Chapter 3: Describing Relationships

Key Vocabulary:
- response variable
- explanatory variable
- independent variable
- dependent variable
- scatterplot
- positive association
- negative association
- linear
- correlation
- r-value
- regression line
- mathematical model
- least-squares regression line
- \( \hat{y} \) "y-hat"
- SSM
- SSE
- \( r^2 \)
- coefficient of determination
- residuals
- residual plot
- influential observation

3.1 Scatterplots and Correlation (pp.142-163)
1. What is the difference between a response variable and an explanatory variable? A response variable measures an outcome of a study. An explanatory variable may help explain or influence changes in a response variable.

2. How are response and explanatory variables related to dependent and independent variables? Explanatory variables are often referred to as independent variables and response variables are often referred to as dependent variables.

3. When is it appropriate to use a scatterplot to display data? When the goal is to display or compare the relationship of two quantitative variables.

4. Which variable always appears on the horizontal axis of a scatterplot? The explanatory variable

5. You can describe the overall pattern of a scatterplot by the direction, form, and strength.

6. Explain the difference between a positive association and a negative association. When the overall pattern moves from upper left to lower right, it’s called a negative association. When the overall pattern moves from lower left to upper right, it’s called a positive association.
7. What does correlation measure? Correlation measures the strength of the linear relationship between two quantitative variables.

8. What is true about the relationship between two variables if the r-value is:
   a. Near 0? - very weak to almost no linear relationship
   b. Near 1? - very strong to almost perfectly linear relationship
   c. Near -1? – very strong to almost perfectly linear relationship
   d. Exactly 1? – perfect linear relationship
   e. Exactly -1? – perfect linear relationship

9. Is correlation resistant to extreme observations? Explain. Correlation is not resistant to extreme observations, therefore outliers can greatly change the value of the correlation.

10. What does it mean if two variables have high correlation? If there is a high correlation, then the two variables closely agree and the scatter plot of the quantitative variables tend to follow a moderate to strong linear pattern.

11. What does it mean if two variables have weak correlation? If there is a weak correlation, then the two quantitative variables do not closely agree and the scatter plot of the variables tend to follow a scattered or non-linear pattern.

12. What does it mean if two variables have no correlation? If there is no correlation, then the strength of a linear relationship between two quantitative variables is non-existent.

13. Explain why two variables must both be quantitative in order to find the correlation between them. Scatterplots are the only choice for displaying the relationship between two quantitative variables. Quantitative variables are needed because correlation is measured numerically using the following interval: \(-1 \leq r \leq 1\)

14. Does a correlation close to -1 or 1 always guarantee a linear relationship? Explain. No. A scatterplot with a correlation that’s close to -1 or 1 can have a curved form so it is necessary to always plot the given data. It is important to remember that you can calculate a correlation for any scatter plot; however r only measures straight-line or linear relationships.
3.2 Least-Squares Regression (pp.164-197)

1. What is a regression line? A straight line that describes how a response variable y changes as an explanatory variable x changes. Basically, it is a line used to model the data.

2. What is extrapolation and why is this dangerous? Extrapolation involves using a regression line to make predictions far outside the range of values of x that are used to obtain the line. Proceed with caution when making these predictions – often times they are not accurate.

3. What is a least-squares regression line? The least-squares regression line of y on x is a type of regression line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. (also called ‘line of best fit’) In other words, a least-squares regression line makes the errors in predicting y as small as possible by minimizing the sum of the squares of the residuals.

4. What is the formula for the equation of the least-squares regression line? Define each variable. Least-squares line: \( \hat{y} = a + bx \); y-hat is the predicted value of y given a specific x; a is the y intercept – the predicted value of y when x is 0; b is the slope – the amount by which y is predicted to change when x increases by 1 unit.

5. The least-squares regression line always passes through the point (x bar, y bar).

6. What is a residual? A residual is the difference between an observed value of y and the value of y predicted by the regression line; i.e. (residual = observed y – predicted y = y - \( \hat{y} \)) In essence, a residual is an observable estimate of the unobservable statistical error.

7. What special property do the residuals have? The mean of the residuals will always equal zero. Why do they have this property? Because the sum of the residuals always equal zero.

8. What is a residual plot? A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. In other words, it is a scatter plot of the residuals against x (the explanatory variable) and turns the regression line horizontal.

9. How can you calculate residuals on your calculator and use this to produce a residual plot? To calculate residuals on your calculator: With data in Lists 1 and 2, enter y hat in highlighted L3 but replace the x in y hat with (L1) and press enter. Next with L4 highlighted, enter L2 minus L3 and press enter to see the residuals displayed in L4. To produce a residual plot: Create (turn on) a plot with L1 as the x variable and L4 as the y variable. Use ZOOM 9 to see the residual plot. NOTE: To compute the standard deviation of the residuals, calculate a 1-VAR STAT on the residual list.

10. If a least-squares regression line fits the data well, what characteristics should the residual plot exhibit? Residual plots help us assess how well a regression line fits the data. If the points in a residual plot are randomly dispersed around the horizontal axis, then a linear regression model is appropriate for the data; otherwise a non-linear model is appropriate.
11. What does the standard deviation of the residuals tell us? The standard deviation is roughly the average distance of the actual values from the least-squares regression line. It tells us the average distance of actual values from their expected values and thus measures prediction error.

12. How is the coefficient of determination defined? As $r^2$ – which is the square of the correlation coefficient, $r$. The coefficient of determination is equal to the percent of variation in one variable that is accounted for (predicted) by the other variable.

13. If $r^2 = 0.95$, what can be concluded about the relationship between $x$ and $y$? 95% of the variation in $y$ is explained by the linear model relating $y$ to $x$. Remember, $r^2$ is a reported measure of how successful the regression model was in explaining the response variable.

14. What are three limitations of correlation and regression? 1. The distinction between explanatory and response variables is important in regression; reversing $x$ and $y$ will yield a different least-squares regression line. 2). Correlation and regression lines describe only linear relationships. 3). Correlation and least-squares regression lines are not resistant.

15. Under what conditions does an outlier become an influential observation? An observation is influential for a statistical calculation if removing it noticeably changes the result of the calculation. Points that are outliers in the $x$ direction of a scatterplot are often influential for the least-squares regression line. Influential points often have small residuals, because they pull the regression line toward themselves.

16. What is a lurking variable? A hidden variable in a study, research, experiment, etc. that may affect the predictor variables given. In other words, a lurking variable is not among the explanatory or response variables in a study, but it may influence the variation in the response variable.

17. Why does association not imply causation? A strong association between two variables is not enough evidence to draw conclusions about cause and effect.