**Reading 15**  
Forecasting - Simple and Multiple Regression

**Simple Linear Regression:**

Regression sounds rather Freudian. Maybe it is. If, however, I teach you something about Freudian regression I would have to bill you and you probably pay enough for this course as is, so I will limit our discussion to simple forecasting regression. Simple regression is not too complicated. In fact, the application is very simple and the results are most rewarding when you identify variables that enable you to make better business forecasts. Estimating and forecasting are quite valuable tools for developing some credibility in your company and the business community. Simple regression analysis begins by working with the relationship between two variables. One of the variables is called a dependent variable and the other is called the independent variable. The idea is to find two variables that are related to each other (cause and effect). The movement in one variable (independent) will be used to predict the movement in the other variable (dependent).

For example, suppose that you wanted to forecast the average annual sales of your retail store based on the square footage. The two variables of interest are average annual sales and square footage of the retail store. Let's further assume that the relationship between the two could be shown to be a positive linear relationship. Let's further assume that I have developed, using the method of least squares, a predicting equation using the two variables. The predicting or forecasting equation is stated as follows.

\[ \hat{Y} = 901.247 + 1.686X \]

Here \( \hat{Y} \) is equal to average annual sales in (000). \( \hat{Y} \) is also the **dependent variable**. The **independent variable** is \( X \) which is the square footage of the retail store.

In order to use the predicting equation, I must know or assume a value for \( X \) which would be the square footage of a chosen retail store. With this information, I can determine a value for the dependent variable, average annual sales.

Let's assume that I want to open a store in downtown Fort Worth that is 4,000 square feet. This is \( X \), the independent variable. Substituting 4,000 for \( X_1 \), what is the forecast of my average annual sales? I will simply substitute 4 (000) for \( X \) in the predicting equation. The result will be a forecast of the annual sales of 907.991 or $907,991. \[ 901.247 + 1.686 (4) \]

It looks rather easy, doesn't it?
You ask, “Is the predicting equation always given to me?”

That would make life too easy, so the short answer is “No”. But take heart as Mel Gibson might say, I can show you how to develop a predicting equation from your own raw data set.

Okay, you say. This is the 10th lecture and there are only two more to go so let's get started.

Do I detect a bit of glee? Ignoring that emotion, I move on ever so intently.

The process I will use is called the method of least squares (MLS) or ordinary least squares (OLS) methods. Both are the same methods just given different names. I will address this process shortly, but first I need to introduce you to some additional concepts.

There are two basic models used in simple linear regression. Notice that I have used the term "linear". I will restrict solutions to a straight-line relationship. There are certainly curvilinear relationships, but for this course I will set them aside other than to tell you they exist. For pictures of those relationships, most statistics textbooks show you pictures and tell you the treatment methods associated with curve-linear forecasting. There are two

**Deterministic Model**

The first model I will show you is called a deterministic model. The best way to look at this model is to view it graphically.
From this graphic picture, you can see that all of the data points fall on a straight line. Because they all fall on the straight line, this is called a deterministic model. Each and every point can be determined. The $Y_{\text{hat}}$ value is rent and the $X$ variable is store sales. The actual equation would look like the following.

$$\hat{Y} = b_0 + b_1 (X)$$

Where $b_0$ is the y-intercept and $b_1$ is the slope of the line.

Here the slope $b_1$ is equal to 10% or 0.10 and it is positive (notice the plus sign). Positive slope is upward and to the right. I can develop negative slope predicting equations, which are just as valid as a positive sloped predicting equation. For example, a negative slope would exist if I developed a relationship between number of months a person owns exercise equipment and the number of hours of exercise per week. The supposition would be that when a person first purchases exercise equipment, the motivation is present to exercise for a longer period of time than when the person owns the equipment for a number of months. Somehow our best intentions deteriorate over a period of time. This relationship reflects a negative relationship which might be straight-line.

In my deterministic example from just above, my monthly rent is dependent on my sales. I will pay a base rent of $500 per month (if sales is zero), but I have a variable clause in my lease that requires me to pay based rent plus 10% of my
sales. If my sales were $1,000 for the month, my rent would then be $600. If my sales were $2,000 per month, then my rent would be $500 + (0.10) ($2,000) or $500 + $200 = $700 for the month. Given any sales value, I can predict my rent. This relationship is a positive linear relationship. Because all of the points would fall on a straight line, the model is referred to as a deterministic model.

**Slope of the Line**

Another example of negative slope is price elasticity. Price would be the vertical scale (Y-axis) and quantity would be the horizontal scale (X-axis). The slope of the line is from the left to the right in a downward direction. When price declines, the quantity purchased goes up. The amount by which the quantity increases is totally dependent on the slope of the line. A line that tends toward vertical is an inelastic demand. A line that tends toward the horizontal is considered to be elastic demand. Inelastic demand simply tells you that a big change in price will result in a small change in quantity. The measurements of inelastic relationships are measured between 0 and -1. An elastic demand simply tells you that a change in price will result in a much bigger change in quantity purchases. The measurements of elastic relationships are measured above -1. Notice that I have used a negative sign to describe the values, however, in practice economists ignore the negative signs and refer to the measurement as 1 or 3 or 0.5. This is done simply to prevent a discussion of which is larger a negative 4 or a negative 5. As is applies to elasticity, the higher the value the more elastic (sensitive) is the product to price changes. In real business situations, we do not often run into many deterministic models. Price elasticity is not often a deterministic model, but it does show that negative relationships can be meaningful. What we most often see is a random model, which is described below.

**Probabilistic or Stochastic or Random Model**

All three terms apply to the same type of model. Here all of the points do not fall on a straight line. The data set is random and when plotted the result will be a scatter diagram much like the one shown in the following graph. This graph reflects a study of 10 pizza restaurants where the student enrollment in nearby colleges or universities seems to influence the sales at the pizza restaurants. I want to open an eleventh restaurant, but would like to be able to predict my sales with a reasonable degree of accuracy.
As I look at the data points, a positively sloped, straight-line relationship between the two variables seems plausible. I need a procedure that enables me to develop a predicting equation. I need to fit a straight-line to the scatter points. The straight line will enable me to forecast my sales expected in any new restaurant I want to establish. With this study, I am asserting that there is a relationship between pizza store sales and student population on college campuses. What I am trying to establish is that the sale of pizza near the college campuses is directly related to the student population at the school (cause and effect). Movement in enrollment (cause) can predict sales movement (effect). The raw data for the 10 different pizza restaurants is tabulated below.
<table>
<thead>
<tr>
<th>Store Number</th>
<th>Student Enrollment (000) X</th>
<th>Restaurant Sales (000) Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>88</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>118</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>117</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>137</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>157</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>169</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>149</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td><strong>140</strong></td>
<td><strong>1,300</strong></td>
</tr>
<tr>
<td>Mean = 14</td>
<td>Mean = 130</td>
<td></td>
</tr>
</tbody>
</table>

This table reflects that Store #1 was located in a town where the student enrollment at a nearby college was 2,000. Store #1 experienced $58,000 annual sales. The rest of the table should be reasonable simple to read in the same manner. For calculation ease, the 000 are removed from the data set.

If I can show a relationship between these two variables (student population and pizza restaurant sales), I can use that relationship to forecast the expected sales for a new restaurant. An accurate forecast will help me in planning the size of location, the number of employees, the capital required, the equipment required, etc.

As you look back at the scatter diagram shown in the graph, there appears to be a positive relationship. It appears that a straight line might fit the data set and best describe the relationship. Now the next problem is to actually fit the straight line to the data set.

Remember, the general equation for a straight line is as follows:

\[
\hat{Y} = b_0 + b_1 (X)
\]
To fit a straight line, I must calculate the y-intercept \( (b_0) \) and the coefficient associated with the slope of the line \( (b_1) \). How I do that is by using a technique called the method of least squares. I am interested in the variation between the straight-line and the actual values. In reality there are three y-values of interest.

\[
\begin{align*}
Y & = \text{Actual Sales and is represented by the dots on the scatter diagram.} \\
\hat{Y} & = \text{Predicted Sales and is represented by the straight line on the scatter diagram.} \\
\bar{Y} & = \text{Average Sales of the entire data set.}
\end{align*}
\]

The actual data points as represented by \( Y \) is not on a straight line; therefore, there will be variability between the actual data and the average data as represented by \( \bar{Y} \). Carrying that thought one step further, you will have variability between \( Y \) and \( \hat{Y} \) and lastly variability between \( \hat{Y} \) and \( \bar{Y} \). Since all of the points do not fall on a straight line like the deterministic model, the variability allows us to measure how well a straight line might describe the relationship between the actual data points and the predicting equation line as represented by \( \hat{Y} \). These three sets of variation for the basis for determining two important values to measure how well the straight line fits the actual data points (goodness of fit) and how strong is the fit (correlation). If the actual data points are close to the predicting line, there is a good fit and we can be more comfortable with using the predicting equation as a forecasting model. If the actual data points are widely dispersed from the predicting line, our forecast will be more prone to error and less reliable for use as a forecasting tool. Right now I need you to understand that this variation exists.

Remember, I told you that all I need to do was to determine a Y-intercept \( (b_0) \) and the coefficient of the slope of the line \( (b_1) \). There are two ways to accomplish this. First I can make the calculations manually, which I am going to show you. Second and more effectively, I can use Excel, which I will also show you.

**Manual Approach to the Method of Least Squares or the Ordinary Least Squares Method:**

“Lotsa” columns are the first approach to make these calculations manually. Notice how I skillfully worked in my fluent Italian, since this is a pizza restaurant study. Of course you real Italians are probably just shaking your head about now. Moving on, let's look at a completed table.

**Pizza Restaurant Study**
The Data and The Analysis:

Table 1:

<table>
<thead>
<tr>
<th>Restaurant Number</th>
<th>Student Enrollment - X (000)</th>
<th>Annual Sales - Y(000)</th>
<th>XY</th>
<th>X^2</th>
<th>Y^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>58</td>
<td>116</td>
<td>4</td>
<td>3,364</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>105</td>
<td>630</td>
<td>36</td>
<td>11,025</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>88</td>
<td>704</td>
<td>64</td>
<td>7,744</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>118</td>
<td>944</td>
<td>64</td>
<td>13,924</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>117</td>
<td>1,404</td>
<td>144</td>
<td>13,689</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>137</td>
<td>2,192</td>
<td>256</td>
<td>18,769</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>157</td>
<td>3,140</td>
<td>400</td>
<td>24,649</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>169</td>
<td>3,380</td>
<td>400</td>
<td>28,561</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>149</td>
<td>3,278</td>
<td>484</td>
<td>22,201</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>202</td>
<td>5,252</td>
<td>676</td>
<td>40,804</td>
</tr>
<tr>
<td>Totals</td>
<td>140</td>
<td>1300</td>
<td>21,040</td>
<td>2,528</td>
<td>184,730</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sum X &= 140 \\
\sum Y &= 130 \\
\sum XY &= 21,040 \\
\sum X^2 &= 2,528 \\
\sum Y^2 &= 184,730 \\
\bar{X} &= 14 \\
\bar{Y} &= 130 \\
\end{align*}
\]

You should not have any trouble completing a table like Table 1. XY is the product of X times Y. \(X^2\) and \(Y^2\) are the square of X and Y. The totals at the bottom of the column are sums of the columns. The averages of X and Y are shown below the table. Assuming you could replicated this table, let’s move on to the actual calculations.

Predicting Equation:

General Form: \[\hat{Y} = b_0 + b_1 (X)\]

Where
\[ b_1 = \frac{\sum XY - (\sum X)(\sum Y)}{n} \]
\[ b_0 = \overline{Y} - b_1 \overline{X} \]

Looks pretty challenging, doesn’t it? Remember the \( \Sigma \) (Capital Greek Sigma) means “sum of a column headed …..”

Next I will take the values created in Table 1 above and substitute them into the formulas. By going back and forth from the Table to this solutions you should be able to follow my substitutions and calculations. Solving for \( b_1 \) and \( b_0 \) we have the following:

\[
b_1 = \frac{21,040 - (140)(1,300)}{10} = \frac{21,040 - 18,200}{2,528 - 1,960} = \frac{2,840}{568} = 5.0 \]  

\[b_0 = 130 - 5(14) = 60 \]  

[This too is in (000), so this is actually 60,000.]

Now I simply substitute these two values into the general form of the equation to develop my predicting equation. Therefore, the predicting equation is . . .

\[ \hat{Y} = b_0 + b_1 (X) \]  

General Form

\[ \hat{Y} = 60 + 5 (X) \]  

Specific Predicting Equation for the Pizza Restaurant Study.

My cause and effect statement can be made as follows:

Sales = 60 + 5 (Student Enrollment)

Movement in Sales (the dependent variable) can be explained by movement in Student Enrollment (the independent variable). Student Enrollment increases (cause) result in Sales increases (effect). Conversely, Student Enrollment decreases (cause) result in Sales decreases (effect). These two variables seem
to move together in the same direction. This indicates the slope must be positive. Going back and looking at the sign of the slope of the equation, we find that it is \[ \hat{Y} = 60 + 5 (X) \]. The +5 is the slope of the predicting equation thus the relationship is positive with a slope upward to the right. The variables move in the same direction. As one goes up the other goes up. If the slope were negative, the opposite would be true. The variables would move down to the right and the movement would be in the opposite direction. For example, elasticity is negative. A reduction in prices will cause quantity purchased to rise. The sign of the \( b_1 \) coefficient determines if the slope is positive or negative.

Now let's assume that I have a location in mind for my new pizza restaurant. Let us further assume that the student enrollment for the college campus is 16,000. All I need to do is substitute the values I have calculate above into the specific predicting equation for the Pizza Restaurant Study. The result is as follows.

\[ \hat{Y} = 60 + 5 (X) \]

Sales = 60 + 5 (Student Population)

If I assume that the student population is 16,000 for the town in which I am going to build a pizza restaurant, then I can predict my sales for that store.

Sales = 60 + 5 (16) = 140

Since I am working in thousands, the sales is $140,000 given a student population of 16,000.

This value is a point estimate of my expected sales. This is extremely helpful in my planning process. This gives me some idea about what level of sales I might expect given the relationship between Sales and Student Population. This is an example of the random or stochastic model. This type of model is characterized by the fact that all observations do not fall on a straight line.

For another example, let's say that you wanted to use this equation to predict the sales in a college town with a student population of 50,000. What would the expected sales be for this town?

Your math would lead to expected sales of $310,000 \[ \hat{Y} = 60 + 5 (50) \] but your answer would be wrong. To understand why, you need to go back to Table 1 and look at the range of your Student Enrollment data. It ranges from 2,000 to 26,000. You have not data to support that the straight line relationship between these two point can be extended to include a Student Enrollment of 50,000. Beyond the 26,000 we might have a curve-linear relationship. The point
is that we do not know what the relationship is between 26,000 and 50,000, so we better now extrapolate too far beyond the 26,000 else we might have a flawed forecast.

As I have already illustrated, there is some variability in my regression model. Since I have restricted my techniques to linear, I am now wondering how well the model predicts sales. In other words, how well does the movement in student population (the independent variable) predict movement in sales (the dependent variable)? To determine this, I need to develop a measure referred to as Goodness of fit.

**Goodness of Fit:**

My predicting equation is developed by using a procedure known as OLS or MLS (Ordinary Least Squares or Method of Least Squares). This procedure develops a single line, which is a best-fit line. *This single best-fit line minimizes the variation between the actual and the predicted values.*

In order to determine how well the predicting line fits the actual data points, I need to partition my total variation (SST) into two pieces (SSR and SSE). These are read Sum of the Squares Total (SST), Sum of the Square Due to Regression (SSR) and Sum of the Squares Due to Error (SSE). The larges will always be SST. The other two are part of SST. After I determine these value, I can them calculate three values which help me understand how well the predicting equation works as a forecasting model. One of the measures is the coefficient of determination, which is the ratio of (SSR ÷ SST). Another is the standard error (Se), which is the error due to regression, which is the square root of [SSE ÷ (n-2)]. The third is the coefficient of correlation, which is the square root of the coefficient of determination. The first two are measure of the Goodness of Fit. The third is a measure of the strength of the relationship between the two variables. I can write the variation in the form of an equation.

\[ \text{SST} = \text{SSR} + \text{SSE} \]

What this means is that if I know two of these values, then I know the third. However, the other side of the coin is this. If I miscalculate one of the values, the other value is automatically incorrect. So I guess I better address the calculation of all three of the values. Let's look at the derivation of SST, SSR and SSE.

\[ \text{SSE} = \sum (Y - \hat{Y})^2 \quad \text{Actual less the predicted value squared.} \]

\[ \text{SST} = \sum (Y - \bar{Y})^2 \quad \text{Actual less the average value squared.} \]

\[ \text{SSR} = \sum (\hat{Y} - \bar{Y})^2 \quad \text{Predicted less the average value squared.} \]
One method of determining these values is to develop the following tables. However, this method is really unnecessary, so you can skip the calculations in the tables themselves, but look at the calculation from the tables. Those are important. The tables are for those of you wanting the detail. I will show you a much simpler approach shortly.

**SSE:** Actual less the predicted value, squared.  

<table>
<thead>
<tr>
<th>Predicted Value Using $\hat{Y} = 60 + 5(X)$</th>
<th>$(Y - \hat{Y})$</th>
<th>$(Y - \hat{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>-12</td>
<td>144</td>
</tr>
<tr>
<td>90</td>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>100</td>
<td>-12</td>
<td>144</td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td>324</td>
</tr>
<tr>
<td>120</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>140</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>160</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>160</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>170</td>
<td>-21</td>
<td>441</td>
</tr>
<tr>
<td>190</td>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0</td>
<td><strong>1530</strong></td>
</tr>
</tbody>
</table>

**SST:** Actual less the average value, squared.  

<table>
<thead>
<tr>
<th>Average $(\bar{Y})$</th>
<th>$(Y - \bar{Y})$</th>
<th>$(Y - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>-72</td>
<td>5,184</td>
</tr>
<tr>
<td>130</td>
<td>-25</td>
<td>625</td>
</tr>
<tr>
<td>130</td>
<td>-42</td>
<td>1,764</td>
</tr>
<tr>
<td>130</td>
<td>-12</td>
<td>144</td>
</tr>
<tr>
<td>130</td>
<td>-13</td>
<td>169</td>
</tr>
<tr>
<td>130</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>130</td>
<td>27</td>
<td>729</td>
</tr>
<tr>
<td>130</td>
<td>39</td>
<td>1,521</td>
</tr>
<tr>
<td>130</td>
<td>19</td>
<td>361</td>
</tr>
<tr>
<td>130</td>
<td>72</td>
<td>5,184</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>0</td>
<td><strong>15,730</strong></td>
</tr>
</tbody>
</table>

| Average Deviation   | SSE             | Average Deviation | SST             |
If you remember, I indicated the following relationship exists.

\[ \text{SST} = \text{SSR} + \text{SSE} \]

Since this is true and I know SST and SSE, I can now re-write the formula as follows.

\[ \text{SSR} = \text{SST} - \text{SSE} \]

I can now calculate SSR as follows.

\[ \text{SSR} = 15,730 - 1,530 = 14,200. \]

Remember, this works well as long as you do not make an error in the calculation of SST or SSE. If you miss one of those, SSR will also be incorrect.

To solve for SSR directly, you would create the following table.

**SSR:** Predicted value less the average value, squared.  

<table>
<thead>
<tr>
<th>( (\hat{Y} - \bar{Y}) )</th>
<th>( (\hat{Y} - \bar{Y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-60</td>
<td>3600</td>
</tr>
<tr>
<td>-40</td>
<td>1600</td>
</tr>
<tr>
<td>-30</td>
<td>900</td>
</tr>
<tr>
<td>-30</td>
<td>900</td>
</tr>
<tr>
<td>-10</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td>40</td>
<td>1,600</td>
</tr>
<tr>
<td>60</td>
<td>3,600</td>
</tr>
<tr>
<td><strong>0</strong></td>
<td><strong>14,200</strong></td>
</tr>
</tbody>
</table>

My purpose, long since forgotten by you, is to know how well the predicting line fits the actual data points. I have developed a forecasting model, so how well does it work. This is known as Goodness of Fit. One measure of Goodness of Fit is the Coefficient of Determination (CD).

\[
\text{Coefficient of Determination} \, (\text{designated } r^2) = \frac{\text{SSR}}{\text{SST}} = \frac{14,200}{15,730} = 0.90.
\]

So what does this mean, you ask?
Glad you ask, I reply.

First, the CD is measured between 0 and 1. A CD of zero (0) would mean either no straight-line relationship or a curve-linear relationship. The closer to 1, the better the predicting equation fits the actual data.

The coefficient of determinations is a measure of goodness of fit. In our Pizza Restaurant Study, it may be interpreted as follows.

90% of the variation in sales (dependent variable) is explained by variation in student population (independent variable).

How much variation can I ever explain?

100%.

So if I am explaining 90%, my relationship between the two variables is quite strong. I can only explain another 10 percentage points. Let's look at a scatter diagram complete with all three values I have been discussing the Actual Data, Average Data and Predicted Data. This is a summary presentation of the three types of variation.

From this chart, you can see the three sources of variation. To calculate both measures of goodness of fit and the measure of strength, I need these three variations.

Now for the simpler approach to manual calculations.
Computational Ease for SSR and SST:

Go back and look at Table 1. I had you calculate a $Y^2$ column; however, you have not used it so far, right? (Answer is - right!). Some of you have been wondering why. Some of you did not notice. Some of you did not care, but most of you would really like a kinder, friendlier approach. The easy computations include $Y^2$. What you do is the following.

You calculated a value for $b_1$.

Let’s begin with that value. You may have to go back and find this from your original calculation, but I will repeat it here.

$$b_1 = \frac{2,840}{568}$$

To directly calculate SSR, you begin with the $b_1$ value in the form stated just above. You would then square the numerator of this equation.

$$SSR = \frac{(2,840)^2}{568} = \frac{8,065,600}{568} = 14,200$$

Wow, you exclaim! Professor you are brilliant!

I bow graciously and soak in your accolades. Wait, I humbly offer, there is more.

More, you question? My heart might now be able to stand up under this new information.

Give it a rest!!! Yep, I can also determine SST directly by using the $Y^2$ column previously discussed.

$$SST = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$SST = 184,730 - \frac{(1,300)^2}{10}$$

$$SST = 184,730 - 169,000 = 15,730$$

These calculations are much easier and simpler than using the long-form of the sum of the squares as shown in Tables 2, 3 and 4. You have a choice. You
may work the problems the long way or work the problems the short way. I know which I would use.

As I have previously shown you, the coefficient of determination (goodness of fit measure) is \( \frac{SSR}{SST} = \frac{14,200}{15,730} = 0.90 \).

I have also shown you that the interpretation is **90% of the variation in the dependent variable (sales in this case) is explained by variation in the independent variable (student population in this case).** This is the only acceptable way to explain the coefficient of determination.

I can carry the concept of goodness of fit one step further. I can calculate something called the standard error (\( S_e \)), which is nothing more than the standard deviation of the regression relationship.

**Standard Error of the Estimate**

The standard error can be determined as follows.

\[
S_e = \sqrt{\frac{SSE}{n-2}}
\]

\[
S_e = \sqrt{\frac{1,530}{10-2}}
\]

\[
S_e = \sqrt{191.25}
\]

\[
S_e = 13.829
\]

I determined SSE by subtracting the two known values SST and SSR (15,730 – 14,200).

That is a fairly simple calculation, but what do I do with it?

Remember in the example above I posed the question of what would my sales be if I placed a store in a town with a student population of 16,000. The answer was that I expected my sales to be $140,000. The $140,000 is determined by simply inserting 16 into the predicting equation, which was

\[
\hat{Y} = 60 + 5 (X).
\]
The $140,000 is a point estimate, an estimate of the sales at a given point. I wonder if I can do anything more to enhance my sales estimate, since I know that the probability of exactly hitting the estimate of $140,000 is close to zero?

The answer is yes. I will show you a quick and dirty use of the standard error. There is a more accurate approach, which I will cover shortly – global predicting values and individual predicting values. Just be aware there are more sophisticated techniques to be used. Since I am working with sales dollars, the data is continuous. Since it is continuous, I can use the empirical rule interpretation and develop a range for my sales. The Empirical Rule tells me if I move 1 standard deviation (standard error in this case) in either direction from the mean, I will encompass 68.3% of my observations. Two standard deviations (standard error) in either direction from the mean will encompass 95.5% of my observations, etc. Let's assume that I want to be 68.3% sure my sales will fall between two values. I would simply do the following.

$140,000 \pm (1)(13,829) = $121,171 to $151,829 \quad (This \ approach \ gives \ you \ good \ planning \ information.)$

So I can be 68.3% certain my sales will fall between $126,171 and $151,829. This helps me better plan my staffing, equipment purchases, tables and chairs, rental space and cash requirements. This tool is excellent in aiding you in forecasting no matter what the two variables may be.

I am sure most of you are sitting at your computer with hand allegorically raised to pose the next question. Your brain waves have conveyed to Carnak that the next answer to the question is "correlation coefficient". The all seeing and all knowing Carnak says the question is "How strong is my relationship between sales and student enrollment?"

Strength is a different measure. Strength is measured by using something called the coefficient of correlation.

**Coefficient of Correlation**

The coefficient of correlation is the square root of the coefficient of determination. The purpose of the coefficient of correlation is to measure the strength of the relationship.

The coefficient of correlation is measured between the values of -1 to +1. A coefficient of correlation of -1 says that the relationship is a perfect, **negative** linear relationship. A coefficient of correlation of +1 says that the relationship is a perfect, **positive** linear relationship. Both of these would be deterministic models with each and every observation falling on the straight line. A value closer to zero
indicates the relationship is non-linear, but tends toward a curvilinear relationship, which is beyond the scope of this teaching at this time.

In this instance, the coefficient of correlation is as follows.

\[ r^2 = 0.90 \text{ from previous calculations} \]

\[ r = \sqrt{0.90} = 0.95. \]

The strength of the relationship is excellent, close to the best of 1.0. However, there is one thing. When I take the square root of any number, the outcome can be a plus or a minus.

How do I determine which this one is?

I must go back to my predicting equation and look at the sign that accompanies the slope. If the sign is negative, then r is also negative. If the sign is positive, then r is also positive. In this case, the sign of the slope is positive; therefore, the sign of r is also positive. I have repeated the predicting equation below so you won't have to scroll up to find it. Notice that the sign of 5 is positive. This concept is a repeat of what we have already covered.

\[ \hat{Y} = 60 + 5(X) \]

**Caveats:**

I have already shown you one warning. Do not go beyond the range of your data set in forecasting. One of my previous questions asked you to calculate my sales using the predicting equation if I had a student population of 50,000. Your mathematical answer was $310,000. While mathematically correct, I pointed out that I did not know what happened to the relationship beyond the last data point of 26,000. Beyond that point, the relationship could be curvilinear or could have a different slope. I just do not know, so I must be careful to not go beyond the data in the data set. Does this mean that I could not forecast for a student population of 27,000 since this value is beyond the last data point of 26,000? No, it does not. What I would do is restrict my forecast to no more than 10% above my last data point. Using this rule of thumb, I could develop a forecast up to 28,600. Beyond that I would be very nervous about the outcome.

Secondly, regression and correlation cannot determine cause and effect. You must use common sense when you are determining the variables. For example, the number of storks in London may correlate very nicely to the number of babies being born in London, but my mother told me there is no relationship between these two variables. In short, a number of variables might yield high correlation
values, but this does not mean that there is a cause and effect relationship. Use common sense.

Thirdly, along the same line, you can get spurious correlation. Two unrelated variables can yield a high correlation coefficient. The number of elephants born in the Kansas City Zoo may have a high correlation to the tonnage of salmon caught by sports fisherman in Alaska. A high correlation does not mean the variables are related. Once again use common sense.

Fourth, my predicting equation in the pizza restaurant study has been developed using 10 pizza restaurants that are located near a college campus. My results are very encouraging. My coefficient of determination is 0.90 (close to 1.0, which is the best). However, remember the idea of random selection where each and every element in the population has an equal and independent chance of being selected? How do you suppose that plays into my selection of the 10 pizza restaurants? To have a random selection, I would need to select the 10 pizza restaurants from the 100's in the population. What does it do to my study if I don't do this? Well it might invalidate my 0.90 coefficient of determination or my coefficient of correlation of +.95. Suppose that my 10 restaurants are not selected at random from among the population. The real results from the population might not show such a strong relationship. Do you suppose there is a way I can test this issue?

Of course, you reply, or you would not have brought it up. How astute of you, I reply.

**Beta and Rho Testing:**

There are two tests I can conduct to see if my findings are supportable in the population. One is called the beta test and the other is called the rho test. Both beta and rho are Greek letters designating a test (beta) for the coefficient of determination and a separate test (rho) for the coefficient of correlation. Usually you do not have to develop both tests, since the findings of one should be identical to the findings for the other in simple regression. However, the results in multiple regression may be different for the two values, so that is why the two exist.

Let's look at the calculations for beta and rho for the pizza restaurant study. The question I want to ask myself is can I support the conclusion that pizza sales are a function of student enrollment which will yield a coefficient of determination of .90 and a coefficient of correlation of +.95.

The beta test is appropriate for testing the coefficient of determination. Here if I can assert that the slope of the actual but unknown population regression line is zero, I can then assert there is no relationship between pizza sales and student enrollment. The hypothesis to be tested is as follows:
\[ H_0: \beta_1 = 0 \] (beta)

\[ H_a: \beta_1 \neq 0 \]

Here the \( \beta_1 \) is representative of the population coefficient of the slope of my equation.

To test this hypothesis, I will use a \( t \)-test.

\[
t_{TEST} = \frac{b_1 \sqrt{SS_X}}{S_e}
\]

Where \( SS_X \) is equal to the following.

\[
SS_X = \sum x^2 - \frac{\left(\sum x\right)^2}{n}
\]

If you will notice \( SS_X \) is the denominator of \( b_1 \) equation. You may have to scroll up to verify what I am telling you.

The solution:

\[
t_{test} = \frac{5 \sqrt{568}}{13.829} = \frac{(5)(23.8328)}{13.829} = \frac{119.1638}{13.829} = 8.61
\]

This is my calculated \( t \)-value. For all testing I match a calculated value to a critical value.

At the alpha level of 0.05, I will look up the critical \( t \)-value in the \( t \)-table in a textbook. The critical \( t \)-value is 2.306 using \( n - 2 \) degrees of freedom. Check this out before proceeding.

Since the calculated value of \( t \) is 8.61 is greater than the critical \( t \)-value of 2.306, I would reject the null hypothesis.

Great, you say, but what does this mean?

Glad you asked, I respond. Since I am rejecting the null, I can conclude that the population coefficient, \( \beta_1 \), is significantly different from zero. Since I reject the null, I can support that the population value would be different than zero. This allows me to assert that my pizza restaurant study can be support in the
population. My coefficient of determination of 0.90 is supportable in the general population.

Let's test the coefficient of correlation. Here I use rho to represent the population correlation coefficient. Here the hypothesis being tested is as follows:

\[ H_0: \rho = 0 \] (rho)

\[ H_a: \rho \neq 0 \]

Here the formula is:

\[ t_{\text{TEST}} = \frac{\rho \sqrt{n - 2}}{\sqrt{1 - \rho^2}} \]

The solution:

\[ t_{\text{test}} = \frac{0.95\sqrt{10 - 2}}{\sqrt{(1 - 0.95)^2}} = \frac{(0.95)(2.8284)}{0.3122} = \frac{2.687}{0.3122} = 8.61 \]

Notice that the calculated t-test value is the same for beta or rho testing for simple regression (8.61). When you make these calculations for multiple regression, the calculations may differ, but for simple regression they should be the same.

Here again my conclusion is to reject the null hypothesis and conclude that the alternate is true. This means that my Pizza Restaurant Study can support a positive coefficient of correlation.

Wondering why I have to go to this trouble to test the population values?

I would refer you to the comments I made a bit earlier about random sampling. I may have a sample that does not represent the population. The sample I chose may be biased, so I should always look at least one of these two calculations. Does this absolutely mean that bias does not exist? Of course not, in statistics nothing is absolutely proven, but I can make assertions from these tests.
More Accurate Uses of the Standard Error:

Remember, I gave you a quick and dirty method of using the standard error in the pizza restaurant. I used the empirical rule since I was working with dollars (continuous data) to fit an interval around the point estimate of the sales of $140,000 given student enrollment of 16,000. I determined that I was 68.3% sure the sales should fall between $126,171 and $153,829. I do feel more comfortable about this range since my tests of beta and rho have shown to support my assertions about the relationship between sales and student enrollment.

Now let's look at two approaches that are also useful. The standard error can be used in estimating two different values. One value is considered to be a global value (referred to as a confidence interval) and one is considered to be an individual interval (referred to as a prediction interval). Let me explain through an example and then through a real life problem.

The Example: Let's say that I want to develop an estimate of the salary of all retail executives with 20 years experience. To do this I would calculate a confidence interval (global implications). However, if I want an estimate of the salary of Curtis Bender, a particular retail executive with 20 years experience, I would calculate a prediction interval (individual implications).

The Real Life Problem: Let's say that the Bradford Electric Company is studying the relationship between kilowatt-hours per month (in thousands) and the number of rooms in a private, single-family residence. A random sample of 10 homes yields the following information.

<table>
<thead>
<tr>
<th>House Included in Study</th>
<th>Number of Rooms</th>
<th>KW Hours (000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Totals</td>
<td>91</td>
<td>74</td>
</tr>
</tbody>
</table>

Rather than have you make the calculations, I am going to give you some of the data you need.
The Predicting Equation: \( \hat{Y} = 1.333 + 0.667(X) \)

\[
\begin{align*}
\bar{X} & = 9.1 \\
\sum X^2 & = 895 \\
\sum X & = 91 \\
S_e & = 0.898 \\
N & = 10
\end{align*}
\]

Using the predicting equation, the question is “What is the predicted KWH usage for a 6 room house?”

\[
\hat{Y} = 1.333 + 0.667(6) = 5.333 \text{ KWH}
\]

At an alpha level of 0.05, calculate both the confidence interval and the prediction interval. Remember the confidence interval is for any six-room house while the prediction interval is for a specific six-room house.

I will first address the confidence interval (global values).

The formula looks horrific. Go take some Rolaids and come back.

\[
\hat{y} \pm t_\alpha S_e = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum X^2 - (\sum X)^2}}
\]

Oh, well, guess the Rolaids did not help because the formula is still the same. Let's tackle it anyway.

The calculations are really not that difficult as long as you are careful about inserting values.

\[
5.333 \pm (2.306)(0.898) \sqrt{\frac{1}{10} + \frac{(6 - 9.1)^2}{895 - (91)^2}}
\]
\[
5.333 \pm (2.306)(0.898) \sqrt{0.10 + \frac{(9.61)}{895 - \frac{(8,281)}{10}}}
\]

\[
5.333 \pm (2.306)(0.898) \sqrt{0.10 + \frac{(9.61)}{895 - 828.1}}
\]

\[
5.333 \pm (2.306)(0.898) \sqrt{0.10 + 0.1436}
\]

\[
5.333 \pm (2.306)(0.898)0.2436
\]

\[
5.333 \pm (2.306)(0.898)(0.4936)
\]

\[
5.333 \pm 1.0221
\]

**4.3109 to 6.3551** This is the 95% global estimate of the KWH usage of all 6-room houses.

Okay, you say. All of the numbers came from the ones given except for the 2.306, so where did you get that number?

Oh, I reply. This is a t-test with a sample size of 10. The alpha is 0.05 and the degrees of freedom are \(n - 2\) or 8. You can verify the 2.306 in the t-table in the back of your book. Check it out. Confidence intervals are two tailed tests, so look under the two tailed alpha value of 0.05 and the degrees of freedom of 8.

**Interpretation:** I am 95% confident that the mean usage of all six-room houses (global) is between 4.3 and 6.4 KWH (rounded). This is known as a **confidence interval** because is applies to all six-room houses.

To determine an interval for a **specific** six-room house located at 1212 Avenue H, I would calculate a **predicting interval**.

The formula:

\[
\hat{y} \pm t_{\alpha} \cdot S_e = \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}
\]
Wait a minute you say. This formula looks like the one for the confidence interval just above.

Good observation, I say. They are essential the same with one difference. Can you find the difference?

Yeah, you reply. The formula for the predicting interval has an additional "one" (1) under the square root.

That's it, I confirm, how brilliant of you.

I can begin my calculations for the predicting interval by adding one to the .2435 as shown in the fifth step of the calculation for the confidence interval just above.

\[
5.333 \pm (2.306)(0.898)\sqrt{1.2436}
\]

\[
5.333 \pm (2.306)(0.898)(1.1152)
\]

\[
5.333 \pm 2.3205
\]

3.0125 to 7.6535 This is the predicting interval for a particular 6 room house located at 1212 Avenue H.

*Interpretation:* I am 95% confident that the mean usage of a particular (located at 1212 Avenue H for instance) six-room house is between 3.01 and 7.65 KWH (rounded). This is known as a predicting interval and is broader than the mean usage for all six-room houses. The predicting interval should be wider (broader) than the confidence interval because it is much more difficult to accurately forecast the mean usage of a particular six-room house than it is for all six-room houses.

**Multiple Regression:**

In the preceding portions of this lecture, I have been dealing with simple regression. Simple regression is where there is one dependent variable (Y) and one independent variable (X). Let's say that I calculate a coefficient of determination of .85 when I develop a simple regression relationship between sales and disposable income. Let's say that I am not totally satisfied with the .85. The .85 tells me that 85% of the variation in Y (sales price of a home) is explained by variation in X (home size in square feet). I want to be able to explain more, so I add another independent variable to the equation (age of the house). In fact, I find several independent variables that I think are important such as number of bedrooms, number of bathrooms and garage size. By adding
the additional independent variables to the equation, I will establish a multiple regression predicting equation that would look like the following.

\[ Y_{\text{hat}} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots + b_nX_n. \]

There is still only **one dependent variable**. **All of the X-variables are independent variables**. Let us assume that I have added the second variable and then calculate a coefficient of determination for the one dependent variable and the two independent variables. The actual model building would be a process of using a computer-generated program such as Excel to develop the correlation coefficient for each one of the independent variables as it relates to the dependent variable. The Excel process also allows for the development of the relationship between each pair of independent variables. You will use the t-test and determine if the correlation coefficient is statistically significant at the specified degrees of freedom. Having said all of this, relax, I am not going to specifically address the calculations in detail using a manual approach.

**Predicting Equation:**

The Excel program will also develop the coefficient for each of the independent variables for you. Let's assume the specific formula for the house sales price problem is as follows.

\[ Y_{\text{hat}} = 31,327.60 + 63.10 \text{(sq. ft.)} - 1,144.40 \text{(Age)} - 8,410.40 \text{(bedrooms)} + 3,522.00 \text{(bathrooms)} + 28,203.50 \text{(garage)}. \]

The question is how do you interpret the results? All of the coefficients come directly from an Excel table. First, the y-intercept is $31,327.60, which means that if the sq. ft., age, number of bedrooms, number of bathrooms and garage are all set at zero, the sales price of the home is $31,327.60.

The first independent variable, square footage, and the coefficient of +63.10 can be explained as follows. Every time the square footage increases, the price of the house will increase by $63.10, holding all other independent variables constant.

The second independent variable, age, and the coefficient of -1,144.40 can be explained as follows. Every time the age of the house decreases by one year the price of the house will decrease by $1,144.40, holding all other independent variables constant.

The third independent variable, number of bedrooms, and the coefficient of -8,410.40 can be explained as follows. Every time the number of bedrooms drops by one, the sales price of the house will drop by $8,410.40, holding all other independent variables constant.
The fourth independent variable, number of bathrooms, and the coefficient of +3,522.00 can be explained as follows. Every time the number of bathrooms increase by one, the sales price of the house will increase by $3,522.00, holding all other independent variables constant.

The fifth and final independent variable, number of garages, and the coefficient of +28,203.50 can be explained as follows. Every time the number of cars able to locate in the garage increases by one, the sales price of the house will increase by $28,203.50, holding all other independent variables constant.

This equation, like the simple regression equation, can be used to estimate the sales price of a house. Let's look at one example. Assume the following;

Square Footage is 2,100.
Age is 15 years.
Number of Bedrooms is 4.
Number of Bathrooms is 3.
Size of Garage is 2 cars.

What do you think the sales price of this house should be?

Making the substitutions, I have the following.

\[
\text{Sales Price} = 31,127.60 + 63.10 (2,100) -1,144.40 (15) - 8,410.40 (4) + 3,522.00 (3) + 28,203.50 (2).
\]

\[
\text{Sales Price} = \$179,802.70. \text{ Rounded to } \$180,000.
\]

**Coefficient of Determination:**

The multiple regression equations for determining SST and SSR are rather complicated without the use of computer software, so I will not go through those calculations. The Excel output will always give you two different coefficients of determination. One is called \( R^2 \) and the other is called \( R-bar^2 \). The wording in the Excel printout will read R-Square and Adjusted R-Square.

While \( r^2 \) measures the coefficient of determination for simple regression, \( R^2 \) is the measure of the coefficient of determination for multiple regression.

\[
R^2 = \frac{SSR}{SST}
\]

\[
R^2 = 1 - \frac{SSE}{SST}
\]
However, I run into a problem when I use either of the above relationships. A mathematical phenomena occurs as I add independent variables. The value of $R^2$ will continue to rise even if there is no relationship between the new independent variable and dependent variable. I have heard of horror stories of some marketing research projects which have up to 150 independent variables in their model. This can be very misleading, if the following adjusted $R^2$ has not been used.

I must use something called R-bar squared. The adjustment includes $k$ in the relationship where $k$ is the number of independent variables. Each time I add an independent variable, I lose a degree of freedom. This means that the multiple coefficient of determination will decline or stay even, if the new independent variable has no linear relationship and explanatory power. The new $R$-bar$^2$ is stated as follows.

$$
R^{-2} = 1 - \frac{SSE}{SST} \left( \frac{n - k - 1}{n - 1} \right)
$$

This is read R-bar squared.

The goal in developing a good multiple regression equation is to have as few independent variables as possible to explain as much of the variation as possible. This is called a Parsimonious model.

There is a second problem associated with multiple regression. This problem is multi-co-linearity, meaning that one or more of the independent variables are linearly related and the correlation between those two variables is extremely high. In reality there must be some sort of linear relationship between all of the independent variables, because movement in one or two is used to predict the movement in another. It is only logical that there is some linear relationship between the independent variables. There is no magic cut off point in determining the degree of relationship, unfortunately. The only real solution is to eliminate or combine the offending independent variables. This is done by using t-test for the matched pairs. Sometimes you can change the units of measure. For example, you might be using sales and square foot as two independent variables. You might want to combine the variables to sales per square foot. This would eliminate one independent variable thus eliminating the possibility of multi-co-linearity for the two variables.

I still have the same issue with the multiple regression equation as I did with the simple regression equation. I have only given you partial data and have walked you through the interpretation. What I have not told you is that the information for
the multiple regressions was developed from a sample. Any time I use a sample, I must question if the predicting equation is supportable in the population. This leads to additional testing using the F-test statistic to determine if the model is significant. This information is also contained in the Excel printout under the ANOVA heading. Additionally beta and rho testing are also helpful in making sure the sample you have taken and the conclusions you have reached can be supported in the population. One final word about these verification processes. I must conclude that the standard error is not too large so as to make an interval unhelpful. For example, let's suppose that the standard error in this instance is $54,700. On a rough basis, I then could conclude that the point estimate of $179,802.70 plus and minus $54,700.00 might yield too large an interval to be of any real help ($234,500 to $125,102) in predicting the sales price of a home. Common sense rules quite often.

**Dummy Variables:**

Remember I have two types of variables - quantitative and qualitative. The quantitative variable is not difficult to mathematically manipulate, but the qualitative variable has no value that can be manipulated. Variables that are not expressed in a quantitative fashion are difficult to use in the calculation procedures. Examples of these variables are gender, hair color, religious preference, etc. In order to offset this problem, a dummy variable is developed. The dummy variable accounts for the qualitative nature of a variable and incorporates its explanatory power into the model. If I had two data sets, one for females and one for males, I could use a zero to represent the female and a one to represent the male (although you ladies taking this class might object to being called a zero). In reality I could use the zero for the males and one for the females, which I suspect will better please you ladies. The point is this. By using something called a dummy variable, I can develop a predicting equation for both the males and the females separately. This process helps overcome our inability to use qualitative variables in our studies.

**Curve-Linear Relationships:**

Many of the relationships in dependent variables and independent variables cannot be explained by straight-line relationships. Often the relationship is that of a curve-linear relationship. For example, let's assume I am interested in the relationship between taxes and population. If I measure taxes in millions of dollars (dependent variable) and population in millions of people (independent variable), the data will probably support that as the population increases, the taxes will increase at a rate faster than linear. This is logical. When this case exists, the straight-line linear equations will not work. Most often a polynomial (usually second degree) will best describe the relationship. The polynomial
simply says that as the independent variable increases the dependent variable increases at an increasing rate. The equation is as follows.

\[ \hat{y} = b_0 + b_1X + b_2X^2 \]

Most statistics textbooks address the issue of curve-linear relationships. All I will mention to you is that they exist. Again, if you run into this type of relationship in your analysis, Excel will be very helpful in working with the data set.

**Use of Excel for Regression Model Building:**

Let’s try working a problem using Excel exclusively. Open the second document and follow it through. It follows a simple regression solution then adds another independent variable to allow you to work the problem as a multiple regression problem.

You will find this in the next reading section.