Chapter 3: Describing Relationships

1. The following scatterplot describe the relationship between height (in cm) and foot length (also in cm) for 12 randomly selected students from the British Census @ Schools database. Which of the following is the best description of this relationship?

A. There is some random variation, but otherwise height is directly proportional to foot length.  
AR. Incorrect. In statistics, a description of the relationship between two quantitative variables should address the form, direction, and strength of the relationship, along with any observation that do not fit the overall pattern.
B. There is a roughly linear relationship between height and foot length.  
BR. Incorrect. In statistics, a description of the relationship between two quantitative variables should address the form, direction, and strength of the relationship, along with any observation that do not fit the overall pattern. This description only addresses form.
*C. There is a moderately weak, positive linear relationship between height and foot length.  
CR. Correct. This description addresses the form (linear), direction (positive), and strength (moderately weak) of the relationship. There are no observations that do not fit the overall pattern.

2. A sociologist is studying the relationship between early childhood nutrition and academic achievement in middle school among children in a certain city. Which of the following statements about the variable “early childhood nutrition” is correct?

A. Early childhood nutrition is a response variable.  
AR. Incorrect. The sociologist want to know if childhood nutrition has an impact on academic achievement.
*B. Early childhood nutrition is an explanatory variable.  
BR Correct. The sociologist want to know if childhood nutrition has an impact on academic achievement.
C. Since there is not a clear explanatory-response relationship in this scenario, we cannot classify early childhood nutrition as either explanatory or response.  
CR Incorrect. The sociologist want to know if childhood nutrition has an impact on academic achievement, so there is a clear explanatory-response relationship.
3. A sociologist studying the relationship between early childhood nutrition and academic achievement in middle school among children in a certain city finds that the correlation between these two variables is 0.86. Which of the following conclusions can he draw from this study?
A. Ensuring good nutrition in early childhood will increase academic achievement for middle school students.
AR. Incorrect. Correlation does not imply causation. There may be other variables involved that are more important. For example, both of these variables may be associated with the presence of more books in the home.
*B. Children in this city who have a healthy diet in early childhood tend to do better in middle school.
BR. Correct. While this correlation does not establish a causal link between these two variables, it is close to +1, implying a strong association.
C. Since the correlation is so low, no conclusions can be drawn.
CR. Incorrect. A correlation of 0.86 is close to the maximum possible value of +1, suggesting a relatively strong association between these two variables.

4. Which of the following quantities is minimized by the least-squares regression line?
*A. The sum of the squared differences between observed values of the response variable and values of the response variable predicted by the model.
AR. Correct. The least-squares line is the line that makes the sum of the squared residuals as small as possible.
B. The sum of the squares of perpendicular distances between all data points and the regression line.
BR. Incorrect. The least-squares line is based on distances between observed values and the regression line, but not the perpendicular distance.
C. The sum of the squared differences between observed values of the explanatory variable and values of the explanatory variable predicted by the model.
AR. Incorrect. The least-squares line is based on distances between observed values and the regression line, but not the distances described in this choice.

5. Which of the following statements about the slope of the least-squares regression line is true?
*A. It has the same sign as the correlation coefficient $r$.
AR. Correct. Recall that $b = r(s_y/s_x)$, where $b$ is the slope of the line, $r$ is the correlation coefficient, and the ratio of the sample standard deviations $s_y/s_x$ is always positive. Therefore, $b$ and $r$ must always have the same sign.
B. The square of the slope equals the proportion of the variation in the response variable that is explained by the explanatory variable.
BR. Incorrect. The square of the correlation $r$ has this property, not the square of the slope of the line. Although the slope and the correlation have the same sign, their numerical values and other properties are different.
C. It is unitless.
CR. Incorrect. The units of the slope are units of $y$ divided by units of $x$. The correlation coefficient, not the slope, is unitless. Although the slope and the correlation have the same sign, their numerical values and other properties are different.
6. The points in the scatterplot represent paired observations \((x, y)\) where \(x\) is an individual’s weight and \(y\) is the time (in seconds) it takes for walking on a treadmill to raise the individual’s pulse rate to 140 beats per minute. The open circles correspond to females and the dark squares to males.

From the scatterplot, which conclusion we can make?
A. There is a positive correlation \(r\) between gender and weight, since men tend to weigh more than women.
AR Incorrect. The correlation \(r\) measures association between two quantitative variables. Since “gender” is not quantitative, \(r\) is not an appropriate measure of association between gender and weight.
*B. There is a negative correlation \(r\) between weight and time for both males and females.
BR Correct. If one looks only at the o’s corresponding to females, there is a clear downward trend indicating a negative correlation between weight and time. The same is true for the +’s corresponding to males.
C. In general, males tend to take less time to have their pulse rate raised to 140 bpm while walking on the treadmill.
CR Incorrect. If this statement were true, then the cluster of +’s corresponding to males would lie distinctly below the cluster of o’s corresponding to females (that is, the +’s would tend to have smaller \(y\)-coordinates than the o’s), clearly not the case here.

7. One of the following is a correct statement involving correlation. The other two contain blunders. Which one is correct?
A. There is a correlation of \(r = 0.54\) between the position a football player plays and his or her weight.
AR Incorrect. This statement contains a blunder because \(r\) measures association between quantitative variables and the variable “position” is categorical.
*B. The correlation between amount of fertilizer and yield of tomatoes was found to be \(r = 0.33\).
BR Correct. Both variables are quantitative in this case, so \(r\) would be an appropriate measure of association. Also, in this case, we would expect the correlation to be positive.
C. The correlation between the gas mileage of a car and its weight is \(r = -0.71\) gallon-pounds.
CR Incorrect. This statement contains a blunder because the correlation \(r\) has no units. If the value had been \(r = 0.71\), then the statement would have been correct.
8. A study showed that students who spend more time studying for statistics tests tend to achieve better scores on their tests. In fact, the number of hours studied turned out to explain 81% of the observed variation in test scores among the students who participated in the study. What is the value of the correlation between number of hours studied and test score?

A. \( r = 0.81 \)  
AR Incorrect. You have misinterpreted \( r^2 \), the proportion of variation in the response variable that can be explained by regression on the explanatory variable, as \( r \).

B. \( r = 0.656 \)  
BR Incorrect. You performed the incorrect operation to transform the given information, which depends on \( r \), into the value of \( r \).

*C. \( r = 0.9 \)  
CR Correct. The given information implies that \( r^2 = 0.81 \). Taking the (positive) square root yields \( r = 0.9 \). (We use the positive root because the first sentence of the problem indicates that the direction of the association between number of hours studied and test score is positive.)

9. The following computer output describes the relationship between \( y = \) height (in cm) and \( x = \) foot length (also in cm) for 12 randomly selected students from the British Census @ Schools database. The scatterplot for this relationship show a roughly linear shape.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>117.99</td>
<td>28.39</td>
<td>4.16</td>
<td>0.002</td>
</tr>
<tr>
<td>Foot length (cm)</td>
<td>1.878</td>
<td>1.155</td>
<td>1.63</td>
<td>0.135</td>
</tr>
</tbody>
</table>

\( S = 7.39858 \quad R^2 = 20.9\% \quad R^2(\text{adj}) = 13.0\% \)

Which of the following is an equation of least-squares regression line for these data?

*A. Height = 117.99 + 1.878( Foot length)  
AR Correct. The “Coef” column lists the y-intercept in the “Constant” row of the “Coef” (Coefficient) column and the slope in the row designated by the explanatory variable, foot length.

B. Foot length = 117.99 + 1.878(Height)  
BR Incorrect. When “Foot length” is listed as a predictor, it is the explanatory variable.

C. Height = 1.878 + 117.99(Foot length)  
CR Incorrect. Slope is the coefficient listed in the “Foot length” row, and the y-intercept is in the “Constant” row.
10. The following computer output describes the relationship between $y =$ height (in cm) and $x =$ foot length (also in cm) for 12 randomly selected students from the British Census @ Schools database. The scatterplot for this relationship show a roughly linear shape.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>117.99</td>
<td>28.39</td>
<td>4.16</td>
<td>0.002</td>
</tr>
<tr>
<td>Foot length (cm)</td>
<td>1.878</td>
<td>1.155</td>
<td>1.63</td>
<td>0.135</td>
</tr>
</tbody>
</table>

$s =$ 7.39858   R-Sq = 20.9%  R-Sq(adj) = 13.0%

Which of the following is the correct interpretation of the number $s =$ 7.39858?

A. For each one-centimeter increase in foot length the model predicts an increase in height of 7.39858 centimeters.
AR. Incorrect. This is an interpretation of the slope of the regression line, which in this case is 1.878.
*B. If we use the regression equation to predict height from foot length, our predictions will be, on average, off by 7.39858 centimeters.
BR. Correct. $s$ is the standard deviation of residuals, and describes typical prediction error.
C. The typical difference between the height of a student and the mean student height is 7.39858.
CR. Incorrect. This is an interpretation of the standard deviation of height.

11. Below is a residual plot for the regression of the number of employees of Microsoft Inc. on year for the years from 1976 to 1989. (Note that this is a residual plot, not a scatterplot!)

There is no observed value for the year 1983. If we were to use this regression to predicted the number of employees in 1983, which of the following is most likely to describe the accuracy of our prediction?

*A. Too high
AR. Correct Since all the residuals between 1980 and 1987 are negative, predicted values for number of employees are higher than observed values. It’s likely that this will be true for 1983 as well.
B. Too low
BR. Incorrect. Use the sign of the residuals for years near 1983 to compare observed values for number of employees to predicted values.
C. About right
CR. Incorrect. Use the sign of the residuals for years near 1983 to compare observed values for number of employees to predicted values.

12. Ms. Kreppel is interested in the relationship between her students’ final exam scores and their scores on a pre-test they took at the beginning of the year. Below is a scatterplot showing this relationship for the 18 students in her class.

![Scatterplot](image)

How would the slope of the least-squares regression line change if the individual whose point is circled were removed from the data set?
* A. The slope would increase.
AR. Correct. With this “influential point” removed from the data, the remaining points exhibit a larger change in final exam grade for a given change in pre-test grade.
B. The slope would decrease.
BR. Incorrect. With this “influential point” removed from the data, the remaining points exhibit a larger change in final exam grade for a given change in pre-test grade.
C. The slope would be unchanged
CR. Incorrect. The slope of the least-squares line is strongly influenced by points that lie outside the general pattern and have high residuals.

13. Ms. Kreppel is interested in the relationship between her students’ final exam scores and their scores on a pre-test they took at the beginning of the year. A scatterplot of the data for the 18 students in her class shows linear relationship for these variables. The equation of the least-squares regression line is

\[
\text{Final Exam} = 34.2 + 0.60(\text{Pre-test})
\]

Which of the following is a correct interpretation of the slope of this regression model?
A. For each one-unit increase in final exam score, the model predicts, on average, a 0.60 unit increase in pre-test score.
AR. Incorrect. Reconsider which variable is explanatory and which is response!
* B. For each one-unit increase in pre-test score, the model predicts, on average, a 0.60 unit increase in exam score.
BR. Correct. Slope describes the amount by which the response variable is expected to change for a one-unit change in the explanatory variable.
C. About 60% of the variation in exam score that is accounted for by the regression of exam score on pre-test score.
14. Ms. Kreppel is interested in the relationship between her students’ final exam scores and their scores on a pre-test they took at the beginning of the year. A scatterplot of the data for the 18 students in her class shows linear relationship for these variables. The equation of the least-squares regression line is $\text{Final Exam} = 34.2 + 0.60(\text{Pre-test})$. One student scored a 76 on the pre-test and an 82 on the final exam. Which of the following is that student’s residual?

*A. 2.2
AR. Correct. Residual = $y_{\text{observed}} - \hat{y} = 82 - (34.2 + 0.6(76)) = 2.2$
B. –7.4
BR. Incorrect. Did you confuse the explanatory and response variables?
C. –2.2
CR. Recall that Residual = $y_{\text{observed}} - \hat{y}$.
13. A residual plot displays a “reverse fan” arrangement, with the spread of points about the line (residual = 0) gradually decreasing from left to right (that is, as \( x \) increases). Which statement would be a correct interpretation of this plot?

A. The original data display a nonlinear relationship (curved pattern of association).
AR Incorrect. If the association were nonlinear, then the residual plot would display a curved pattern of some sort. The residual plot described here may not result from such an association.

B. Predictions using the regression line will be more reliable for small \( x \) than for large \( x \).
BR Incorrect. According to the plot, the linear fit is better for large \( x \) than for small \( x \), so predictions will be more reliable for large \( x \) than for small \( x \).

*C. Predictions using the regression line will be more reliable for large \( x \) than for small \( x \).
CR Correct. According to the plot, the linear fit is better for large \( x \) than for small \( x \), so predictions will be more reliable for large \( x \) than for small \( x \).