Chapter 5

CHAPTER 5

RELATIONSHIPS BETWEEN QUANTITATIVE VARIABLES

In this chapter, we will learn how to describe the relationship between two quantitative variables. Remember (from Chapter 2) that the terms quantitative variable and measurement variable are synonyms for data that can be recorded as numerical values and then ordered according to those values. The relationship between weight and height is an example of a relationship between two quantitative variables.

The questions we ask about the relationship between two variables often concern specific numerical features of the association. For example, we may want to know how much weight will increase on average for each one-inch increase in height. Or, we may want to estimate what the college grade point average will be for a student whose high school grade point average was 3.5.

We will use three tools to describe, picture and quantify the relationship between two quantitative variables:
- **Scatter plot**, a two-dimensional graph of data values.
- **Correlation**, a statistic that measures the strength and direction of a linear relationship between two quantitative variables.
- **Regression equation**, an equation that describes the average relationship between a quantitative response variable and an explanatory variable.

5.1 Looking for Patterns with Scatter Plots

A scatter plot is a two dimensional graph of the measurements for two numerical variables. A point on the graph represents the combination of measurements for an individual observation. The vertical axis, which is called the y-axis, is used to locate the value of one of the variables. The horizontal axis, called the x-axis, is used to locate the values of the other variable.

As we learned in Chapter 2, when looking at relationships we can often identify one of the variables as an explanatory variable that may explain or cause differences in the response variable. The term dependent variable is used as a synonym for response variable. In a scatter plot, the response variable is plotted on the vertical axis (the y-axis), so it may also be called the y variable. The explanatory variable is plotted along the horizontal axis (the x-axis) and may be called the x variable.

**Questions to Ask about a Scatter Plot**
- What is the average pattern? Does it look like a straight line or is it curved?
- What is the direction of the pattern?
- How much do individual points vary from the average pattern?
- Are there any unusual data points?
Example 1. Height and Hand-Span
Table 5.1 displays the first 12 observations of a data set that includes the heights (in inches) and fully stretched hand-spans (in centimeters) of 167 college students. The data values for all 167 students are the raw data for studying the connection between height and hand-span. Imagine how difficult it would be see the pattern in the data if all 167 observations were shown in Table 5.1. Even when we just look at the data for 12 students, it takes a while to confirm that there does seem to be a tendency for taller people to have larger hand-spans.

Figure 5.1 is a scatter plot that displays the hand-span and height measurements for all 167 students. The hand-span measurements are plotted along the vertical axis (y) and the height measurements are plotted along the horizontal axis. Each point represents the two measurements for an individual.

Table 5.1 Hand-Spans and Height

<table>
<thead>
<tr>
<th>Ht.(in)</th>
<th>Span(cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>23.5</td>
</tr>
<tr>
<td>69</td>
<td>22.0</td>
</tr>
<tr>
<td>66</td>
<td>18.5</td>
</tr>
<tr>
<td>64</td>
<td>20.5</td>
</tr>
<tr>
<td>71</td>
<td>21.0</td>
</tr>
<tr>
<td>72</td>
<td>24.0</td>
</tr>
<tr>
<td>67</td>
<td>19.5</td>
</tr>
<tr>
<td>65</td>
<td>20.5</td>
</tr>
<tr>
<td>76</td>
<td>24.5</td>
</tr>
<tr>
<td>67</td>
<td>20.0</td>
</tr>
<tr>
<td>70</td>
<td>23.0</td>
</tr>
<tr>
<td>62</td>
<td>17.0</td>
</tr>
</tbody>
</table>

and so on, for n=167 observations

Figure 5.1. Hand-Span and Height

We see that taller people tend to have greater hand-span measurements than shorter people do. When two variables tend to increase together, as they do in Figure 5.1, we say that they have a **positive association**. Another noteworthy characteristic of the graph is that we can
describe the general pattern of this relationship with a straight line. In other words, the hand-span and height measurements may have a linear relationship.

**Positive and Negative Association**
- Two variables have a **positive association** when the values of one variable tend to increase as the values of the other variable increase.
- Two variables have a **negative association** when the values of one variable tend to decrease as the values of the other variable increase.

**Example 2. Driver Age and the Maximum Legibility Distance of Highway Signs**
In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of thirty drivers could read a newly designed sign. The thirty participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers, and wanted to examine the relationship between age and the sign legibility distance.

Table 5.2 lists the data and Figure 5.2 shows a scatter plot of the ages and distances. The sign legibility distance is the response variable so that variable is plotted on the y-axis (the vertical axis). The maximum reading distance tends to decrease as age increases, so there is a negative association between distance and age. This is not a surprising result. As a person gets older, his or her eyesight tends to get worse so we would expect the distances to decrease with age.

The researchers collected the data to determine numerical estimates for two questions about the relationship:
- How much does the distance decrease when age is increased?
- For drivers of any specific age, what is the average distance at which the sign can be read?

We’ll examine these questions in the next section. For now, we simply point out that the pattern in the graph looks linear, so a straight line equation that links distance to age will help us answer these questions.

<table>
<thead>
<tr>
<th>Age</th>
<th>Distance</th>
<th>Age</th>
<th>Distance</th>
<th>Age</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>510</td>
<td>37</td>
<td>420</td>
<td>68</td>
<td>300</td>
</tr>
<tr>
<td>20</td>
<td>590</td>
<td>41</td>
<td>460</td>
<td>70</td>
<td>390</td>
</tr>
<tr>
<td>22</td>
<td>560</td>
<td>46</td>
<td>450</td>
<td>71</td>
<td>320</td>
</tr>
<tr>
<td>23</td>
<td>510</td>
<td>49</td>
<td>380</td>
<td>72</td>
<td>370</td>
</tr>
<tr>
<td>23</td>
<td>460</td>
<td>53</td>
<td>460</td>
<td>73</td>
<td>280</td>
</tr>
<tr>
<td>25</td>
<td>490</td>
<td>55</td>
<td>420</td>
<td>74</td>
<td>420</td>
</tr>
<tr>
<td>27</td>
<td>560</td>
<td>63</td>
<td>350</td>
<td>75</td>
<td>460</td>
</tr>
<tr>
<td>28</td>
<td>510</td>
<td>65</td>
<td>420</td>
<td>77</td>
<td>360</td>
</tr>
<tr>
<td>29</td>
<td>460</td>
<td>66</td>
<td>300</td>
<td>79</td>
<td>310</td>
</tr>
<tr>
<td>32</td>
<td>410</td>
<td>67</td>
<td>410</td>
<td>82</td>
<td>360</td>
</tr>
</tbody>
</table>
5.2 Describing Linear Patterns with a Regression Line

Scatter plots show us a lot about a relationship, but we often want more specific numerical descriptions of how the response and explanatory variables are related. Imagine, for example, that we are examining the weights and heights of a sample of college women. We might want to know what the increase in average weight is for each one-inch increase in height. Or, we might want to estimate the average weight for women with a specific height, like 5’10’’.

Regression analysis is the area of statistics used to examine the relationship between a quantitative response variable and one or more explanatory variables. A key element of regression analysis is the estimation of an equation that describes how, on average, the response variable is related to the explanatory variables. This regression equation can be used to answer the types of questions that we just asked about the weights and heights of college women.

A regression equation can also be used to make predictions. For instance, it might be useful for colleges to have an equation for the connection between verbal SAT score and college grade point average (GPA). They could use that equation to predict the potential GPAs of future students, based on their verbal SAT scores. Some colleges actually do this kind of prediction to decide whom to admit, but they use a collection of variables to predict GPA. The prediction equation for GPA usually includes high school GPA, high school rank, verbal and math SAT scores, and possibly other factors such as a rating of the student’s high school or the quality of an application essay.

There are many types of relationships and many types of regression equations. The simplest kind of relationship between two variables is a straight line, and that’s the only type we will discuss here. Straight-line relationships occur frequently in practice, so this is a useful and important type of regression equation. Before we use a straight line regression model, however, we should always examine a scatterplot to verify that the pattern actually is linear. We remind you of the music preference and age example where a straight line definitely does not describe the pattern of the data.

Interpreting the Regression Equation and Regression Line

When the best equation for describing the relationship between x and y is a straight line the resulting equation is called the regression line. This line is used for two purposes:
- to estimate the average value of y at any specified value of x
- to predict the value of y for an individual, given that individual's x value
Example 1 Revisited. Describing Height and Hand-Span with a Regression Line

In Figure 5.1, we saw that the relationship between hand-span and height has a straight-line pattern. Figure 5.6 displays the same scatterplot as Figure 5.1, but now a line is shown that describes the average relationship between the two variables. To determine this line, we used statistical software (Minitab) to find the “best” line for this set of measurements. We’ll discuss the criterion for “best” later. Presently, let’s focus on what the line tells us about the data.

The line drawn through the scatterplot is the regression line and it describes how average hand-span is linked to height. For example, when the height is 60 inches, the vertical position of the line is at about 18 centimeters. To see this, locate 60 inches along the horizontal axis (x axis), look up to the line, and then read the vertical axis to determine the hand-span value. The result is that we can estimate that people 60 inches tall have an average hand-span of about 18 centimeters (roughly 7 inches). We can also use the line to predict the hand-span for an individual whose height is known. For instance, someone 60 inches tall is predicted to have a hand-span of about 18 centimeters.

Figure 5.6 Regression Line Describing Hand-Span and Height

If we estimate the average hand-span at a different height, we can determine how much hand-span changes, on average, when height is varied. Let’s use the line to estimate the average hand-span for people who are 70 inches tall. We see that the vertical location of the regression line is somewhere betweeen 21 and 22 centimeters, perhaps about 21.5 centimeters (roughly 8.5 inches). So, when height is increased from 60 inches to 70 inches, average hand-span increases from about 18 centimeters to about 21.5 centimeters.

The average hand-span increased by 3.5 centimeters (about 1.5 inches) when the height was increased by 10 inches. This is a rate of 3.5/10 = 0.35 centimeters per one inch increase in height, which is the slope of the line. For each one-inch difference in height, there is about a 0.35 centimeter average difference in hand-span.
Algebra Reminder
The equation for a straight line relating $y$ and $x$ is:

$$y = b_0 + b_1 x$$

where $b_0$ is the "y-intercept" and $b_1$ is the slope. When $x = 0$, $y = y$-intercept.

The slope of a line can be determined by picking any two points on the line, and then calculating

$$\text{slope} = \frac{\text{difference between } y \text{ values}}{\text{difference between } x \text{ values}} = \frac{y_2 - y_1}{x_2 - x_1}$$

The letter $y$ represents the vertical direction and $x$ represents the horizontal direction. The slope tells us how much the $y$ variable changes for each increase of one unit in the $x$ variable.

We ordinarily don’t have to “read” the regression line as we just did. Statistical software will tell us the regression equation, the specific equation used to draw the line.

For the hand-span and height relationship, the regression equation determined by statistical software is:

$$\text{Hand-span} = -3 + 0.35 \text{ Height}.$$ 

When emphasis is on using the equation to estimate the average hand-spans for specific heights, we may write:

$$\text{Average Hand-span} = -3 + 0.35 \text{ Height}$$

When emphasis is on using the equation to predict an individual hand-span, we might instead write:

$$\text{Predicted Hand-span} = -3 + 0.35 \text{ Height}$$

In most situations, the correct statistical interpretation of a regression equation is that it estimates the average value of a response variable ($y$) for individuals with a specific value of the explanatory variable ($x$). The equation $\text{Hand-span} = -3 + 0.35 \text{ Height}$ tells us how to draw the line, but not all individuals follow this pattern exactly. Look again at Figure 5.6, in which we see that the line describes the overall pattern, but we also see substantial individual deviation from this line.

Let’s use the regression equation to estimate the average hand-spans for some specific heights.

- For height=60, average hand-span $= -3 + 0.35(60) = -3 + 21 = 18$ cm.
- For height=70, average hand-span $= -3 + 0.35(70) = -3 + 24.5 = 21.5$ cm

In the equation, the value 0.35 multiplies the height. This value is the slope of the straight line that links hand-span and height. Consistent with our estimates above, the slope in this example tells us that hand-span increases by 0.35 centimeters, on average, for each increase of one inch in height. We can use the slope to estimate the average difference in hand-span for any difference in height. If we consider two heights that differ by 7 inches, our estimate of the difference in hand-spans would be $7 \times 0.35 = 2.45$ centimeters, or approximately one inch.

The Equation for the Regression Line
All straight lines can be expressed by the same formula in which $y$ is the variable on the vertical axis and $x$ is the variable on the horizontal axis. The equation for a regression line is:

$$\hat{y} = b_0 + b_1 x.$$ 

In any given situation, the sample is used to determine numbers that replace $b_0$ and $b_1$.

- $\hat{y}$ is spoken as “$y$-hat” and it is also referred to either as predicted $y$ or estimated $y$.
- $b_0$ is the intercept of the straight line. The intercept is the value of $y$ when $x = 0$.
- $b_1$ is the slope of the straight line. The slope tells us how much of an increase (or decrease) there is for the $y$ variable when the $x$ variable increases by one unit. The sign of the slope tells us whether $y$ increases or decreases when $x$ increases.
Interpreting a Regression Line

- \( \hat{y} \) estimates the average \( y \) for a specific value of \( x \). It also can be used as a prediction of the value of \( y \) for an individual with a specific value of \( x \).
- The slope of the line estimates the average increase in \( y \) for each one unit increase in \( x \).
- The intercept of the line is the value of \( y \) when \( x=0 \). Note that interpreting the intercept in the context of statistical data only makes sense if \( x=0 \) is included in the range of observed \( x \)-values.

Example 2 Revisited: Driver Age and the Maximum Legibility Distance of Highway Signs

The regression line \( \hat{y} = 577 - 3x \) describes how the maximum sign legibility distance (the \( y \) variable) is related to driver age (the \( x \) variable). Statistical software was used to calculate this equation and to create the graph shown in Figure 5.7. Earlier, we asked these two questions about distance and age:

- How much does the distance decrease when age is increased?
- For drivers of any specific age, what is the average distance at which the sign can be read?

The slope of the equation can be used to answer the first question. Remember that the slope is the number that multiplies the \( x \) variable and the sign of the slope indicates the direction of the association. Here, the slope tells us that, on average, the legibility distance decreases 3 feet when age increases by one year. This information can be used to estimate the average change in distance for any difference in ages. For an age increase of 30 years, the estimated decrease in legibility distance is 90 feet because the slope is –3 feet per year.

The question about estimating the average legibility distances for a specific age is answered by using the specific age as the \( x \) value in the regression equation. To emphasize this use of the regression line, we write it as:

\[
\text{Average distance} = 577 - 3\, \text{Age}
\]

Here are the results for three different ages:

<table>
<thead>
<tr>
<th>AGE</th>
<th>AVERAGE DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>577 - 3(20) = 517 feet</td>
</tr>
<tr>
<td>50</td>
<td>577 - 3(50) = 427 feet</td>
</tr>
<tr>
<td>80</td>
<td>577 - 3(80) = 337 feet</td>
</tr>
</tbody>
</table>

The equation can also be used to predict the distance measurement for an individual driver with a specific age. To emphasize this use of the regression line, we write the equation as:

\[
\text{Predicted distance} = 577 - 3\, \text{Age}
\]

For example, we can predict that the legibility distance for a 20-year old will be 517 feet and for an 80-year old will be 337 feet.
The Least Squares Line, Errors and Residuals

We can use statistical software to estimate the regression line, but how does the computer find the best equation for a set of data? The most commonly used method is called least squares and the regression line determined by this method is called the least squares line.

The phrase “least squares” is actually a shortened version of “least sum of squared errors.” This criterion focuses on the differences between the values of the response variable (y) and the regression line. The response variable is emphasized because we often use the equation to predict that variable for specific values of the explanatory variable (x). Therefore, we should minimize how far off the predictions will be in that direction.

For any given line, we can calculate the predicted value of y for each point in the observed data. To do this for any particular point, we use the observed x value in the equation. We then determine the prediction “error” for each point. An “error” is simply the difference between the observed y value and the predicted value  \( \hat{y} \). These errors are squared and added up for all of the points in the sample. The least squares line minimizes the sum of the squared errors. This terminology is somewhat misleading, since the amount by which individual differs from the line is seldom due to "errors" in the measurements. A more neutral term for the difference ( \( y - \hat{y} \) ) is that it is the residual for that individual.

<table>
<thead>
<tr>
<th>The Least Squares Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>When we use a line to predict the values of y, the sum of squared differences between the observed values of y and the predicted values is smaller for the least squares line than it is for any other line.</td>
</tr>
</tbody>
</table>

There is a mathematical solution that produces general formulas for the slope and intercept of the least squares line. These formulas are used by all statistical software, spreadsheet programs, and statistical calculators. To be complete, we include the formulas. You won’t need them if you use the computer to do a regression analysis.
Formulas for the slope and intercept of the least squares line

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

\[ b_0 = \bar{y} - b_1\bar{x} \]

- \( x_i \) represents the x measurement for the \( i^{th} \) observation.
- \( y_i \) represents the y measurement for the \( i^{th} \) observation.
- \( \bar{x} \) represents the mean of the x measurements.
- \( \bar{y} \) represents the mean of the y measurements.

Example 2 Revisited. Errors for the Highway Sign data

The least squares regression line for Example 2 is \( \hat{y} = 577 - 3x \) where \( y \) = maximum sign legibility distance and \( x \) = driver age. For this equation, the calculation of the errors and the squared errors for the first three data points shown in Table 5.2 is:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>( \hat{y} = 577 - 3x )</th>
<th>error = ( y - \hat{y} )</th>
<th>squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>510</td>
<td>577 - 3(18) = 523</td>
<td>510 - 523 = -13</td>
<td>169</td>
</tr>
<tr>
<td>20</td>
<td>590</td>
<td>577 - 3(20) = 517</td>
<td>590 - 517 = 73</td>
<td>5329</td>
</tr>
<tr>
<td>22</td>
<td>516</td>
<td>577 - 3(22) = 511</td>
<td>516 - 511 = 5</td>
<td>25</td>
</tr>
</tbody>
</table>

This process can be carried out for all 30 data points. The sum of the squared errors is smaller for the line \( \hat{y} = 577 - 3x \) than it would be for any other line.
Chapter 5

5.3 Measuring Strength and Direction with Correlation

The linear pattern is so common that a statistic was created to characterize this type of relationship. The statistical correlation between two quantitative variables is a number that indicates the strength and the direction of a straight-line relationship.

- The strength of relationship is determined by the closeness of the points to a straight line.
- The direction is determined by whether one variable generally increases or generally decreases when the other variable increases.

As used in statistics, the meaning of the word “correlation” is much more specific than it is in everyday life. A statistical correlation only describes linear relationships. Whenever a correlation is calculated, a straight line is used as the frame of reference for evaluating the relationship. When the pattern is nonlinear, as it was for the music preference data shown in Figure 5.3, a correlation is not an appropriate way to measure the strength of the relationship.

Correlation is represented by the letter \( r \). Sometimes this measure is called the "Pearson product moment correlation" or the "correlation coefficient." The formula for correlation is complicated. Fortunately, all statistical software programs and many calculators provide a way to easily calculate this statistic.

### A Formula for Correlation

\[
r = \frac{1}{n-1} \sum_{i} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

- \( n \) is the sample size.
- \( x_i \) is the x measurement for the \( i^{th} \) observation.
- \( \bar{x} \) is the mean of the x measurements.
- \( s_x \) is the standard deviation of the x measurements.
- \( y_i \) is the y measurement for the \( i^{th} \) observation.
- \( \bar{y} \) is the mean of the y measurements.
- \( s_y \) is the standard deviation of the y measurements.

### Interpreting the Correlation Coefficient

Some specific features of the correlation coefficient are:

- Correlation coefficients are always between \(-1\) and \(+1\).
- The magnitude of the correlation indicates the strength of the relationship, which is the overall closeness of the points to a straight line. The sign of the correlation does not tell us about the strength the linear relationship.
- A correlation of either \(+1\) or \(-1\) indicates that there is a perfect linear relationship and all data points fall on the same straight line.
- The sign of the correlation indicates the direction of the relationship. A positive correlation indicates that the two variables tend to increase together (a positive association). A negative correlation indicates that when one variable increases the other is likely to decrease (a negative association).
- A correlation of 0 indicates that the best straight line through the data is exactly horizontal, so knowing the value of x does not change the predicted value of y.
Example 1 Revisited. The correlation between hand-span and height
The relationship between hand-span and height appears to be linear, so a correlation is useful for characterizing the strength of the relationship. For these data, the correlation is \( r = +0.74 \), a value that indicates a somewhat strong positive relationship. A look back at Figure 5.1 shows us that average hand-span definitely increases when height increases, but within any specific height there is some natural variation among individual hand-spans.

Example 2 Revisited. The Correlation between Age and Sign Legibility Distance
The correlation for the data shown in Figure 5.2 is \( r = -0.8 \), a value that indicates a somewhat strong negative association between the variables.

The Interpretation of \( r^2 \), the Squared Correlation
The squared value of the correlation is frequently used to describe the strength of a relationship. A squared correlation, \( r^2 \), always has a value between 0 and 1, although some computer programs will express its value as a percentage between 0 and 100%. By squaring the correlation, we retain information about the strength of the relationship, but we lose information about the direction.

Researchers typically use the phrase “proportion of variation explained by \( x \)” in conjunction with the squared correlation, \( r^2 \). For example, if \( r^2 = 0.60 \) (or 60%), the researcher may write that the explanatory variable explains 60% of the variation in the response variable. If \( r^2 = 0.10 \) (or 10%), the explanatory variable only explains 10% of the variation in the response variable. This interpretation stems from the use of the least squares line as a prediction tool.

Let’s consider Example 1 again. For that example, the regression equation is
\[
\text{hand-span} = -3.5 + 0.35 \text{ height}.
\]
The correlation is 0.74 and \( r^2 = (0.74)^2 = 0.55 \) (or 55%). We can say that height explains 55% of the variation in hand-span, but what does it mean to say this?

Suppose that we ignore the height information when we make predictions of hand-span. In other words, suppose we don’t use the least squares line to predict hand-span. For the 167 students in the data set, \( \bar{y} \), the average hand-span is about 20.8 centimeters. If we ignore the least squares line, we could use this value to predict the hand-span for any individual, regardless of his or her height. Our prediction “equation” is simply \( \text{hand span} = 20.8 \). For both this equation and the least squares equation involving height, we can compute the sum of squared differences between the actual hand-span values and the predicted values.

- The sum of squared differences between observed \( y \) values and the sample mean \( \bar{y} \) is called the total variation in \( y \) or sum of squares total and is denoted as SSTO.
- The sum of squared differences between observed \( y \) values and the predicted values based on the least squares line is called the sum of squared errors and is denoted by SSE. Remember that errors are sometimes called residuals, and a synonym for sum of squared errors is residual sum of squares.

Whenever the correlation is not 0, the least squares line will produce generally better predictions than the sample mean so SSE will be smaller than SSTO. The squared correlation expresses the reduction in squared prediction error as a fraction of the total variation. This leads to the formula
\[
r^2 = \frac{\text{SSTO} - \text{SSE}}{\text{SSTO}}
\]
It can be shown (using algebra) that this quantity is exactly equal to the square of the correlation.

Let’s consider \( r^2 \) for Examples 5 and 7. In Example 5, the correlation between left and right hand-spans is 0.95 so \( r^2 = 0.90 \), or about 90%. This indicates that the span of one hand is very predictable if we know the span of the other hand (see Figure 5.8). In Example 7, the correlation between television viewing hours and age is only \( r = 0.12 \). The squared correlation is about 0.014. As we can see from the scatter plot in Figure 5.10, knowing a person’s age doesn’t help us predict how much television he or she watches per day.