

# Chapter 13

## Correlation and Regression

### 13.1 Study Suggestions

Like Chapter 12, this chapter stresses interpretation over computation. For example, at no place in Chapter 13 are you required to compute the correlation coefficient.

Make sure you understand the properties of straight lines that are reviewed in Section 13.1. In particular, practice substituting a number for  $x$  in an equation to obtain a value for  $y$ , and practice using a graph of a line to determine the value of  $y$  that corresponds to a particular value of  $x$ .

The scatterplot is the preeminent device introduced in Chapter 13. The scatterplot is of independent interest, and also must be examined to determine whether a more sophisticated data analysis procedure is reasonable.

The analysis of bivariate numerical data should begin with an inspection of each variable individually, using the methods of Chapter 12. The actual bivariate analysis begins with the examination of the scatterplot. (Although I do not require my students to create scatterplots, I do make sure that they understand how to determine the  $x$  and  $y$  values of points, and how to place a particular case's point in a scatterplot.)

Various insights can be gained into the relationship between  $X$  and  $Y$  by examining a scatterplot. Some of these insights are easier to illustrate for large data sets, and some are easier for small data sets. Thus, Chapter 13 includes a very large data set, the batting averages study with  $n = 124$  cases, and a number of small data sets, most notably the studies of spiders and of crickets.

The analyst first examines the scatterplot for the

presence of isolated cases. Like classifying outliers in Chapter 12, declaring a case to be isolated is a subjective endeavor. Large data sets are better than small data sets for illustrating the process of identifying isolated cases. In my opinion, the batting average study has three isolated cases, but a person can reasonably disagree. At this point, a tremendous advantage of the batting averages study emerges—each case has a name. This allows one to point to the three isolated cases and say, “This is Wade Boggs, this is Don Mattingly, and this is Floyd Rayford.” The existence of names for the cases has two advantages, one minor and one very helpful. First, the minor advantage is that the names help personalize the study and make it seem more real. Second, the isolated cases have a big impact on the analysis, and I will refer to them often. It is very convenient to be able to say, “Refer to Wade Boggs,” instead of “Refer to the case with  $x = 0.368$  and  $y = 0.357$ .”

After the search for isolated cases, the analyst must decide whether the relationship revealed by the scatterplot is linear or curved. The methods of Chapter 13 are appropriate and useful only if the relationship is linear.

The formula for the correlation coefficient is motivated by dividing the scatterplot into quadrants determined by the lines  $x = \bar{x}$  and  $y = \bar{y}$ , and then counting the number of points that fall in each quadrant. As mentioned earlier, I do not require students to compute the correlation coefficient by hand, but it is essential that they understand its properties listed at the end of Section 13.2 of the text.

The key result in Chapter 13 is the formula for the regression line. Instead of jumping immediately to the best line, I find it essential to begin with a com-

parison of two lines. After all, if a student does not understand why and how one line is better than another, how will that student understand that one line can be better than *all* others? To this end, I owe a great debt of gratitude to my former student Susan Robords. Susan's data on cricket chirps and temperature, and the line purported to be "correct" in a published source that she found and reported, provide a fascinating illustration of the major ideas behind the regression line. The analysis of Susan's data illustrate the important items listed below.

- Lines in a statistics course are different than lines in a mathematics course. In math, a line extends forever, but in statistics, a line is restricted to a particular finite interval of values of  $x$ . For example, Susan's data give us no insight into what it means (regarding temperature) if a cricket fails to chirp.
- Because a statistics course restricts a line to a finite range of values of  $x$ , it can be difficult to compare lines by comparing their slopes and intercepts. For example, the line Susan found in her readings and the regression line have very different slopes and intercepts, but Figure 13.20 on page 466 of the text shows that the lines, when restricted to the values of  $x$  in the data set, are very similar.
- The comparison of two lines can be visualized as a comparison of squared vertical distances in a scatterplot.
- The regression line is the best line in the overall sense of the principle of least squares. The regression line does not necessarily give good predictions for every (or even any) case, and, if compared to some other line, it does not necessarily give the better predicted value for every case.
- The slope and intercept of the regression line are *jointly* important because both are needed to obtain predicted values. The slope is important *individually* because it measures how a unit change in  $x$  influences the predicted value of the response. The intercept, however, is not important *individually* unless  $x = 0$  is in the range of the data.

- In view of the previous item, I prefer the alternate form of the regression line,

$$\hat{y} = \bar{y} + r\left(\frac{s_Y}{s_X}\right)(x - \bar{x}),$$

because every number in this form of the equation is always of interest individually.

- The third form of the regression line,

$$\frac{\hat{y} - \bar{y}}{s_Y} = r\left(\frac{x - \bar{x}}{s_X}\right),$$

is not recommended for computing predicted values, but is useful because it reveals another property of the correlation coefficient and aids understanding of the regression effect.

A natural question arises. The regression line is the best prediction line, but is it any good? The coefficient of determination and the distribution of the residuals provide relative and absolute answers to this question, respectively.

The coefficient of determination has two uses. First, it indicates how much improvement is obtained by using  $X$  to make predictions as compared to not using  $X$  to make predictions. Second, since the coefficient of determination equals the square of the correlation coefficient, we have another feature of the correlation coefficient, namely a precise way to interpret its square.

The distribution of the residuals allow the researcher to make a subjective assessment of the absolute usefulness of the regression line. The residual plot is shown to be a useful way to check the assumption of a linear relationship between  $X$  and  $Y$ .

The chapter ends with an examination of the impact on the analysis of a single isolated case, and the related, though easier to overlook, problem of having data from two sources.

## 13.2 Solutions to Odd-Numbered Exercises

### Solutions for Section 13.1

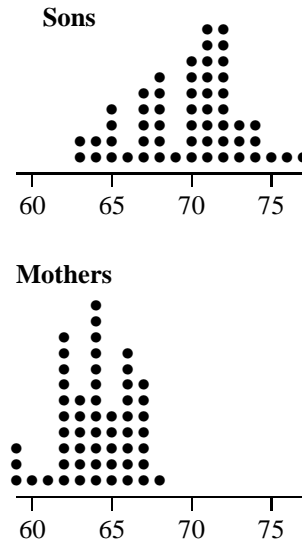
1. For  $x = 0$ ,  $y = 3 + 5(0) = 3$ . For  $x = -1$ ,  $y = 3 + 5(-1) = -2$ . For  $x = 2$ ,  $y = 3 + 5(2) = 13$ .

3. The picture in the text indicates that  $x = 0$  yields  $y = 3$ ,  $x = -1$  yields  $y = -2$ , and  $x = 2$  yields  $y = 13$ .

**Solutions for Section 13.2**

1. (a) It may or may not be true that persons who have seen a movie and have been interviewed are representative of the population of persons who eventually see that movie. Remember, however, that just because people who chose to see *Child's Play 2* enjoyed the movie, it does not follow that a person who dislikes "horror" movies would enjoy the film!
- (e) There is one isolated case,  $x = 1.7$  and  $y = 2$ , for *Bonfire of the Vanities*. The remainder of the cases exhibit a weak increasing linear relationship. The movie *Havana*, with  $x = 3.2$  and  $y = 2.7$ , could possibly be labeled as isolated. The relationship is increasing, but much too weak to yield  $r = 0.83$ . Thus, by elimination, I conclude  $r = 0.36$ .
5. (a) This example illustrates how difficult it is to identify isolated cases when there is little data. An argument could be made that the lightest spider is an isolated case, or that the spider with the highest heart rate is an isolated case, or, if one were particularly adventuresome, that the three cases in the lower right corner of the scatterplot are isolated cases. Personally, I do not want to label 3 out of 8 observations as being isolated! My choice for an isolated case is the lightest spider.
- (b) Clearly, the correlation coefficient is very close to zero, so I conclude that  $r = 0.055$ .
9. (a) The dot plots are below. The distribution of sons' heights is skewed to the left, with no outliers, and a deep valley at 69 inches. The distribution of mothers' heights is not markedly skewed, has no outliers, and has four peaks. The sons are taller and have

greater spread in their heights than the mothers.



- (b) The case with  $X = 59$ ,  $Y = 76$  is the most isolated case. The cases with  $(x = 60, y = 72)$ ,  $(x = 62, y = 75)$ , and  $(x = 67, y = 77)$  could also be labeled isolated.
- (d)  $r = 0.41$  because the other values are clearly too small or too large.

**Solutions for Section 13.3**

1. (b) The regression line is
- $$\hat{y} = 3.32 + 0.247(x - 2.52), \text{ or}$$
- $$\hat{y} = 2.70 + 0.247x.$$

The predicted value is

$$\hat{y} = 3.32 + 0.247(3.0 - 2.52) = 3.44$$

for *Ghost*, and

$$\hat{y} = 3.32 + 0.247(3.2 - 2.52) = 3.49$$

for *Havana*.

**Solutions for Section 13.4**

3. (a) The two smallest values of the residuals might be classified as outliers. Except for the possible outliers, the scatterplot of the residuals versus  $X$  looks fine—that is, there is no curvature.

- (b) This question is tricky. From Figure 13.30, the smallest residual is approximately equal to  $-1.1$ . From Figure 13.31, this residual belongs to a movie for which  $X = 1.7$ . Next, from Table 13.4 on page 456, there are two movies with  $X = 1.7$ : *Almost an Angel* and *Bonfire of the Vanities*. Because these two movies have the same value of  $X$ , they have the same value of  $\hat{y}$ . Thus, the movie with the smaller  $y$  will have the smaller value of  $e = y - \hat{y}$ . *Almost an Angel* has  $y = 2.8$  while *Bonfire of the Vanities* has  $y = 2.0$ . Thus, *Bonfire of the Vanities* has the smaller residual of these two movies and the smallest residual of all movies.

By similar reasoning, *Havana* has the second smallest residual.

### 13.3 Exam Questions

- Sally selects a random sample of 100 college students and finds the correlation coefficient between height (in inches) and weight (in pounds) to equal 0.6. Ralph selects his own random sample of size 100 from the same population but measures height in centimeters and weight in kilograms. The value of the correlation coefficient for Ralph's data \_\_\_\_\_.
  - is smaller than 0.6
  - is 0.6
  - is larger than 0.6
  - cannot be determined from the information given.
- A Cartesian coordinate system contains a scatterplot of a set of data and a graph of the regression line. Each case is represented by an 'O' in the scatterplot. The 'O' for a particular case is above the regression line.
  - The residual for this case is smaller than 0.
  - The residual for this case is equal to 0.
  - The residual for this case is larger than 0.
  - The residual for this case cannot be determined from the information given.
- A Cartesian coordinate system contains a scatterplot of a set of data and a graph of the regression line. Each case is represented by an 'O' in the scatterplot. The 'O' for a particular case lies exactly on the regression line.
  - The residual for this case is smaller than 0.
  - The residual for this case is equal to 0.
  - The residual for this case is larger than 0.
  - The residual for this case cannot be determined from the information given.
- An analyst obtains  $r = 0.6$ . If a subject's value of  $x$  equals  $\bar{x} + 2s_X$ , then the subject's predicted value of  $y$  equals \_\_\_\_\_.
  - $\bar{y} + 2s_Y$
  - $\bar{y} + 1.2s_Y$
  - $\bar{y} + 2s$
  - $\bar{y} + 1.2s$
- Which of the following statements is correct?
  - The value  $r = -0.5$  reflects a stronger linear relationship than does the value  $r = 0.6$ .
  - The value  $r = 0.6$  reflects a stronger linear relationship than does the value  $r = -0.5$ .
  - The values  $r = 0.6$  and  $r = -0.5$  reflect linear relationships of the same strength.
- The regression line for a set of data is found to be:
 
$$\hat{y} = 80 + 3(x - 50).$$
  - A particular case has  $x = 55$ ; find the regression line prediction of this case's value of the response.
  - Refer to part (a). If the actual response for the case is 100, compute the value of the residual.

7. A scatterplot suggests that there is a linear relationship between two variables  $X$  and  $Y$ . Computations yield:

$$n = 101, \bar{x} = 20, \bar{y} = 80, s_Y = 8, s_X = 2,$$

$$r = 0.5, \text{ and } s = 5.22.$$

What is the equation of the regression line?

8. A researcher has values of  $X$  and  $Y$  for a number of subjects. The researcher also has two rules, or formulas, that yield predicted values  $\hat{y}$  and  $\check{y}$ . You are further told that

$$\sum (y - \hat{y})^2 = 50 \text{ and}$$

$$\sum (y - \check{y})^2 = 50.$$

According to the Principle of Least Squares, which prediction rule is better,  $\hat{y}$  or  $\check{y}$ ?

9. Suppose that a regression line is  $\hat{y} = 50 + 2x$ .

True or false? The number 50 in the equation of the regression line can be interpreted as the predicted value of  $y$  for  $x = 0$ .

10. A regression analysis is performed with  $n = 6$  cases. Five of the residuals are  $-3, -2, 0, +4,$  and  $+1$ . How many cases (of the six) fall exactly on the regression line?
11. A Cartesian coordinate system contains a scatterplot of a set of data, a graph of the regression line, and the four quadrants determined by the lines  $x = \bar{x}$  and  $y = \bar{y}$ . Which (one or more) of the following must be true? (Hint: Recall that the graph of the regression line must include the point  $(\bar{x}, \bar{y})$  and draw a picture.)

- (a) If  $r < 0$ , then each case in the first quadrant has a positive residual.
- (b) If  $r < 0$ , then each case in the third quadrant has a negative residual.
- (c) If  $r > 0$ , then each case in the second quadrant has a positive residual.
- (d) If  $r > 0$ , then each case in the fourth quadrant has a negative residual.

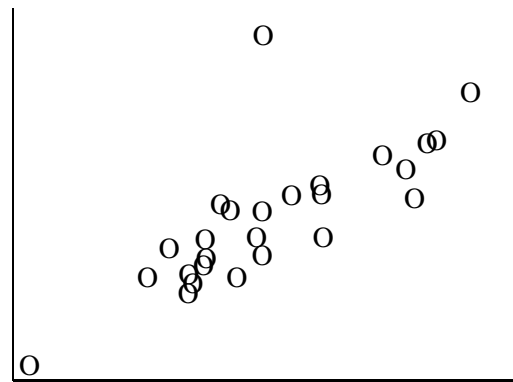
12. During summer school, 1995, 20 students completed Statistics 301. Each student took a midterm exam and a final exam. Summary statistics of the exam scores are given below.

Exam	Mean	Standard Deviation
Midterm	26.83	6.22
Final	31.65	6.82

In addition, the correlation coefficient equals 0.764. One of the students, Kelly, scored 25 on the midterm and 35 on the final.

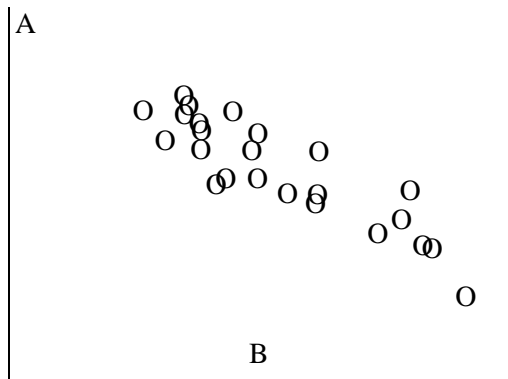
- (a) Obtain the regression line for using the midterm score to predict the final score.
- (b) Use your answer to (a) to predict Kelly's score on the final exam.
- (c) Refer to part (b). Compute the residual for Kelly.
- (d) In the scatterplot, in which quadrant is the circle (or whatever symbol is used) for Kelly? (Hint: Recall that quadrants are defined on pages 450 and 451 of the text.)

13. Below is a scatterplot of data.



- (a) A researcher labels exactly one of these cases to be isolated and an outlier. Put an "X" through that case.
- (b) A researcher labels exactly one of these cases to be isolated, but not an outlier. Circle that case.

14. Below is a scatterplot of data from 26 cases.



You are given that the correlation coefficient of these data equals  $-0.762$ . Define the following three data sets:

- Data set 2: The data set pictured above with the case marked “A” deleted.
- Data set 3: The data set pictured above with the case marked “B” deleted.
- Data set 4: The data set pictured above with the case marked “A” and the case marked “B” both deleted.

You are given that the correlation coefficients for data sets 2, 3, and 4, are  $-0.700$ ,  $-0.887$ , and  $-0.913$ , although not necessarily in that order.

Match each data set—2, 3, and 4—with its correlation coefficient. (Hint: The correspondence is exactly as suggested by Figure 13.28 on page 487 of the text and by the first paragraph of Section 13.4.3 on pages 485 and 486 of the text.)

15. Tori has performed a regression analysis, but, unfortunately, her dog chewed-up her results. Two facts remain: for  $x = 5$  the predicted value of  $y$  equals 10 and for  $x = 8$  the predicted value of  $y$  equals 6.

Which one of the following statements is true?

- (a) The correlation coefficient for Tori’s data is negative.

- (b) The correlation coefficient for Tori’s data is zero.
- (c) The correlation coefficient for Tori’s data is positive.
- (d) There is insufficient information to determine whether the correlation coefficient for Tori’s data is negative, zero, or positive.

16. Brian has performed a regression analysis, but, unfortunately, his cat clawed-up his results. All Brian can remember is that two different values of  $x$  gave exactly the same predicted value of  $y$ . Which one of the following statements is true?

- (a) The correlation coefficient for Brian’s data is negative.
- (b) The correlation coefficient for Brian’s data is zero.
- (c) The correlation coefficient for Brian’s data is positive.
- (d) There is insufficient information to determine whether the correlation coefficient for Brian’s data is negative, zero, or positive.

17. A researcher has values of  $x$  and  $y$  for four subjects,

$$\begin{array}{r} x : 0 \quad 1 \quad 2 \quad 3 \\ y : 1 \quad 2 \quad 6 \quad 8 \end{array}$$

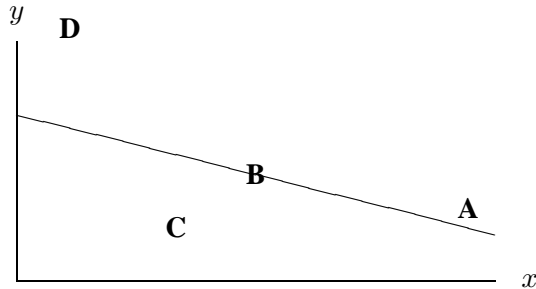
The researcher wants to compare two prediction rules,

$$\hat{y} = 2x, \text{ and } \check{y} = x^2.$$

According to the principle of least squares, which of these prediction rules is superior? To receive credit, you must produce computations to support your answer. (Hint: Refer to cricket data on pages 465–467 of the text.)

18. Refer to the previous question. The researcher discovers that the response corresponding to  $x = 2$  does not equal 6 as given before. Explain why a correction of the response would have no effect on the decision as to which prediction rule is superior.

19. A regression line for a set of data is graphed below. In addition, four of the cases from the data set are plotted. (Other cases are not shown.)



Use the above picture to answer the following questions. For each question sort the four cases, A, B, C, and D, from smallest to largest on the feature indicated. Be careful! There will be no partial credit given.

- Sort the cases by their value of  $x$ .
  - Sort the cases by their value of  $y$ .
  - Sort the cases by their value of  $\hat{y}$ .
  - Sort the cases by the absolute value of their residual.
  - Sort the cases by the value of their residual.
20. Richard performs a regression analysis and obtains three correlation coefficients. First, he obtains the correlation coefficient for  $x$  and  $y$  (this is the one the text calls simply *the* correlation coefficient). Next, he obtains the correlation coefficient for  $x$  and  $e$ . Finally, he obtains the correlation coefficient for  $x$  and  $\hat{y}$ . Unfortunately, he misplaces two of the three correlation coefficients. The one he has equals 0.73.

Determine which correlation coefficient he has and then determine the two missing correlation coefficients. Explain your answer.

21. You are given the following information about a regression analysis.
- A case with  $x = 20$  and  $y = 47$  has a residual of  $-3$ .
  - A case with  $x = 30$  has  $\hat{y} = 62$ .

- $s_y/s_x = 2$ .

Determine the value of the correlation coefficient,  $r$ , and the  $y$ -intercept,  $b_0$ .

## 13.4 Solutions to Exam Questions

- (d) Almost certainly Ralph has different people in his sample than Sally had.
- (c)
- (b)
- (b)
- (b) The value  $r = 0.6$  reflects a stronger linear relationship than does the value  $r = -0.5$  because it is further from zero..
- (a)  $\hat{y} = 80 + 3(55 - 50) = 95$ .  
(b)  $e = y - \hat{y} = 100 - 95 = 5$ .
- $\hat{y} = 80 + 2(x - 20)$  or  $\hat{y} = 40 + 2x$ .
- According to the Principle of Least Squares, the two prediction rules are equally good (or bad).
- False. (If  $x = 0$  is not included in the range of the data, then the data should not be used to predict the response when  $x = 0$ .)
- Because the residuals sum to 0, the sixth residual equals 0. Thus, two residuals equal 0 and these two cases fall exactly on the regression line.
- All four statements are true.
- (a) The slope of the regression line is

$$0.764\left(\frac{6.82}{6.22}\right) = 0.838.$$

The regression line is

$$\hat{y} = 31.65 + 0.838(x - 26.83), \text{ or}$$

$$\hat{y} = 9.17 + 0.838x.$$

- (b) The predicted final score for Kelly is

$$\hat{y} = 9.17 + 0.838(25) = 30.12.$$

(c) Kelly's residual is

$$e = 35 - 30.12 = 4.88.$$

(d) For Kelly,  $x < \bar{x}$  and  $y > \bar{y}$ ; thus, Kelly is in the second quadrant.

13. (a) The case with the largest  $Y$  value.

(b) The case with both the smallest  $X$  and  $Y$  values.

14. Deleting A will make the relationship weaker, while deleting B will make the relationship stronger. It follows that

- Data set 2 has  $r = -0.700$ .
- Data set 3 has  $r = -0.913$ .
- Data set 4 has  $r = -0.887$ .

15. (a). The slope of Tori's line is negative; thus, the correlation coefficient is negative.

16. (b). The slope of Brian's line is zero; thus, the correlation coefficient is zero.

17. First, construct the table

$x$	$y$	$\hat{y}$	$(y - \hat{y})^2$	$\ddot{y}$	$(y - \ddot{y})^2$
0	1	0	1	0	1
1	2	2	0	1	1
2	6	4	4	4	4
3	8	6	4	9	1
			9		7

The rule  $\ddot{y}$  is superior to  $\hat{y}$  because  $7 < 9$ .

18. For  $x = 2$ ,  $\hat{y} = \ddot{y} = 4$ ; thus, each predictor will have the same error and the same squared error.

19. (a)  $D < C < B < A$ .

(b)  $C < A < B < D$ .

(c)  $A < B < C < D$ .

(d)  $B < A < C < D$ .

(e)  $C < B < A < D$ .

20. It is stated in the text that the correlation coefficient for  $x$  and  $e$  is always 0. The values of  $x$  and  $\hat{y}$  all lie on the regression line. Thus, the correlation coefficient of these variables is  $-1$ ,

0, or 1. As a result, the given 0.73 must be the correlation coefficient for  $x$  and  $y$ , and, because it is greater than 0, we can conclude that the correlation coefficient for  $x$  and  $\hat{y}$  is 1.

21. It follows that  $x = 20$  yields  $\hat{y} = 50$ . Thus, the slope of the regression line is

$$(62 - 50)/(30 - 20) = 1.2.$$

But the slope also equals

$$r(s_Y/s_X) = 2r; \text{ thus, } r = 0.6.$$

Finally, the regression line is

$$\hat{y} = b_0 + b_1x.$$

Plugging in, for example,  $x = 20$  and  $\hat{y} = 50$  yields

$$50 = b_0 + 1.2(20) \text{ or } b_0 = 26.$$