11.1 Creating the Least Squares Equation

Probabilistic Models

In this section, we will try to model the relationship between two variables. In algebra, you worked with many models that were determinstic in nature. For example, the model: \( y = 1.06x \) is a deterministic model that will give the after tax price for an item purchased in Florida. \( X \) here represents the pre-tax price of an item. \( Y \) is the final price of the item post-tax. This model is deterministic because given a pre-tax price we can exactly (there is no error in this prediction) determine the value of the item after tax. Recall that the \( y \) variable is called the dependent variable because it depends upon the independent variable \( x \).

Deterministic models are great when we can get them, but many times we do not know all of the factors affecting the dependent variable. In those cases, we will not be able to predict \( y \) without error. This means we will need to create a probabilistic model:

\[
y = \text{deterministic model} + \text{Random error}
\]

In the probabilistic model graph, the difference between the height of the line and the height of our individual points is due to the random error.
In this chapter, we look at the simplest form of a probabilistic model:

**A First-Order (Straight-Line) Probabilistic Model**

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

where: \( y \) = Dependent variable & \( x \) = independent variable

\[ E(y) = \beta_0 + \beta_1 x \] is Deterministic component,

\( \beta_1 \) = slope, \( \beta_0 \) = y-intercept, and \( \varepsilon \) = random error component.

Using data to come up with estimates of the parameters \( \beta_0 \) and \( \beta_1 \) in order to form an equation is called **regression analysis**. The goal of **regression analysis** is to find the straight line that comes closest to all of the points in a scatter plot simultaneously.
Fitting the Model: The Least Squares Approach

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

To fit the straight line model we need to find a way to estimate the unknown parameters: \( \beta_1 \) \& \( \beta_0 \).

Consider this simple example:

In earlier research it was found that the body-mass index (BMI, weight scaled for height) was the main determinant of female physical attractiveness as judged by men. This raised the question: can women's attraction to men be so easily explained? To explore this hypothesis, researchers replicated the experiments on female attractiveness, but substituted male bodies and female raters.

The results show that a woman's ratings of male attractiveness can be explained by simple physical characteristics, in particular the waist-to-chest ratio. Women prefer men who have a narrow waist and a broad chest and shoulders. Below are some of the waist-to-chest ratios and the attractiveness scores issued by the female judges:

<table>
<thead>
<tr>
<th>Subject</th>
<th>WCR</th>
<th>Attractiveness Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.70</td>
<td>6.5</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>5.3</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>0.85</td>
<td>3.8</td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
<td>3.5</td>
</tr>
</tbody>
</table>

If we assume an adequate model for this situation is the first-order linear model \( E(y) = \beta_0 + \beta_1 x \), we can try to use the sample data to estimate the missing parameters \( \beta_1 \) \& \( \beta_0 \).

Consider the scatter plot of the data below:
We could try to fit some arbitrary line to the points above. For Example:
It looks like this line passes through two points: (0.70, 6.5) and another point of (0.85, 3.8) which gives a slope of \( \frac{y_2 - y_1}{x_2 - x_1} = \frac{6.5 - 3.8}{0.70 - 0.85} \approx -18 \). Using the point-slope formula from Algebra, we get the linear equation: \( y = -18x + 19.1 \). This model was obtained visually (i.e.-we guessed). We could have made several other guesses at the appropriate equation, so we should assess our guess.

Let’s then compare the observed and predicted values for the visual model we found. In the table below, the X and Y are the actual values represented by the dots in our graph above. The \( \tilde{y} \) (y-tilda) is the value that results when we plug the x value from the left most column of the table into our model. The first of the last two columns gives us the difference between the actual y-value from the point and the predicted value from our line. We call that difference the error of our prediction. For example, our line model says that when x is 0.75 we should have y at 5.6, but in reality the y value at x = 0.75 was 5.3. This means our error is -0.3. The last column squares these differences (or errors).

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>( \tilde{y} = -18x + 19.1 )</th>
<th>( y - \tilde{y} )</th>
<th>( (y - \tilde{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>6.5</td>
<td>6.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.75</td>
<td>5.3</td>
<td>5.6</td>
<td>-0.3</td>
<td>0.09</td>
</tr>
<tr>
<td>0.80</td>
<td>4.0</td>
<td>4.7</td>
<td>-0.7</td>
<td>0.49</td>
</tr>
<tr>
<td>0.85</td>
<td>3.8</td>
<td>3.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.90</td>
<td>3.5</td>
<td>2.9</td>
<td>0.6</td>
<td>0.36</td>
</tr>
<tr>
<td><strong>Sum of Errors:</strong></td>
<td></td>
<td></td>
<td><strong>-0.4</strong></td>
<td><strong>0.94</strong></td>
</tr>
</tbody>
</table>

One way to determine quantitatively how well a straight line fits a set of points is to note the extent to which the data points deviate from the line. The quantity \( \sum (y - \tilde{y}) \) in the table above gives us the total deviation between our observed values and our predicted values. The \( \sum (y - \tilde{y}) \) should equal zero (the sum of errors should equal zero). In our visual model, that is not the case, which is something that would eliminate it as a candidate model for this set of data points. We will want all of our prediction lines to have the property that the sum of errors equals zero. This will ensure on average our error of prediction is zero. The quantity \( \sum (y - \tilde{y})^2 \) is called the sum of squares of the errors (SSE) gives another measure of deviation which gives a greater emphasis to larger deviations from the line.
Remember we visually selected the model (line) above, so it is no wonder that the sum of errors was not zero. However, there are usually multiple models possible that have the property that \( \sum (y - \hat{y}) = 0 \). Since, we can find more than one model with the property \( \sum (y - \hat{y}) = 0 \), so we need an additional criteria to choose the best fitting line. It turns out that it can be shown that there is one (and only one) line for which the SSE is a minimum. This line is called the \textbf{least squares line}.

The \textbf{least squares line} has the following properties:

1. The sum of errors (SE) equals zero.
2. The sum of squared errors (SSE) is smaller than that for any other straight line model.

The following formulas will give the Least Squares Estimates for the population \( \beta_1 \) & \( \beta_0 \). (We will use a “hat” symbol to denote the estimates, that is to say \( \hat{\beta}_1 \) estimates \( \beta_1 \) and \( \hat{\beta}_0 \) estimates \( \beta_0 \)).

\textbf{Formulas for the Least Squares Estimates}

\textit{Slope:} \( \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \)

\textit{y-intercept:} \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \)

where \( SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \left( \frac{\sum x_i}{n} \right) \left( \sum y_i \right) \)

and \( SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \left( \frac{\sum x_i}{n} \right)^2 \)

\textbf{Example 160:} Now let us find the least squares prediction line for our WCR vs. Attractiveness Rating example above:

Let \( E(y) = \beta_0 + \beta_1 x \) be our straight line model where \( y = \text{attractiveness rating} \) on an 8-point scale and \( x = \text{waist-to-chest ratio} \).
Preliminary computations for the male attractiveness problem

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i^2$</th>
<th>$x_iy_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>6.5</td>
<td>0.49</td>
<td>4.55</td>
</tr>
<tr>
<td>0.75</td>
<td>5.3</td>
<td>0.5625</td>
<td>3.975</td>
</tr>
<tr>
<td>0.80</td>
<td>4.0</td>
<td>0.64</td>
<td>3.2</td>
</tr>
<tr>
<td>0.85</td>
<td>3.8</td>
<td>0.7225</td>
<td>3.23</td>
</tr>
<tr>
<td>0.90</td>
<td>3.5</td>
<td>0.81</td>
<td>3.15</td>
</tr>
</tbody>
</table>

Totals

$\sum x_i = 4$  \hspace{1cm} $\sum y_i = 23.1$  \hspace{1cm} $\sum x_i^2 = 3.225$  \hspace{1cm} $\sum x_iy_i = 18.105$

Using the numbers from above we can get:

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_iy_i - \left(\frac{\sum x_i}{n}\right)\left(\frac{\sum y_i}{n}\right) = 18.105 - (4)(23.1)/5 = -0.375$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2 = 3.225 - \frac{(4)^2}{5} = 3.225 - 3.2 = 0.025$$

then \( \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = -0.375/0.025 = -15 \)

and \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \left(\frac{\sum y_i}{n}\right) - \left(\hat{\beta}_1\frac{\sum x_i}{n}\right) = \frac{23.1}{5} + 15\left(\frac{4}{5}\right) = 16.62 \)

The least squares line is then given by:

\[ \hat{y} = 16.62 - 15x \]
Let us find the SSE for this line to determine if it beats our visual model:

\[
\hat{y} = 16.62 - 15x
\]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>(\hat{y} = 16.62 - 15x)</th>
<th>((y - \hat{y}))</th>
<th>((y - \hat{y})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>6.5</td>
<td>6.12</td>
<td>0.38</td>
<td>0.1444</td>
</tr>
<tr>
<td>0.75</td>
<td>5.3</td>
<td>5.37</td>
<td>-0.07</td>
<td>0.0049</td>
</tr>
<tr>
<td>0.80</td>
<td>4.0</td>
<td>4.62</td>
<td>-0.62</td>
<td>0.3844</td>
</tr>
<tr>
<td>0.85</td>
<td>3.8</td>
<td>3.87</td>
<td>-0.07</td>
<td>0.0049</td>
</tr>
<tr>
<td>0.90</td>
<td>3.5</td>
<td>3.12</td>
<td>0.38</td>
<td>0.1444</td>
</tr>
</tbody>
</table>

**Sum of Errors:** 0 0.683

We can now confirm our Least Squares Model is better fitting than our visual model because our LSM has a Sum of Errors = 0, and even if both models had that trait, the least squares model has a lower SSE.

Finally, we can now use the above model for prediction.

**Example 161** What would the average attractiveness score be for a 45 inch chest and a 32 inch waist (which produces a WCR of 0.71)?
Example 162: Find the least squares prediction line for the following pre-owned Corvette data.

<table>
<thead>
<tr>
<th>Age</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>27000</td>
</tr>
<tr>
<td>6</td>
<td>26000</td>
</tr>
<tr>
<td>5</td>
<td>27500</td>
</tr>
<tr>
<td>2</td>
<td>40500</td>
</tr>
<tr>
<td>2</td>
<td>36400</td>
</tr>
<tr>
<td>5</td>
<td>29500</td>
</tr>
<tr>
<td>4</td>
<td>33500</td>
</tr>
<tr>
<td>5</td>
<td>30800</td>
</tr>
<tr>
<td>1</td>
<td>40500</td>
</tr>
<tr>
<td>4</td>
<td>30500</td>
</tr>
</tbody>
</table>

Example 163: What would the average price be for a Corvette that was three years old?

11.2 Finding S for the Random Error Terms

Model Assumptions

Our model in the above section had a deterministic component and a random error component. In this section, we will consider the assumptions that we make about that random error component ($\epsilon$).

Assumption 1: The mean of the probability distribution for ($\epsilon$) is 0. That is why $E(y) = \beta_0 + \beta_1 x$.

Recall the original model was $y = \beta_0 + \beta_1 x + \epsilon$.

Assumption 2: The variance for ($\epsilon$) is a constant denoted by $\sigma^2$. No matter what x value we use in the model the distribution of the random error has the same variance.

Assumption 3: The probability distribution of ($\epsilon$) is normal.

Assumption 4: The values of ($\epsilon$) associated with any two observed values of y are independent.

Study the figure below: notice how the shape of the error distribution is normal and the variance is also constant.
An estimator of $\sigma^2$

$$S^2 = \frac{SSE}{\text{Degrees of Freedom for Error}} = \frac{SSE}{n-2}$$

*(We used up two degrees of freedom estimating our two parameters $\beta_1$ & $\beta_0$. That is why d.f. = $n - 2$.)

where:

$$SSE = \sum (y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

where:

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\frac{\sum y_i}{n}\right)^2$$

and recall $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$

Finally, if we want an estimate of $\sigma$, we will use $\sqrt{S^2} = \sqrt{\frac{SSE}{n-2}}$

Let's find $S^2 = \frac{SSE}{n-2}$ for the following example:
Example 164: Find $S^2 = \frac{SSE}{n-2}$ for the data below:

<table>
<thead>
<tr>
<th>Number of Weeks on a Low Carb Diet</th>
<th>Weight Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$Y$</td>
</tr>
<tr>
<td>1</td>
<td>-5</td>
</tr>
<tr>
<td>2</td>
<td>-9</td>
</tr>
<tr>
<td>3</td>
<td>-12</td>
</tr>
<tr>
<td>4</td>
<td>-17</td>
</tr>
<tr>
<td>5</td>
<td>-20</td>
</tr>
</tbody>
</table>

Steps to finding $S^2$, an estimate of $\sigma^2$:

1. Create and fill in the preliminary calculation table below:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i^2$</th>
<th>$y_i^2$</th>
<th>$x_iy_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>$\sum x_i$</td>
<td>$\sum y_i$</td>
<td>$\sum x_i^2$</td>
<td>$\sum y_i^2$</td>
</tr>
</tbody>
</table>

2. Calculate $SS_{xy} = \sum x_iy_i - \frac{(\sum x_i)(\sum y_i)}{n}$

3. Calculate $SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

4. $SS_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$

5. Find $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$

6. Calculate $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$

7. Finally, $S^2 = \frac{SSE}{n-2}$
Interpretation of $s$, the estimated standard deviation of $(\epsilon)$:

We expect approximately 95% of the observed $y$ values to lie within $2s$ of their respective least squares predicted $y$ – values, $\hat{y}$.

For example, if $s = 0.103$ we would expect 95% of all of our observed values to lie within 0.206 of the least squares line.

Incidentally, the difference ($y - \hat{y}$) between an observed sample $y$-value and the value of $\hat{y}$, which is the value of $y$ that is predicted by using the regression equation, is called a residual for a sample of paired $(x, y)$ data.

Example 165: Find the $S^2 = \frac{SSE}{n-2}$ for the following data and state the largest deviation we would expect between any of the actual data points and our least squares line.

<table>
<thead>
<tr>
<th>Female Height (in Meters)</th>
<th>Ideal Height of Mate</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>1.52</td>
<td>1.69</td>
</tr>
<tr>
<td>1.60</td>
<td>1.74</td>
</tr>
<tr>
<td>1.68</td>
<td>1.80</td>
</tr>
<tr>
<td>1.75</td>
<td>1.93</td>
</tr>
<tr>
<td>1.83</td>
<td>2.00</td>
</tr>
</tbody>
</table>

11.3 Finding the Standard Error of the Slope Estimator

In order to perform a hypothesis test or form a confidence interval to make an inference about $\beta_1$ (the slope), we need to know the sampling distribution of our estimator $\hat{\beta}_1$.

**Sampling Distribution of $\hat{\beta}_1$**
If we make the four assumptions about $\varepsilon$ (see section 11.2), the sampling distribution of the least squares estimator $\hat{\beta}_1$ of the slope will be normal with mean $\beta_1$ (the true slope) and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{ss}}}$$

We estimate $\sigma_{\hat{\beta}_1}$ by $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{ss}}}$ and refer to this quantity as the estimated standard error of the least squares slope $\hat{\beta}_1$ (recall $s = \sqrt{\frac{SSE}{n-2}}$).

**Example 165.5** Use the data and the results from example 165 to find the standard error of the slope estimator ($s_{\hat{\beta}_1}$).

Now that we know the sampling distribution of $\hat{\beta}_1$, we can perform our hypothesis test.

### 11.4 Hypothesis Tests about the Slope $\beta_1$

**Making Inferences about $\beta_1$, our slope**

Recall $\beta_1$ is our slope for the linear model: $y = \beta_0 + \beta_1 x + \varepsilon$.

![Positive Relationship](image1.png)

![No Relationship](image2.png)

![Negative Relationship](image3.png)
If the true value of the slope is equal to zero \( y = \beta_0 + \beta_1 x + \epsilon \) becomes \( y = \beta_0 + 0 x + \epsilon = \beta_0 + \epsilon \), this means that \( x \) has no role in predicting \( y \). If that is the case, our model is not useful. For this reason, we will want to test the claim that the slope is equal to zero. We would like to reject that claim because if we are unable to reject it we have a useless model.

### A Test of Model Utility: Simple Linear Regression

**One Tailed Test**

\[
H_a : \beta_1 \geq 0 \quad \text{or} \quad H_a : \beta_1 \leq 0
\]

\[
H_a : \beta_1 < 0 \quad \text{or} \quad H_a : \beta_1 > 0
\]

Test Statistic: 
\[
t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S / \sqrt{SSxx}}
\]

Rejection region: \( t < -t_a \)

Or \( (t > t_a \) when \( H_a : \beta_1 > 0 ) \)

**Two Tailed Test**

\[
H_a : \beta_1 = 0
\]

\[
H_a : \beta_1 \neq 0
\]

Rejection region: \(|t| > t_{a/2}\)

Where \( t_a \) and \( t_{a/2} \) are based on \((n-2)\) degrees of freedom

Now let’s do an example:

**Example 166:** At the 1% significance level, test the claim that there is a positive linear relationship between a mother’s height and her daughter’s height.

<table>
<thead>
<tr>
<th>Mother’s Height</th>
<th>Daughter’s Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>58.6</td>
</tr>
<tr>
<td>67</td>
<td>64.7</td>
</tr>
<tr>
<td>64</td>
<td>65.3</td>
</tr>
<tr>
<td>60</td>
<td>61.0</td>
</tr>
<tr>
<td>65</td>
<td>65.4</td>
</tr>
<tr>
<td>59</td>
<td>60.9</td>
</tr>
</tbody>
</table>
**Step 1:** Form the claim symbolically

\[ \beta_1 > 0 \]

**Step 2:** Get your Hypotheses

\[ H_0 : \beta_1 \leq 0 \]
\[ H_A : \beta_1 > 0 \]

**Step 3:** Find \( s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}} \) where \( s = \sqrt{\frac{SSE}{n-2}} \), \( \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \), and \( SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} \)

\[ s_{\hat{\beta}_1} = 0.3579 \]

**Step 4:** Find the Test Stat

\[ t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1} = \sqrt{s} / \sqrt{SS_{xx}}} = 1.719 \]

**Step 5:** Find your critical t-value by looking up \( \alpha = 1\% \) in one tail 4 df \( (n-2) / 2 = 3.747 \)

**Step 6:** Form your initial conclusion: Do not reject the null.

**Step 7:** State your final conclusion

At the 1% significance level, there is not sufficient evidence to support the claim that there is a positive relationship between a mother's height and her daughter's height.

**Example 167:** At the 1% significance level, test the claim that there is a linear relationship between the number of casino employees (in thousands) working in Mississippi and the crime rate per thousand.

<table>
<thead>
<tr>
<th>Number of casino employees</th>
<th>Crime rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.35</td>
</tr>
<tr>
<td>18</td>
<td>1.63</td>
</tr>
<tr>
<td>24</td>
<td>2.33</td>
</tr>
<tr>
<td>22</td>
<td>2.41</td>
</tr>
<tr>
<td>25</td>
<td>2.63</td>
</tr>
<tr>
<td>29</td>
<td>2.93</td>
</tr>
</tbody>
</table>

![Pie chart showing total crime in Mississippi (1999-2008)]
11.5 Confidence Interval for the Slope $\beta_1$

It is also possible to form an interval estimator of $\beta_1$:

**A 100(1-\alpha)% Confidence Interval for the Sample Linear Regression Slope $\beta_1$**

$$\hat{\beta}_1 \pm t_{n/2}s_{\hat{\beta}_1}$$

Where the estimated standard error of $\hat{\beta}_1$ is calculated by $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

And $t_{n/2}$ is based on (n-2) degrees of freedom.

**Example 168**: Using the data from above, we calculated $s_{\hat{\beta}_1} = 0.01257$ and $\hat{\beta}_1 = 0.11729$. Use the formula, $\hat{\beta}_1 \pm t_{n/2}s_{\hat{\beta}_1}$ to form a 99% confidence interval to estimate the true slope $\beta_1$ for the Mississippi Casino/Crime rate data above.

11.6 Finding $r$, the Coefficient of Correlation

**The Coefficient of Correlation**

The coefficient of correlation, $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ is a measure of the strength of the linear relationship between two variables $x$ and $y$.

Recall that the numerator for $\hat{\beta}_1$ is $SS_{xy}$, this is the same as the numerator for $r$. This means that when $SS_{xy} = 0$, both $r$ and $\hat{\beta}_1$ will be equal to zero. When $SS_{xy} = 0$, there is no linear relationship between $x$ and $y$. 
Values of $r$ and their implications:

People sometimes misinterpret $r$. Please remember that if $r = 0$ it does not mean there is no relationship between $x$ and $y$ it just means there does not seem to be a linear relationship between them. Look at drawing f. above, it clearly has a relationship (perhaps quadratic), but it has no linear relationship. Also, if $|r|$ is close to one, it does not mean that $x$ causes $y$ or that $y$ causes $x$. It only means there is some linear relationship between the two variables, but the relationship could be due to some other unknown cause. For example, in the 1970’s it might have been possible to show a positive correlation between number of hours spent flying and the incidence of lung cancer. This doesn’t mean that flying causes lung cancer. In fact, in the 1970’s people were allowed to smoke on planes. This meant frequent flyers were inundated with second hand smoke, which was more likely the cause of the higher rates of cancer.
Example 169: Calculate the correlation coefficient for the following data set which gives the waist measurement and body fat percentage for males who weigh 165 pounds. Do these quantities have a linear relationship? If so, is it positive or negative?

<table>
<thead>
<tr>
<th>Male Waist (in inches)</th>
<th>Body fat %</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>8.5</td>
</tr>
<tr>
<td>32</td>
<td>12.0</td>
</tr>
<tr>
<td>34</td>
<td>18.5</td>
</tr>
<tr>
<td>36</td>
<td>25.0</td>
</tr>
<tr>
<td>38</td>
<td>27.0</td>
</tr>
</tbody>
</table>

Here are some more examples of $r$ values and scatter plots:

(a) Positive correlation: $r = 0.851$

(b) Positive correlation: $r = 0.991$

(c) Perfect positive correlation: $r = 1$

(d) Negative correlation: $r = -0.702$

(e) Negative correlation: $r = -0.965$

(f) Perfect negative correlation: $r = -1$
How do we know if $r$ is close enough to -1 or 1 to conclude that linear correlation exists? What about the natural variability that will occur from sample to sample? Well, normally in this case we would conduct a hypothesis test. This is actually quite simple. I have included a flow chart that describes the process. The only thing you need to know is that the population symbol for the correlation coefficient is “rho” ($\rho$).
Once we know if there seems to be enough evidence to assume there is linear correlation we can start to use our model:

### 11.7 Finding r-squared the Coefficient of Determination

**The coefficient of determination**

Another way to measure the usefulness of the model is to measure the contribution of $x$ in predicting $y$. To do this, we calculate how much the errors of prediction of $y$ were reduced by using the information provided by $x$.

Recall that $SS_{yy} = \sum (y_i - \bar{y})^2 = \text{total sample variation}$ of the observations around the sample mean for $y$, and $SSE = \sum (y_i - \hat{y}_i)^2 = \text{the remaining unexplained sample variability}$ after fitting the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. 

Given any value of one variable, the best predicted value of the other variable is its sample mean.
Then recall that if the model is \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \), and \( x \) contributes nothing to the prediction of \( y \), the best model becomes: \( \hat{y} = \bar{y} \). (If \( x \) does not contribute to the prediction of \( y \) then why is it that the best model becomes \( \hat{y} = \bar{y} \)? Answer: Mathematically, it is because the slope \( \hat{\beta}_1 \) for the model will be zero.

Intuitively, imagine if I pointed to a guy on the street and said to you, “what’s his IQ given that he wears a size ten shoe.” You would ignore the shoe size info because it’s probably useless here. Then the most logical guess for the guy’s IQ is the average IQ since you have no other useful information.)

Therefore, if \( x \) does not contribute to the prediction of \( y \), \( SS_{yy} \approx SSE \). However, if \( x \) does contribute to the prediction of \( y \), \( SS_{yy} > SSE \). Then a simple measure of the usefulness of the model could be formed as follows:

The **coefficient of determination** is

\[
 r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} = \frac{\text{Explained sample variability}}{\text{Total sample variability}}
\]

The **coefficient of determination**, \( r^2 \), represents the proportion of the total sample variability around the mean of \( y \) that is explained by the linear relationship between \( x \) and \( y \) (see the diagram below).
Example 170: The following table gives male human heights and shoe sizes, find both $r$ and $r^2$:

<table>
<thead>
<tr>
<th>Heights</th>
<th>65</th>
<th>72</th>
<th>60</th>
<th>59</th>
<th>59</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoe Sizes</td>
<td>6.5</td>
<td>9.5</td>
<td>7.5</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Steps to finding $r$ and $r^2$:

1. Create and fill in the preliminary calculation table below:

<table>
<thead>
<tr>
<th></th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i^2$</th>
<th>$y_i^2$</th>
<th>$x_iy_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Calculate $SS_{xy} = \frac{\sum x_iy_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n}$

3. Calculate $SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$

4. $SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$

5. Find $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$

6. Calculate $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$

7. Calculate $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ and calculate $r^2 = \frac{SS_{xy} - SSE}{SS_{yy}}$
Example 171 Use the following data to find \( r \) and \( r^2 \) (the table below will help you considerably)

### Data

<table>
<thead>
<tr>
<th>( x )</th>
<th>3</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

| Table 11-1 Finding Statistics Used to Calculate \( r \) |
|-------|---|---|---|---|
| \( x \) | \( y \) | \( x \cdot y \) | \( x^2 \) | \( y^2 \) |
| 3     | 5  | 15 | 9  | 25 |
| 1     | 8  | 8  | 1  | 64 |
| 3     | 6  | 18 | 9  | 36 |
| 5     | 4  | 20 | 25 | 16 |
| **Total** | **12** | **61** | **44** | **141** |
| \( \Sigma x \) | \( \Sigma y \) | \( \Sigma xy \) | \( \Sigma x^2 \) | \( \Sigma y^2 \) |

On the exam, you may have some of the work done for you like in this example.

### 11.8 Using the Model to Create an Estimation Interval

**Using the model for estimation and prediction**

Two common uses of our regression model are:

1. To estimate the mean value of \( y \) for a specific \( x \) value (section 11.8)
2. To predict an individual value of \( y \) for a specific \( x \) value (section 11.9)

In this section, we will only consider the first of the two above cases.

The formula below is the standard deviation of the sampling distribution of the estimator \( \hat{y} \), when \( \hat{y} \) is being used to estimate the mean value of \( y \) for a specific \( x \) value:

\[
\text{Standard error of } \hat{y} : \quad \sigma_{(\hat{y})} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

Recall that we use \( S = \sqrt{S^2} = \sqrt{\frac{SSE}{n-2}} \) to estimate the value of \( \sigma \).
Now that we know what our sampling error will be, we can form:

Our estimation interval to estimate the average value of $y$ for a specific value of $x$

$$\hat{y} \pm t_{a/2} S \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{a/2}$ is based on $n-2$ degrees of freedom.

**Example 171.5:** A manager is worried about hiring older workers because he fears that they may be more likely to call out sick. He decides to look at the attendance records for sample of employees from the previous year. The data is given below. Find a 95% confidence interval for the average number of sick days used by a 53 year old employee.

<table>
<thead>
<tr>
<th>Employee Age</th>
<th>18</th>
<th>28</th>
<th>38</th>
<th>48</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sick Days Last Year</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

$S = 0.31623$, $\hat{\beta}_0 = 21.14$, $\hat{\beta}_1 = -0.33$, $\bar{x} = 38$, $SS_{xx} = 1,000$

11.9 Using the Model to Create a Prediction Interval

The formula below is the standard deviation of the prediction error for the predictor $\hat{y}$, when $\hat{y}$ is being used to estimate the value of $y$ for a specific $x$ value:

Standard error of prediction: $\sigma_{(\hat{y} - \hat{y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$
Recall that we use \( S = \sqrt{S^2} = \sqrt{\frac{SSE}{n-2}} \) to estimate the value of \( \sigma \). Now that we know what our sampling error will be, we can form:

Our prediction interval to estimate a specific \( y \) value for a given \( x \) value

\[
\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

where \( t_{\alpha/2} \) is based on \( n-2 \) degrees of freedom.

**Example 172:** Find the 95% prediction interval for the height of a woman given that she has a shoe size of 6.

<table>
<thead>
<tr>
<th>Heights (Y)</th>
<th>65</th>
<th>72</th>
<th>60</th>
<th>59</th>
<th>59</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoe Size</td>
<td>6.5</td>
<td>9.5</td>
<td>7.5</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

\( S = 4.1924, \ \hat{\beta}_0 = 44.421, \ \hat{\beta}_1 = 2.617, \ \bar{x} = 7.1, \ SS_{xx} = 10.7 \)
**Step 1:** Use the least squares line to find \( \hat{y} \)

\[
\hat{y} = 44.421 + 2.617(6) = 60.123
\]

**Step 2:** Find \( t_{\alpha/2} \)

Our degrees of freedom is \( n - 2 = 3 \), so our \( t_{\alpha/2} = 3.182 \)

**Step 3:** Find

\[
S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

\[
4.1924 \sqrt{1 + \frac{1}{5} + \frac{(6 - 7.1)^2}{10.7}} \approx 4.804
\]

**Step 4:** Find the Margin of Error = ME = \( t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \)

\[
= 3.182 (4.804) = 15.286
\]

**Step 5:** Finish by getting: \([ \hat{y} - ME, \hat{y} + ME] \)

\[
= [60.123 - 15.286, 60.123 + 15.286]
\]

\[
= [44.837, 75.409]
\]

The interval above is not very good. It is very wide. How could we improve our results? The margin of error determines the interval width, so we need to reduce the size of:

\[
t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

If our sample size was larger we would have more degrees of freedom which would make \( t_{\alpha/2} \) smaller, also the quantity \( 1 / n \) would be smaller. We could get a shorter interval by picking an \( x_p \) that is closer to \( \bar{x} \). Finally, if SSE was smaller, \( S \) would be smaller, but to get a smaller SSE we need a better model. Recall however that the least squares line has the minimum SSE. This means that we need a better predictor variable (\( x \)), or we need to use a more complicated model.
Let’s study the confidence interval widths and the prediction interval widths by looking at the following figure and by looking at the formulas themselves. Try to think about how these intervals can be made more narrow (i.e.—better).

Now, that we have spent a good deal of time talking about linear regression, let’s layout some important guidelines.

**Guidelines for Using the Regression Equation**

1. If there is no linear correlation, don’t use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data.*
3. A regression equation based on old data is not necessarily valid now.
4. Don’t make predictions about a population that is different from the population from which the sample data were drawn.

*We should expand on the second guideline given above. If the data used to create our model used \( x \) values ranging from 6 to 20, don’t try to make predictions about \( x \) values such as 1 or 45 because they are outside of the interval \([6, 20]\) that we used to create our model.
Estimating $y$ beyond the range of values associated with the observed values of $x$ can lead to large prediction errors.

Beyond the range of observed $x$ values, the relationship may look very different.

Example 173: Consider the following data:

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

$a$. Find the least squares line

$b$. Calculate SSE

$c$. Calculate $s^2$

$d$. Find a 95% confidence interval for the mean value of $y$ when $x_p = 2.5$

$e$. Find a 95% prediction interval for the value of $y$ when $x_p = 4.6$

$f$. Would it be wise to make a prediction interval for $y$ when $x_p = 10$?
A complete example

How does the proximity of a fire station (x) affect the damages (y) from a fire?

To answer this question we would need some real world data. After collecting the data, we would construct a scatter plot to see if it seems appropriate to try to fit a linear model to the data. Based on the scatter plot below, it seems plausible to try to fit a straight line model to the data.

There are 15 data points here. For example it appears that at 6 miles from the nearest fire station a home or building would have ~$44,000 worth of damage.

In the real world, the calculations are left up to a computer. Here is the output from SAS a popular software package:
The data produces the following estimates (in thousands of dollars):

\[ \hat{\beta}_0 = 10.28 \]
\[ \hat{\beta}_1 = 4.91 \]

The estimated damages equal $10,280 + $4910 for each mile from the fire station, or

\[ \hat{y} = 10.28 + 4.92x \]

The estimate of the standard deviation for our random error component of the model ($\varepsilon$) is highlighted above under root MSE: $s = 2.31635$. Most of the observed fire damages will be within $2s \approx 4.64$ thousand dollars of their respective predicted values when using the least squares line.

We would now like to check the usefulness of the model: Test that the true slope $\beta_1$ is 0.
\[ H_0 : \beta_1 = 0 \]
\[ H_a : \beta_1 > 0 \]

You can see (next to the Distance variable) on our computer output that SAS automatically performs a two-tailed test, with a reported p-value < .0001. The one-tailed p-value is < .00005, which provides strong evidence to reject the null. Next to that line in the SAS printout you will see a 95% confidence interval on $\beta_1$ from the SAS output is $4.071 \leq \beta_1 \leq 5.768$. This confirms that $\beta_1$ does not seem to be zero.

Another important measure of the model's utility is our coefficients of determination and correlation. Both of the values below indicate a strong linear relationship:

- The coefficient of determination, $r^2$, is .9235. (also highlighted on the SAS output)
- The coefficient of correlation, $r$, is $r = \sqrt{r^2} = \sqrt{.9235} = .96$

Now, we can use our model for predictions. You will see at the bottom of the SAS output we have asked the software to predict the damages for a fire that occurs 3.5 miles from the fire station. The following results were obtained:

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]
\[ \hat{y} \approx 10.2 + 4.92(3.5) = 27.5 \]

95% prediction interval is (22.324, 32.667)

We’re 95% sure the damage for a fire 3.5 miles from the nearest station will be between $22,324 and $32,667. Finally, since the x-values in our sample range from .7 to 6.1, predictions about damages for x-values beyond this range will be unreliable.