- C. The correlation between left foot length and right foot length is 2.35.
- D. The correlation between hair color and age is positive.

KEY: B

72. Correlation and regression are concerned with
- the relationship between two categorical variables.
 - the relationship between two quantitative variables.
 - the relationship between a quantitative explanatory variable and a categorical response variable.
 - the relationship between a categorical explanatory variable and a quantitative response variable.

KEY: B

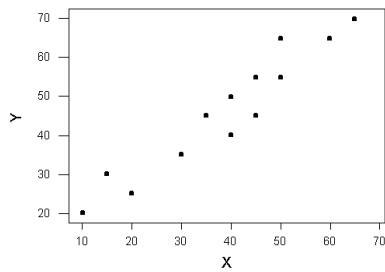
73. A researcher reports that the correlation between two quantitative variables is $r = 0.8$. Which of the following statements is correct?
- The average value of y changes by 0.8 when x is increased by 1.
 - The average value of x changes by 0.8 when y is increased by 1.
 - The explanatory variable (x) explains 0.8 or 80% of the variation in the response variable (y).
 - The explanatory variable (x) explains $(0.8)^2 = 0.64$ or 64% of the variation in the response variable (y).

KEY: D

74. A reviewer rated a sample of fifteen wines on a score from 1 (very poor) to 7 (excellent). A correlation of 0.92 was obtained between these ratings and the cost of the wines at a local store. In plain English, this means that
- in general, the reviewer liked the cheaper wines better.
 - having to pay more caused the reviewer to give a higher rating.
 - wines with low ratings are likely to be more expensive (probably because fewer will be sold).
 - in general, as the cost went up so did the rating.

KEY: D

75. Consider the following graph:



For these data, there is

- positive correlation.
- negative correlation.
- biased correlation.
- zero correlation.

KEY: A

76. Which of the following is a deterministic relationship?
- The relationship between hair color and eye color.
 - The relationship between father's height and son's height.
 - The relationship between height in inches and height in centimeters.
 - The relationship between height as determined with a ruler and height as determined by a tape measure.

KEY: C

Chapter 3

77. The correlation between father's heights and student's heights for a sample of 79 male students is $r = 0.669$. What is the proportion of variation in son's heights explained by the linear relationship with father's heights?
KEY: 44.8%

78. The following output is for a simple regression in which y = grade point average (GPA) and x = number of classes skipped in a typical week. The results were determined using self-reported data for a sample of $n = 1,673$ at a large northeastern university.

The regression equation is $GPA = 3.25 - 0.0614 \text{ Skip_Classes}$
Predictor Coef SE Coef T P
Constant 3.25415 0.01409 230.88 0.000
Skip_Cla -0.061411 0.006593 -9.32 0.000
$S = 0.4378$ R-Sq = 4.9% R-Sq(adj) = 4.9%

- What value is given in the output for R^2 ? Based on this value, explain whether the association between GPA and classes skipped per week is a strong association or a weak association.
KEY: $R^2 = 4.9\%$. This indicates a weak association. Number of classes skipped per week explains only about 5% of the (linear) variation in grade point averages.

Questions 79 and 80: The next two questions are based on a regression equation for 55 college students with x = left forearm length (cm) and y = height. The forearm lengths ranged from 22 cm to 31 cm. The regression equation is

$$\hat{y} = 30.3 + 1.49x.$$

79. The total sum of squares (SSTO) = 1054.8, and the error sum of squares (SSE) = 464.5. What is the value of r^2 , the proportion of variation in y explained by its linear relationship with x ?
KEY: $(1054.8 - 464.5)/1054.8 = 0.56$ or 56%.

80. What is the correlation between left forearm length and height?

KEY: $r = \sqrt{0.56} = 0.748$

Questions 81 and 82: A regression was done for 20 cities with latitude as the explanatory variable (x) and average January temperature as the response variable (y). The latitude is measured in degrees and average January temperature in degrees Fahrenheit. The latitudes ranged from 26 (Miami) to 47 (Duluth). The regression equation is

$$\hat{y} = 49.4 - 0.313x.$$

81. The total sum of squares (SSTO) = 4436.6, and the error sum of squares (SSE) = 1185.8. What is the value of r^2 , the proportion of variation in y explained by its linear relationship with x ?
KEY: $(4436.6 - 1185.8)/4436.6 = 0.733$ or 73.3%.

82. What is the value of the correlation coefficient r ?

KEY: $r = -\sqrt{0.733} = -0.856$.

83. A counselor at a weight-loss clinic wishes to determine if there is a relationship between BMI and depression. The scores of 6 patients are determined. She makes a scatterplot to assess of the relationship appears to be linear. It turns out to be a positive linear relationship. The Sum of Squared Errors is 8.875 and the Total Sum of Squares (SSTO) is 17.5. What is the value of the correlation coefficient between BMI and depression?

KEY: $r^2 = (17.5 - 8.875)/17.5 = 0.493$. $r = +\sqrt{r^2} = +0.70$

84. A psychology student wishes to determine the strength of the relationship between school grades and short term memory. Twenty-two students were given a memory test. SPSS was used to analyze the data. Part of the output is shown below.

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	7.062	1	7.062	5.362	.031 ^a
Residual	26.339	20	1.317		
Total	33.401	21			

Coefficients^a

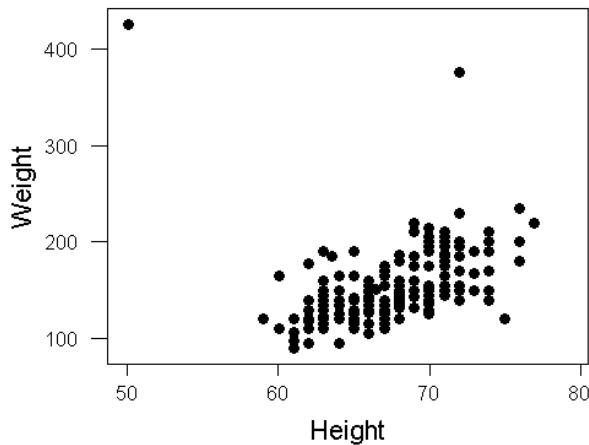
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	6.073	.545		11.142	.000
memorytest	.277	.120	.460	2.316	.031

What is the value of the correlation coefficient between grades and the memory test scores?

KEY: The slope is positive, so we know r is positive. $r^2 = 7.062/33.401 = 0.211$. $r = +\sqrt{r^2} = +0.46$

Sections 3.4 – 3.5

85. A scatterplot of the self-reported weights (y variable) and self-reported heights (x variable) for 176 college students follows.

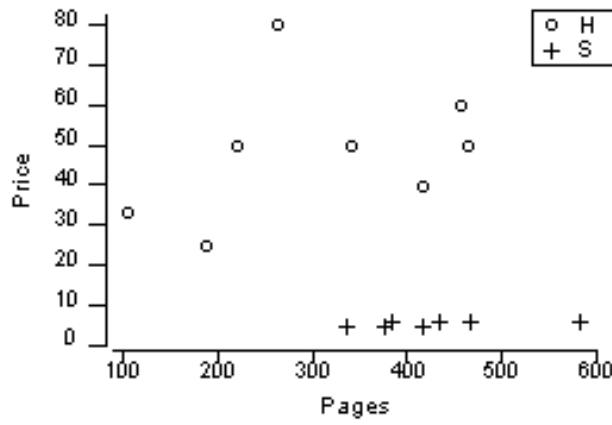


What is the main difficulty with using a regression line to analyze these data?

- A. Presence of one or more outliers
- B. Inappropriately combining groups
- C. Curvilinear data
- D. Response variable is not quantitative

KEY: A

86. A scatterplot of the price of a book (y variable) versus the number of pages in the book (x variable) is shown for 15 books in a professor's office. In addition, the symbol “o” shows that the book was a hardcover book, while the symbol “+” shows that the book was a softcover book.



What is the main difficulty with using a regression line to analyze these data?

- A. Presence of one or more outliers
- B. Inappropriately combining groups
- C. Curvilinear data
- D. Response variable is not quantitative

KEY: B

87. A researcher would like to study the relationship between y = graduating GPA of students at a university and x = major field of study (1 = science, 2 = humanities, etc.). What is a possible difficulty with using a regression line to analyze these data?
- A. Presence of one or more outliers
 - B. Inappropriately combining groups
 - C. Curvilinear data
 - D. Explanatory variable is not quantitative

KEY: D

88. Based on 1988 census data for the 50 States in the United States, the correlation between the number of churches per state and the number of violent crimes per state was 0.85. What can we conclude?
- A. The presence of a lot of churches in a state causes the number of violent crimes in the state to increase.
 - B. The correlation is spurious because of the confounding variable of population size: both number of churches and number of violent crimes are related to the state's population size.
 - C. Since the data comes from a census, or nearly complete enumeration of the United States, there must be a causal relationship between the number of churches and the number of violent crimes.
 - D. The relationship is not causal because only correlations of +1 or -1 show causal relationships.

KEY: B

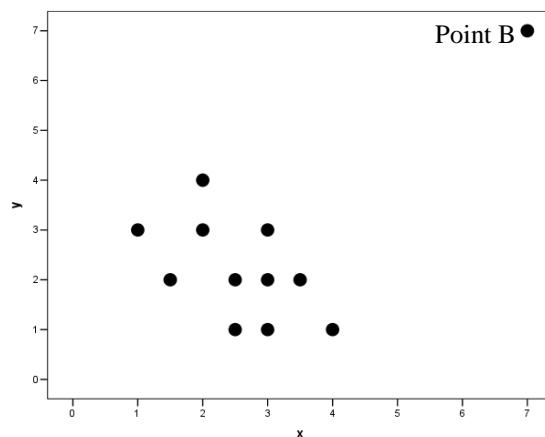
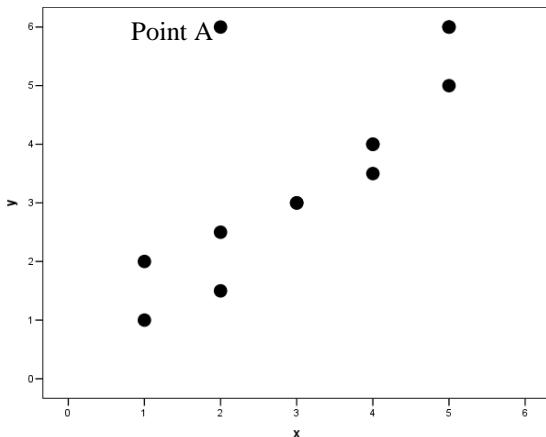
89. What is the effect of an outlier on the value of a correlation coefficient?
- A. An outlier will always decrease a correlation coefficient.
 - B. An outlier will always increase a correlation coefficient.
 - C. An outlier might either decrease or increase a correlation coefficient, depending on where it is in relation to the other points.
 - D. An outlier will have no effect on a correlation coefficient.

KEY: C

90. A scatter plot of number of teachers and number of people with college degrees for cities in Pennsylvania reveals a positive association. The most likely explanation for this positive association is:
- A. Teachers encourage people to get college degrees, so an increase in the number of teachers is causing an increase in the number of people with college degrees.
 - B. Teaching is a common profession for people with college degrees, so an increase in the number of people with college degrees causes an increase in the number of teachers.
 - C. Cities with higher incomes tend to have more teachers and more people going to college, so income is a confounding variable, making causation between number of teachers and number of people with college degrees difficult to prove.
 - D. Larger cities tend to have both more teachers and more people with college degrees, so the association is explained by a third variable, the size of the city.

KEY: D

Questions 91 and 92: Consider the following scatterplots for two quantitative variables y and x . One point has been labeled in each plot.



91. Which point would be the most influential?

- A. Point A
- B. Point B

KEY: B

92. Which point would produce the largest residual?

- A. Point A
- B. Point B

KEY: A

93. A professor found a negative correlation between the number of hours students came to her office hours and their score on her final exam. What possible confounding variable could explain the observed negative correlation?

KEY: Students who were having difficulty in the course could have come to office hours more often than students who were doing well but still not do as well on the final exam.

94. Suppose a study of employees at a large company found a negative correlation between weight and distance walked on an average day. Explain why a conclusion that walking reduces weight cannot be made from this study (i.e. What confounding variable could also explain the observed negative correlation?)

KEY: People who are already in shape could prefer to walk everyday, while people who are overweight may prefer not to walk.

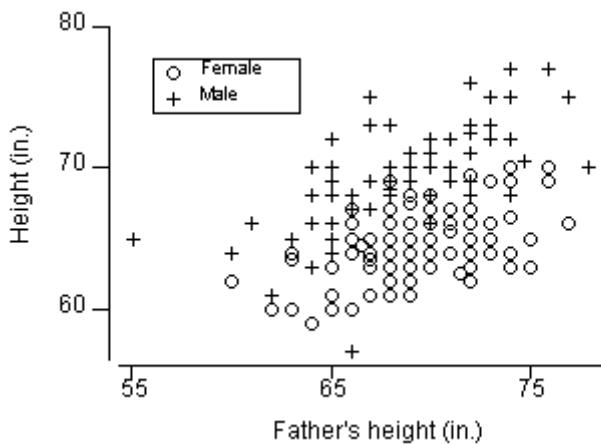
95. A survey of students found a negative correlation between the weekly hours of T.V. watched and the weekly hours spent exercising. One student explained that reducing the hours of T.V. watched (cause) would result in students sleeping longer and having more energy to exercise (effect). Give another explanation with hours of exercise as the cause and hours T.V. watched as the effect.

KEY: Students who exercise could be physically tired and go to bed earlier, thus reducing the number of T.V. hours watched.

96. A researcher found a positive correlation between the number of hours elementary school children spend watching T.V. (weekly) and their reading score on a standardized exam. What possible confounding variable could explain the observed positive correlation?

KEY: The grade or age of the children with older children (say 5th graders) watching more T.V. but would also have a higher reading score than younger children (say 1st graders).

97. The scatterplot below shows student heights (y axis) versus father's heights (x axis) for a sample of 173 college students. The symbol “+” represents a male student and the symbol “o” represents a female student.



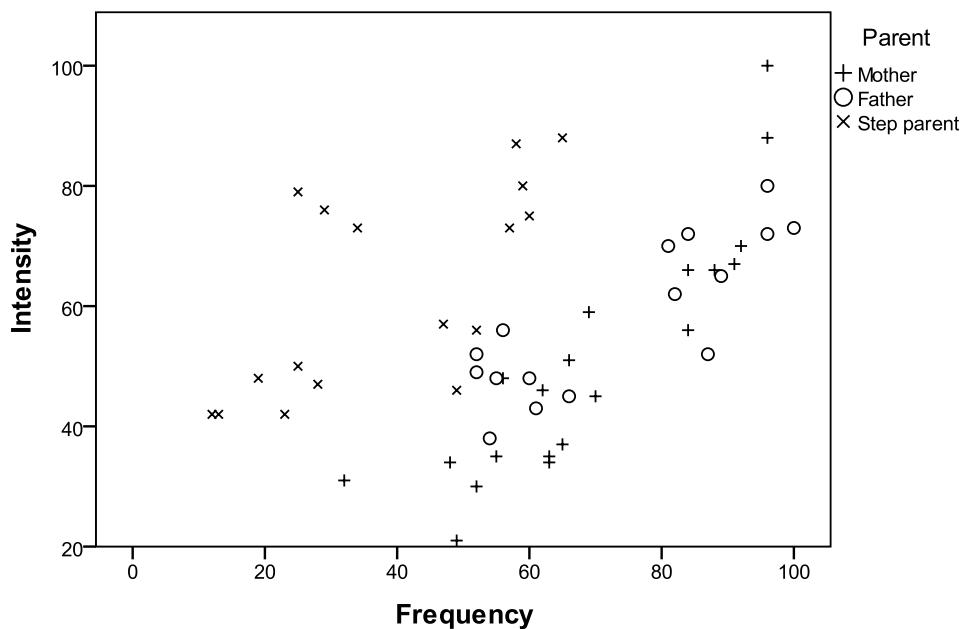
Based on the scatterplot, what is the problem with using a regression equation for all 173 students?

KEY: It appears that the men and women form two separate groups and should have two separate regression lines representing the relationship between father's height and student's height.

98. What are influential observations and where are they generally found?

KEY: Influential observations are generally outliers with extreme x values and their removal would dramatically change the correlation and regression line.

99. The 25 item “Parenting Hassles Scale” asks parents questions about certain situations in their family. They are asked to rate these situations on (1) how often they occur (frequency), and (2) how much they bother them (intensity). The score system for the frequency was 0-4 points (0 = never, 4 = all the time) and for the intensity 1-5 (1 = not much at all, 5 = very much). If an item received a 0 (zero) for frequency, it automatically received a 0 (zero) for intensity. The two variables, frequency and intensity, were formed by taking the total of the corresponding scores from the two rating scales across the various situations. These two variables are then used to create a scatterplot. The symbols indicate whether a mother, a father, or a step parent filled in the questionnaire.



Based on the scatterplot, what is the problem with using one regression equation for all parents combined?
KEY: It appears that the stepparents and biological parents (the group of mothers and fathers combined) form two separate groups and should have two separate regression lines representing the relationship between frequency and intensity.