Excerpt from:

7.1 INTRODUCTION

In Chapter 6 we dealt with linear correlation, a procedure for describing the degree and direction (positive or negative) of the linear relationship between two variables. The focus of this chapter is on using the concepts of linear correlation to develop the concepts of **linear regression**. Linear regression is a procedure for describing, in mathematical terms, the characteristics of the line that best captures the relationship between two variables. There are two important pieces of information that result from this procedure. First, we can use the regression line to predict individuals’ scores on one variable from knowledge of their scores on another variable. Second, we can estimate how accurate our predictions are likely to be.

**LINEAR REGRESSION**: A statistical procedure for predicting one variable from another variable by using a linear prediction rule.

We begin our discussion with an example that illustrates the basic concepts of linear regression. Then we turn to the more technical aspects of regression.

7.2 CONCEPTS OF PREDICTION

Figure 7.1 presents a scatterplot of hypothetical scores on two midterm exams taken by 12 students. Notice that there is a relatively strong, but not perfect, positive correlation between the two sets of scores \( r = .78 \). Now suppose that one student in the class had obtained a score of 20 on the first midterm but, because of illness, was unable to take the second midterm. What prediction could we make about that student’s performance on the second midterm?

For illustrative purposes let us start by ignoring the correlation between the two exams and predicting the student’s score only on the basis of other students’ performance on the second midterm. In this case our best prediction would be the class mean on the second midterm \( \bar{Y} = 15.42 \) because the mean is the value about which the sum of squared deviations between it and all other scores in the distribution is minimized. Moreover, we can estimate the average amount of error in our prediction by computing the class standard deviation \( s_Y = 3.58 \). Recall that the standard deviation is a measure of the amount of error associated with using the mean as a predictor.¹

Let us now predict the student’s score by taking into account the correlation between the two exams. Figure 7.1 displays that correlation by plotting
midterm 2 scores against midterm 1 scores. Referring to Figure 7.1, we locate a score of 20 on the *X* axis. The values directly above are the second midterm scores of those students who obtained 20 on the first midterm, the same score obtained by our ailing student. These values represent the distribution of second midterm scores conditional on having scored 20 on the first midterm. The mean of that conditional distribution (17.33) is the best **predicted score** on midterm 2 for any student who scored 20 on the first midterm.

Clearly the prediction based on the correlation between the two exams (17.33) is different from the prediction that ignored that correlation (15.42). But is it more accurate? The answer is yes because the standard deviation of the conditional distribution (1.53) is smaller than the overall class standard deviation (3.58). That is, the actual midterm 2 scores for students who scored 20 on midterm 1 are closer, on the average, to the mean of their conditional distribution than to the overall class mean. The preceding comparison between using and not using the correlation to predict the midterm 2 score of a student who obtained a score of 20 on midterm 1 is summarized as follows:

<table>
<thead>
<tr>
<th></th>
<th>Predicted score</th>
<th>Prediction error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignoring correlation</td>
<td>15.42</td>
<td>3.58</td>
</tr>
<tr>
<td>Using correlation</td>
<td>17.33</td>
<td>1.53</td>
</tr>
</tbody>
</table>

In sum, when there is a correlation between two variables, we can use that information to make predictions from one variable to another that are
more accurate than predictions that do not take the correlation into account. Next we want to derive a mathematical procedure that is akin to the procedure just described for making predictions from correlated data.

PREDICTED SCORE: The predicted value of one variable for a given value of another variable.

7.3 LINEAR RELATIONSHIPS

Figure 7.2 is a reproduction of Figure 7.1, with the means of each of the conditional distributions of second midterm scores connected to produce a prediction, or regression, line. You can see that this particular regression line is quite jagged and constitutes a very complex curve. If we were to derive a mathematical formula for describing this curve it would be extremely complicated. However, looking at the means of the conditional distributions, it appears that they fall approximately on a straight line. A simpler description, and undoubtedly a more reasonable one, is that the relationship is linear in the whole population of students. However, because each conditional mean is estimated from only a small sample of students, the relationship deviates from a straight line because of sampling fluctuation. Consequently, when a relationship is approximately linear (indicated by a sizable Pearson correlation), we adopt a straight-line model and use it to derive a relatively simple mathematical equation for the relationship.

Figure 7.2 Regression line joining the conditional means of midterm 2 scores for each distinct value of midterm 1.
You should note that because we are adopting a linear model to describe relationships that may not be precisely linear in a particular sample of individuals, predictions based on this model will not necessarily equal conditional means, as they did in the previous illustration. Predictions from a linear model are, in fact, estimates of conditional means based on the assumption that the relationship between two variables is precisely linear in an entire population of individuals.

In the section to follow we will consider the mathematical characteristics of straight lines where all points fall precisely on the line, that is, where the linear relationship is perfect. Although perfect linear relationships are rare in behavioral research, examining them first will help you understand procedures for mathematically describing imperfect linear relationships.

The Formula for a Straight Line

The formula for a straight line expresses a relationship between two variables. Any straight line has the general formula of

\[ Y = a + bX \]  \hspace{0.5cm} (7.1)

This equation states that for any value of variable \( X \), the value of variable \( Y \) that is paired with it on a straight line can be found by multiplying \( X \) by a constant value, \( b \), and adding a second constant value, \( a \), to that product. In this general equation, \( b \) is the slope of the line. Its value indicates how much \( Y \) will change for a one-unit change (increase or decrease) in \( X \). In symbols, given any two points on a line \( (X_1, Y_1) \) and \( (X_2, Y_2) \) the slope of the line is given as

\[ b = \frac{Y_2 - Y_1}{X_2 - X_1} \]

The value of \( a \) indicates the value of \( Y \) when \( X \) is zero. Because the value of \( Y \) when \( X = 0 \) is found on a graph where the line intercepts the \( Y \)-axis, \( a \) is called the \( Y \) intercept.

To illustrate the properties of Formula 7.1, suppose you resolved to get fit in the upcoming year and decided that the best way to do so was to join a local fitness club. However, before joining the club you need to estimate whether you can afford the expense. Assume that the fitness club charges a yearly membership fee of $50 and a session fee of $2 for each time you work out. From this information you can use a linear equation to compute how much keeping fit will cost you per year.

yearly cost = $50 (membership fee) + $2 (session fee) × no. of workouts

\[ Y = $50 + $2X \]

Clearly, because of the membership fee, it will cost you $50 per year even if you never work out at all. Thus, in this example the yearly cost \( [Y] \), when the number of workouts \( [X] \) is zero, is equal to 50. Therefore, \( a = 50 \). For every
additional workout (for each unit increment in \( X \)), your yearly cost \( Y \) will increase by $2. Therefore, \( b = 2 \).

By placing any value for number of workouts into this equation, you can compute the corresponding yearly cost. For example, if you want to know how much working out once a week for a year would cost (52 workouts), your answer would be

\[
Y = \$50 + 2(52)
\]

\[
= \$50 + \$104
\]

\[
= \$154
\]

Figure 7.3 shows a graph of the relationship described by the equation \( Y = 50 + 2(X) \). From it you can see at a glance whether you can afford the expense of using the club some specific number of times.

You should notice several things about Figure 7.3. First, the relationship described is indeed a straight line. Second, the line intercepts the Y-axis at 50, the value of \( a \). Third, there is a positive slope to the line—it increases from the lower left to the upper right—because \( b \) has a positive value (+2). Linear relationships can also have a negative slope, which would be indicated by a negative value of \( b \). For example, in some localities the relationship between car accidents and car insurance is negative. For each additional year of accident-free driving the cost of car insurance goes down. Finally, you should notice that Figure 7.3 illustrates a perfect linear relationship. That is, all points fall exactly on the line. As we mentioned earlier, in behavioral research the relationship between two variables is rarely perfectly linear. Nonetheless, it is possible to determine a linear equation like Formula 7.1 that will allow us to approximate, or predict, values of \( Y \) for different values of \( X \). The procedure for determining the linear equation for an imperfect lin-
ear relationship amounts to determining the values of $a$ and $b$ for the line that best approximates the relationship. The procedure is called linear regression.

### 7.4 LINEAR REGRESSION: THE CRITERION OF LEAST SQUARES

We now know the general formula for a straight line. How can we use it to produce the regression line plotted in Figure 7.4? The first thing to notice is that no single linear equation can be found that reproduces exactly the values of $Y$ from their corresponding values of $X$. The reason, of course, is that the data in Figure 7.4 are not perfectly linear. To reflect this fact, when the general formula for a straight line is used to describe imperfect linear relationships, it is slightly altered to read

$$Y' = a + bX$$  \hspace{1cm} [7.2]$$

Where $Y'$ (pronounced $Y$ prime) refers to the *predicted* value of $Y$.

The basic problem facing us is that any number of lines could be fit to these data, each having different values of $a$ and $b$, and therefore, each producing different predicted values of $Y$. We need a criterion for deciding which of these many lines is in some sense the best line. The criterion chosen is the same criterion used to select the mean as the central location of a univariate distribution, namely, the *criterion of least squares*. Recall that the criterion of least squares specifies the mean as the "best" central location of a univariate distribution because it is the point about which the sum of squared deviations is minimized. That is, the mean is chosen so that $\Sigma(X -$
\( \bar{X} \) is a minimum. Similarly, the criterion of least squares specifies the "best" central location of a bivariate distribution [assuming a linear relationship] as the line about which the sum of squared deviations between it and the actual \( Y \) values is minimized. That is, the regression line is chosen so that \( \Sigma (Y - Y')^2 \) is a minimum.

The regression line drawn in Figure 7.4 was derived by using the least squares criterion. For each value of \( X \), the predicted value of \( Y \) is located on the regression line. The sum of squared deviations between actual \( Y \) values and corresponding \( Y' \) values predicted from this line are smaller than the sum of squared deviations between obtained \( Y \) values and \( Y' \) values predicted from any other line. The obvious question at this point is, how did we arrive at this line? The answer is that we determined the values of \( a \) and \( b \) that satisfied the least squares criterion. How we determined those values is the topic of the next section.

### 7.5 Determining the Regression Coefficients

The values of the \( Y \) intercept, \( a \), and the slope, \( b \), that describe the line that best fits the data in the least squares sense are defined as

\[
b = r_{XY} \frac{s_Y}{s_X}
\]

\[
a = \bar{Y} - b\bar{X} = \bar{Y} - r_{XY} \frac{s_Y}{s_X} \bar{X}
\]

From Formulas 7.2, 7.3, and 7.4 we can derive a formula for computing \( Y' \).

\[
Y' = a + bX
\]

\[
Y' = [\bar{Y} - r_{XY} \frac{s_Y}{s_X} \bar{X}] + [r_{XY} \frac{s_Y}{s_X} X]
\]

which by rearranging terms becomes

\[
Y' = \bar{Y} + r_{XY} \frac{s_Y}{s_X} [X - \bar{X}] = \bar{Y} + b(X - \bar{X})
\]

Focusing for a moment only on the slope of the best fitting line [Formula 7.3], you can see that its value depends partly on \( r \). (See Proof 7.1.) Now looking at Formula 7.5, you can see that when \( r \) is 0.0, the slope is zero and the resulting predicted value of \( Y \) is \( \bar{Y} \) for all values of \( X \).

\[
Y' = \bar{Y} + 0 \frac{s_Y}{s_X} [X - \bar{X}]
\]

\[
Y' = \bar{Y} + 0 = \bar{Y}
\]
For example, if the correlation were zero between the two midterm exams discussed earlier, regardless of which $X$ value is placed in Formula 7.5, the predicted score on midterm 2 would be the mean of $Y = 15.42$.

As the absolute value of $r$ increases so does the absolute value of the slope, and predicted $Y$ values increasingly deviate from $\bar{Y}$ as each value of $X$ deviates from $\bar{X}$. With a perfect correlation, the predicted value of $Y$ deviates from $\bar{Y}$ precisely as much as the corresponding value of $X$ deviates from $\bar{X}$ (relative to their respective standard deviations), and therefore, $Y$ is precisely predictable from $X$. This relationship between correlation and prediction can be seen most clearly by translating Formula 7.5 into standard score terms:

\[
Y' = \bar{Y} + r_{xy} \frac{s_y}{s_x} (X - \bar{X})
\]

\[
Y' - \bar{Y} = r_{xy} \frac{s_y}{s_x} (X - \bar{X})
\]

\[
\frac{Y' - \bar{Y}}{s_y} = r_{xy} \frac{X - \bar{X}}{s_x}
\]

\[
z_{Y'} = r_{xy} z_X
\]  \hspace{1cm} (7.6)

Formula 7.6 shows that when raw scores are transformed into standard scores, $r = b$ and the predicted values of $Y$ in standard score units ($z_{Y'}$) is equal to $r$ times the standard score value of $X(z_X)$. The relation between correlation and regression is an important one and will be touched on throughout this chapter. Now, however, it is time to use the information learned to this point to show how we arrived at the regression line in Figure 7.4.

Table 7.1 reproduces the data plotted in Figure 7.4. We will use it to compute the regression coefficients (Formulas 7.3 and 7.4), produce regression formulas for predicting $Y$ (Formulas 7.2 and 7.5), and show how to plot the regression line.

**REGRESSION COEFFICIENTS:**  The values of $a$ and $b$ derived by using the least squares criterion to determine a best fitting line.

Included in Table 7.1 are the values of $r_{xy}$, $s_x$, $s_y$, $\bar{X}$, and $\bar{Y}$, all the information we need to compute the regression coefficients. Thus,

\[
b = r_{xy} \frac{s_y}{s_x} = .78 \frac{3.57}{5.73} = .485
\]
Table 7.1  COMPUTING REGRESSION STATISTICS
ON TWO MIDTERM EXAMS
(Midterm 1, $X$, is used to predict midterm 2, $Y$.)

<table>
<thead>
<tr>
<th>Student</th>
<th>Midterm 1 ($X$)</th>
<th>Midterm 2 ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>23</td>
</tr>
</tbody>
</table>

Mean 15.8  15.4
Standard deviation 5.7  3.6
$r = .78$

and

$$a = \overline{Y} - b\overline{X} = 15.42 - .485(15.83) = 7.75$$

Putting these values for $a$ and $b$ into Formula 7.2, we find that

$$Y' = a + bX = 7.75 + .485(X)$$

If we use Formula 7.5 instead,

$$Y' = \overline{Y} + b(X - \overline{X}) = 15.42 + .485(X - 15.83)$$

Either formula will produce the same predicted $Y$ scores, and therefore, the same regression line. The advantage of Formula 7.2 is that it literally describes the mathematical characteristics (the values of $a$ and $b$) of the regression line for predicting $Y$. Thus, it is usually used for descriptive purposes. However, to actually compute $Y'$ values it is easier to use Formula 7.5 because computation of $a$ is unnecessary.

To construct our regression line for predicting $Y$ from $X$ all we need to do is take two values of $X$ (since a straight line is completely determined by two points), enter them into either Formula 7.2 or Formula 7.5, and compute two values of $Y'$. We then place these two paired scores, $(X_1, Y'_1)$ and $(X_2, Y'_2)$, on a scatterplot and join the points with a line.

Here is an example of computing $Y'$ values by using both Formula 7.2 and Formula 7.5. The chosen values of $X$ (midterm 1 scores) are 11 and 20.
\[ X \text{ value} | \text{Formula 7.2 for } Y' | \text{Formula 7.5 for } Y' \\
\hline
11 | Y' = 7.74 + .485(11) | Y' = 15.42 + .485(11 - 15.83) \\
| = 13.10 | = 13.10 \\
20 | Y' = 7.74 + .485(20) | Y' = 15.42 + .485(20 - 15.83) \\
| = 17.44 | = 17.44 \\
\]

You can see that both formulas produce the same values of \( Y' \).

Referring to Figure 7.4, locate the values of 11 and 20 on the X-axis. Moving up perpendicularly from these scores to the regression line, you will find that the values of 13.1 and 17.44 fall on the line. Any other value of \( Y' \) can be found by looking at the regression line or by using either Formula 7.2 or Formula 7.5. What are the \( Y' \) values associated with \( X \) values of 8 and 24?

### 7.6 PREDICTION ERROR

Although the regression line provides the best prediction of \( Y \) values from known \( X \) values, rarely do predicted values (\( Y' \)s) actually equal obtained values (\( Y \)s). In fact, only when the correlation between \( X \) and \( Y \) is perfect will \( Y \) and corresponding \( Y' \) values equal one another. In all other cases there will be some deviation between \( Y \) and \( Y' \), in which case we say there is some amount of prediction error associated with the regression line. By estimating the magnitude of this prediction error we can then evaluate how well the regression line fits the data. This evaluation takes the form of comparing the prediction error associated with the regression line to the error associated with using the sample mean as a predictor. The smaller the prediction error of the regression line relative to the prediction error of the mean, the better the regression line fits the data. Let us now develop this idea a little further.

---

**PREDICTION ERROR**: The average difference between predicted scores and obtained scores.

**Variance Error and Standard Error of Estimate**

Consider developing a predictor of students' midterm 2 scores without benefit of knowing their midterm 1 scores. In this case we could not use the regression line based on the correlation between midterms. Instead, we would have to use as our predictor the class mean on midterm 2 because it is our best least squares predictor when no other information is available. The prediction errors associated with using \( \bar{Y} \) to predict \( Y \) are the \( Y - \bar{Y} \) values for each student. The prediction error for student 10 is indicated in Figure 7.5 by...
Figure 7.5 Scatterplot of midterm exam scores showing (a) the line representing the overall mean of midterm 2 across distinct values of midterm 1, (b) the linear regression of midterm 2 on midterm 1, (c) the total deviation scores \( (Y - \bar{Y}) \), (d) the unexplained deviation scores, \( (Y'' - \bar{Y}) \), and (e) explained deviation scores \( (Y' - \bar{Y}) \).

The vertical distance between the student’s observed \( Y \) value and the line representing the value of \( \bar{Y} \) across all values of \( X \). The average prediction error associated with the mean is the sample variance,

\[
 s_Y^2 = \frac{SS_Y}{df} = \frac{\Sigma(Y - \bar{Y})^2}{N - 1}
\]

or the sample standard deviation,

\[
 s_Y = \sqrt{\frac{SS_Y}{df}} = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{N - 1}}
\]

Now let us use the regression line for our predictions. The prediction errors associated with using \( X \) to predict \( Y \) are the \( Y'' - Y' \) values for each student, an example of which is shown for student 10 in Figure 7.5. The average prediction error associated with the regression line is conceptually equivalent to the sample variance and is computed as

\[
 s_{Y''-Y'}^2 = \frac{SS_{Y''-Y'}}{df} = \frac{\Sigma(Y'' - Y')^2}{N - 2}
\]

This statistic is called the variance error of estimate or residual variance. Its square root is called the standard error of estimate:

\[
 s_{Y''-Y'} = \sqrt{\frac{SS_{Y''-Y'}}{df}} = \sqrt{\frac{\Sigma(Y'' - Y')^2}{N - 2}}
\]
STANDARD ERROR OF ESTIMATE:  The standard deviation of the prediction errors in linear regression.

The degrees of freedom for the variance error of estimate and the standard error of estimate are \( N - 2 \). Recall from our discussion of variance in Chapter 4 that the degrees of freedom for a particular statistic are equal to the number of observations \( (N) \) minus the number of parameters that have to be estimated to compute that statistic. In the case of the sample variance, 1 degree of freedom is lost because one parameter, the population mean, has to be estimated from the sample data to compute the sum of squared errors from the mean. Therefore, the degrees of freedom for the sample variance are \( N - 1 \), where \( N \) is the total number of observations and 1 is the number of parameters estimated. In the case of the variance error of estimate, two parameters, \( a \) and \( b \), must be estimated from the sample data to compute the sum of squared errors from the regression line. For this reason the degrees of freedom for the variance estimate of error are \( N - 2 \). The variance error of estimate and the standard error of estimate computed from Table 7.2 are

\[
s^2_{Y-Y'} = \frac{54.46}{10} = 5.45
\]

\[
s_{Y-Y'} = \sqrt{5.45} = 2.33
\]

These values are smaller than the obtained sample variance and standard deviation, indicating that the regression line is a better predictor than the class mean.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Variance error</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{Y} )</td>
<td>( s_Y = 12.81 )</td>
<td>( s_Y = 3.58 )</td>
</tr>
<tr>
<td>Regression line</td>
<td>( s^2_{Y-Y'} = 5.45 )</td>
<td>( s_{Y-Y'} = 2.33 )</td>
</tr>
</tbody>
</table>

Prediction Error and Correlation

Estimating prediction error by using Formulas 7.7 and 7.8 involves calculating \( Y' \), then subtracting it from \( Y \) and squaring the result for every observation, a tedious task that only the most vigilant of us can accomplish without computation error. Fortunately there is an easier way. With some algebraic manipulation, the numerator of Formula 7.7 can be rewritten as

\[
SS_{Y-Y'} = SS_Y(1 - r^2)
\]  

which, when expressed in variance terms instead of sums of squares, becomes

\[
s^2_{Y-Y'} = s^2_Y(1 - r^2) \frac{N - 1}{N - 2}
\]
Table 7.2  EXAMPLE DATA FOR COMPUTING REGRESSION STATISTICS
ON MIDTERM 1 (X) AND MIDTERM 2 (Y) SCORES

<table>
<thead>
<tr>
<th>Student</th>
<th>Midterm 1 (X)</th>
<th>Midterm 2 (Y)</th>
<th>Estimate (Y')</th>
<th>Error (Y - Y')</th>
<th>Error² (Y - Y')²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>11</td>
<td>11.59</td>
<td>-0.59</td>
<td>0.34</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>14</td>
<td>11.59</td>
<td>2.41</td>
<td>5.81</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>10</td>
<td>13.05</td>
<td>-3.05</td>
<td>9.25</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>16</td>
<td>13.05</td>
<td>2.95</td>
<td>8.69</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>12</td>
<td>14.52</td>
<td>-2.52</td>
<td>6.35</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>14</td>
<td>14.52</td>
<td>-0.52</td>
<td>0.27</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>16</td>
<td>15.60</td>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>16</td>
<td>17.45</td>
<td>-1.45</td>
<td>2.10</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>17</td>
<td>17.45</td>
<td>-0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>19</td>
<td>17.45</td>
<td>1.55</td>
<td>2.41</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
<td>17</td>
<td>19.41</td>
<td>-2.42</td>
<td>5.81</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>23</td>
<td>19.41</td>
<td>3.59</td>
<td>12.89</td>
</tr>
</tbody>
</table>

\[ s_y^2 = 12.82 \quad SS_Y = 140.92 \]
\[ s_y = 3.58 \quad SS_Y = 86.46 \]
\[ s_{Y-Y'}^2 = 5.46 \quad SS_{Y-Y'} = 54.46 \]
\[ s_{Y-Y'} = 2.33 \]
\[ r_{XY} = .78 \]

When \( N \) is large, the ratio of \( N - 1 \) to \( N - 2 \) approaches 1, and the formula for the variance error of estimate is simplified to

\[ s_{Y-Y'}^2 = s_y^2(1 - r^2) \quad [7.11] \]

Taking the square root of the variance error of estimate produces the standard error of estimate,

\[ s_{Y-Y'} = s_y \sqrt{(1 - r^2) \frac{N - 1}{N - 2}} \quad [7.12] \]

which with large samples can be simplified to

\[ s_{Y-Y'} = s_y \sqrt{(1 - r^2)} \quad [7.13] \]

The variance error of estimate from Formulas 7.10 and 7.11 is

\[ s_{Y-Y'}^2 = 12.81(1 - .78^2) \frac{11}{10} = 5.45 \]
\[ s_{Y-Y'}^2 = 12.81(1 - .78^2) = 4.96 \]

The standard error of estimate is simply the square root of each of these quantities, that is, 2.33 and 2.23, respectively. With the term \( N - 1/N - 2 \)
included, the correlation formulas agree exactly with the raw score formulas. Without it, the variance error of estimate and standard error of estimate are underestimated, a result you should keep in mind because Formulas 7.11 and 7.13 are frequently used in practice, sometimes erroneously when sample sizes are small.

From these formulas you can easily see the range of possible values for the prediction error associated with the regression line. When the correlation is zero, the prediction error associated with the regression line is essentially equal to the prediction error associated with the mean. That is,

\[ s_{Y-Y'}^2 = s_Y^2 (1 - 0) \frac{N - 1}{N - 2} = s_Y^2 \]

This result should make sense if you remember that when \( r = 0 \), \( Y' = \bar{Y} \) for all values of \( X \). Thus, the prediction errors, \( Y - Y' \), will equal \( Y - \bar{Y} \), and the maximum value of \( s_{Y-Y'}^2 \) must be \( s_Y^2 \).

In contrast, when the correlation is 1.0 or \(-1.0\), all of the data fall precisely on the regression line and there is no prediction error:

\[ s_{Y-Y'}^2 = s_Y^2 (1 - 1) \frac{N - 1}{N - 2} = 0 \]

Again, this should make sense because if \( r = 1.0 \), \( Y' = Y \) for all values of \( X \), and therefore, the prediction errors, \( Y - Y' \), will attain their minimum value of zero.

It should now be apparent that regression and correlation are closely related concepts and that both are related to the concept of variation. We now examine certain aspects of these relationships in order to shed additional light on the interpretation of correlation and regression statistics.

Regression and Variation

Consider once again Figure 7.5. We have already seen that the deviations, \( Y - \bar{Y} \), represent error associated with using the sample mean as a predictor of \( Y \). The value of \( Y - \bar{Y} \) is called a total deviation score. For example, the total deviation score on \( Y \) for student 10 in Table 7.2 is

\[ Y - \bar{Y} = 19.00 - 15.42 = 3.58 \]

If we square each total deviation score and sum, we obtain \( SS_Y \), which in regression analysis we call the total variation in \( Y \). It is the amount of variation in \( Y \) that exists without considering the relationship between \( Y \) and \( X \). In Table 7.2, \( SS_Y = 140.92 \).

We have also seen that the deviations \( [Y - Y'] \) represent error associated with using the regression line as a predictor of \( Y \). The value of \( Y - Y' \) is called the unpredictable deviation score. Student 10's unpredictable deviation score on \( Y \) is

\[ Y - Y' = 19.00 - 17.45 = 1.55 \]