Chapter 8

Simple Regression[1]

So far, we have encountered the idea that variables may be related to each other. Very often, we can use these relationships to determine how one variable that is easier to control affects another variable that we are interested in. To develop such relationships, we can plot the data and try to find an equation that relates the two variables. What we need, though, is a systematic way to decide what the best equation is to fit the data. We will start by using the simplest equations, linear models, to represent the data. The equations for the models will be developed using **least squares regression analysis**. This is a technique in which a line is assumed to exist that fits the data. By manipulating the slope and y-intercept of this line, it can be made to fit the data better. The "best fit" occurs when a certain quantity, **the total squared error**, is made as small as possible. You have already explored this in chapter 7 with the idea of trendlines. Excel calculated all its trendlines using least squares regression.

Not all data looks like a straight line when it is graphed. Since we can always find a "best fit" line for the data, we need some way of determining whether the linear regression equation is a good choice. To decide this, we will make use of several statistical measures and some diagnostic graphs. These will help us answer two important questions: Is the data close enough to linear to make a linear regression equation worth using? If we use the regression equation to predict information, what kind of error can we expect to have in our estimates?

- *As a result of this chapter, students will learn*
  - √ The meaning of R^2 and S_e for regression models
  - √ What residuals are
  - √ What regression models are used for
  - √ How R^2 and S_e are computed
- *As a result of this chapter, students will be able to*
  - √ Write down a regression model based on Excel or StatPro output
  - √ Use a regression model to make predictions
  - √ Explain the meaning of a simple regression model

√ Apply S_e to predictions from a model
√ Use marginal analysis to interpret the slopes of a linear model

8.1 Modeling with Proportional Reasoning in Two Dimensions

At this point, we know a lot about straight lines. What we need to do now is to use this information to build models of the data. These models will allow us to explore what happens to the dependent variable for different values of the independent variable. Right now, we're after models that are proportional. A proportion is simply a ratio (a fraction) between two quantities. Most people use the term **proportional model** interchangeably with **linear model**. These are models in which the change in the y-variable is in a fixed proportion to the change in the x-variable. What this means is that no matter what x-value you are looking at, if you increase it by a fixed amount, any fixed amount, then the change in y is fixed by the constant of proportionality. In this case, the constant of proportionality is the slope of the line.

Consider the cost of manufacturing widgets. (**Widget** is simply a word to describe something non-specific; a widget could be anything: a baseball bat, an engine part, or even a sandwich.) Normally there are fixed costs associated with manufacturing. These costs are constant, regardless of how many widgets you make. The fixed costs include things like payment for the production facilities, coverage of salaried employees, electricity, and other costs that are reasonably constant. Fixed costs look pretty much like a y-intercept on a graph. There are also variable costs in production. These include the cost of the materials to make the widgets and the wages of the employees who make the widgets. They may also include costs for quality control. Clearly, the more widgets you make, the more materials and labor you will use. It is easiest to assume that these variable costs are like the slope of a linear model, so each additional widget adds a certain amount of cost to the total manufacturing costs. Thus, we have a linear model:

Total cost of producing widgets = Fixed Costs + (Variable Costs) * (Number of Widgets)

(There are certainly other ways of modeling cost, but this is the easiest to understand, so it makes a nice starting point.) Suppose your fixed costs are \$1,000 and the variable costs are \$3.50 per widget. If you make 10 widgets, it will cost you \$1,000 + \$3.50 per widget* 10 widgets = \$1,000 + \$35 = \$1,035. If you make five more widgets, for a total of 15 widgets, it will cost \$1,000 + \$3.50 (15) = \$1,052.50, exactly \$17.50 more. Notice that \$17.50 = \$3.50 * 5. In a proportional model, no matter what the current production level is (how many widgets you are making), the model always predicts the same change in y for a fixed change in x. Such models are sometimes called **level independent models**.

What this really means is that whether you are making 10 widgets or 20,000 widgets, if you make 5 more widgets it will cost an additional \$17.50. This is the reason that linear models are so useful; they are easy to interpret. Each coefficient in the equation of the model has a meaning that is easily understood in terms of the problem context. Just to keep you even more confused, economists often refer to marginal costs, the additional cost you will pay in order to make one more widget. For a linear model, the marginal cost is simply the slope.

For more information, see the interactive Excel workbook C08 StepByStep.xls.

8.1.1 Definitions and Formulas

Simple linear regression This is a process for systematically determining the equation of the best-fit line to a given set of (x, y) data. The regression equation is determined by a process called least squares regression and results in a formula to compute the slope and y-intercept of the line that will minimize the "total squared error" of the line. Based on some theoretical calculations with calculus, you can show that the slope, B, of a regression line is given by

$$B = \text{Corr}(X, Y)\frac{\sigma_y}{\sigma_x}$$

where $\text{corr}(X, Y)$ represents the correlation of the variable X with Y and the σ represent the standard deviations of the X and Y variables. Once you have the slope, the y-intercept is easy to find: $A = \bar{Y} - B\bar{X}$, where $\bar{X}$ and $\bar{Y}$ are the means of the X and Y variables.

Proportional Two quantities are proportional when a specific amount of change in one of the quantities results in a certain amount of change in the other quantity given by a fixed multiplicative factor. In mathematical terms, the phrase "the change in y is proportional to the change in x" can be written as $\Delta y \propto \Delta x$. This means that $\Delta y = k\Delta x$ for some constant k that is independent of y and x.

Coefficient A coefficient is a fixed (or constant) number in an algebraic model. For example, linear equations have two coefficients: the slope and y-intercept. Coefficients are sometimes called **constants** or **parameters**. In regression output, a table of coefficients is produced. It is up to you to combine these correctly into the model equation. In the examples below, you will see how to do this.

Constant In most regression output, the y-intercept of the regression line is labeled "constant" in the table of coefficients. More generally, a constant is any value in a formula that is fixed, like the number 2 in the linear relationship $y = 2x + 5$. Sometimes constants are called parameters.

Explanatory Variable This is the variable (or variables, in later chapters) used to explain the results of the model. In simple regression, the x-variable is the explanatory variable. As you can guess, this is just another name for the independent variable.

Response Variable This is the variable that responds to the independent (or explanatory) variable. Thus, it is really the y-variable or dependent variable.

8.1.2 Worked Examples

Example 8.1. Translating Regression Output Into an Equation
Regression output from StatPro's simple regression routine will look like the screen below. This regression output comes from the data on backpacks in example 7 (page 219) from the

last chapter. The data is in "C07 Backpacks.xls". Notice that the output is divided into three areas by headings in bold: summary measures, ANOVA table, and regression coefficients. For right now, we are concerned with the regression coefficients. In the next section we'll come to understand the summary measures (which explain how good and accurate the model is) and a little of the ANOVA table (ANOVA stands for Analysis Of Variance; it is used for computing the summary measures.)

Results of simple regression for Price

Summary measures

Multiple R	0.7022
R-Square	0.4931
StErr of Est	9.8456

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	2640.4857	2640.4857	27.2397	0.0000
Unexplained	28	2714.1810	96.9350		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-30.6751	13.5969	-2.2560	0.0321	-58.5272	-2.8231
Number of Books	1.4553	0.2788	5.2192	0.0000	0.8842	2.0265

From this output, we can easily write down the linear equation that best estimates the price of the backpack, based on the explanatory variable "Number of Books". We know that "Price" is the response variable from the first line of the regression output; it says "Results of simple regression for Price". The dependent variable will always be given here. To write down the regression equation, we need to know only slope and y-intercept.

The y-intercept is the "Coefficient" next to "Constant" in the regression coefficients portion of the output. Thus, this model has a y-intercept of -30.6751. Since the Price variable is in dollars and the y-intercept will have the same units, it probably makes sense to round this to -30.67. The slope of the regression model is the coefficient next to the explanatory variable, in this case "Number of books". So the slope here is 1.4553. Since number of books is typically between 10 and 60, we may want to round off to three decimal places (1.455), so that after multiplication by a number of books the result is a dollar amount.

The final regression equation is then

$$\text{Price (of backpack in dollars)} = -30.67 + 1.455 * \text{Number of Books}.$$

Example 8.2. Interpreting Coefficients of a Regression Model
In the previous example, we developed a regression model from StatPro's regression output. But what does this model mean?

The y-intercept is usually pretty clear: it's the y-value when the x-variable is zero. So, if we were to market a backpack that couldn't hold any books (Number of Books = 0), we could expect that people would not pay any money for it. In fact, the equation predicts that we would have to pay the customer \$30.67 just to take the backpack away! After all, a backpack that doesn't hold any books isn't very useful. We'll get another interpretation of this number in the next example.

To interpret the slope, we need to use some proportional reasoning. Remember that slope is rise over run. "Run" in this case refers to the number of books the backpack will hold, while "rise" refers to the price of the backpack. Our model has slope = rise/run = (change in y)/(change in x) = 1.455. This is marginal analysis: to determine what happens to the value of the dependent variable when x changes by 1 unit. If we design an identical backpack, that can hold one more book, then the price of the backpack will increase by \$1.455. In fact, since linear models are proportional, a 2 unit increase in x will result in a 2 $\times$ \$1.455 = \$2.910 increase in the price.

Another way to see this is to analyze the units of the slope. Since slope is change in y over change in x, it must have the units "y units per x unit". Thus, our slope really means "1.455 dollars per book". For each additional book a backpack can hold, this model predicts a \$1.455 increase in the price of the backpack.

Example 8.3. Calculating the X-intercept (Solving a linear equation)
Now, the previous examples have a y-intercept that doesn't make much sense. After all, who would market a backpack that you have to pay people to use? But let's graph the equation of the model and see if it helps. We know the y-intercept is -30.67, so the line passes through the point (0, -30.67). It has a positive slope (1.455) so it is increasing; this means that the more books the backpack holds, the more it is worth.

Figure 8.1: Example of a straight line predicting the price of a backpack based on the number of books it holds.

Notice that the line passes through the x-axis. Thus, it has an x-intercept. In this case, the x-intercept appears to be around Number of Books = 20, but it's a little higher than

that. Finding the x-intercept will answer the following question: What would be the least number of books the backpack would have to hold in order to be marketable? Let's try to find the x-intercept exactly.

To do this, we take the regression equation, and we plug in everything we know is true about the x-intercept. Right now, we have a guess at its x coordinate, but that's not good enough. We do know one thing for certain, though: it has a y-coordinate of 0. So we can plug in "Price = 0" to our regression equation above to get

$$0 = \text{-}30.67 + 1.455*\text{Number of Books.}$$

Now, we want to use algebra to rearrange the equation to find how many books it takes to make the price zero (that's what the above equation really means). To solve the equation, we simply "undo" what has been done to the number of books. First, the number of books is multiplied by 1.455, then that result is added to -30.67. So, to undo it, we first subtract (-30.67) and then divide by 1.455. But if we do this to the right-hand side of the equation, we must do it to the left-hand side. Zero minus (-30.67) is just 30.67 Dividing this by 1.455, we get approximately 21.079. These steps are shown below.

$$\begin{aligned} 0 + 30.67 &= -30.67 + 1.455 * \text{Books} + 30.67 \\ 30.67 &= 1.455 * \text{Books} \\ \frac{30.67}{1.455} &= \frac{1.455 * \text{Books}}{1.455} \\ 21.079 &= \text{Books} \end{aligned}$$

This means that unless your backpack can hold at least 22 books (the nearest whole number greater than your x intercept of 21.079, since we cannot really have a part of a book), you cannot expect anyone to buy it.

8.1.3 Exploration 8A: Regression Modeling Practice

StateSteins is a tourist trade vendor. The company manufactures hand-crafted beer steins with state logos and images. They have facilities in each of the fifty states (and in Washington, DC). The data file "C08 Profit.xls" contains last year's figures for profit, revenue, cost, number of steins sold, and labor for each of the separate state facilities. Your job is to investigate these data and determine which variable is the best predictor for the company's profits.

1. Formulate and estimate linear regression models to predict profit as a function of each of the four explanatory variables. Be sure that you have the routine construct the important diagnostic graphs (Fitted v. Actual, Residuals v. Fitted) to help in your analysis. You will need to rename the worksheets with these graphs in order to ensure that StatPro will not overwrite them each time you compute a new regression model.

2. Interpret the slopes of each of the four regression models you created. Be sure to include the units for each slope.

3. Examine the diagnostic graphs to see what they tell you about the quality of your model. You will learn more about how to interpret these graphs in the next section, but for now, see what you can learn from them.

8.1.4 How To Guide

Simple Regression with StatPro

StatPro makes regression analysis fairly easy. Follow the steps below with any set of data to develop a regression model along with its associated measures and graphs.

1. Select the region of the worksheet that contains the data

2. Select the StatPro routine to apply to the data.

   The routine for simple regression is under "Regression/ Simple..." Note that there are other types of regression. In this text, we'll focus on "simple" and "multiple". "Stepwise" and "block" regression are similar to each other and are a modification of multiple regression. "Forward" and "backward" regression are related to time series analysis.

3. Verify that the data region is correct.

   At this point, the simple regression routine will warn you that it has not been designed to deal with certain problems in the data. Just click "OK" and move on.

4. Select the variables to which the routine will be applied.

   StatPro will give you two screens on which to select variables. The first screen allows you to select the dependent (response) variable. The second allows you to select the explanatory (independent) variable.

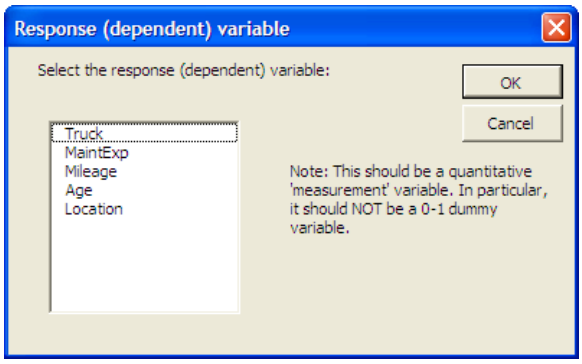

Figure 8.2: Selecting the variables in the regression routine of StatPro.

5. Fill in the details of the routine.

   For simple regression, the routine will ask you which diagnostic graphs you want produced (the dialog box is shown below). Each of these graphs will be discussed later in the text. Usually, though, you will only need the first two options ("Fitted versus actual" and "Residuals versus fitted"). These two scatterplots allow you to determine whether the model, in this case a linear model, is a good choice for your data. Explanations of these graphs appear in a later chapter; for now, we'll rely on some other tools.

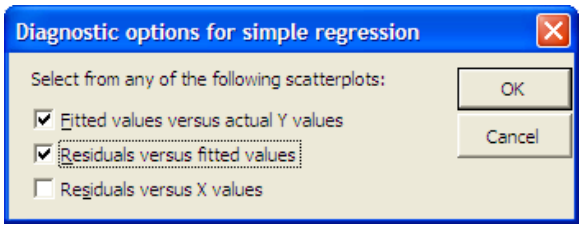

Figure 8.3: Diagnostic options that are useful when running regression.

6. Select the placement for the output of the routine.

   For regression output, we definitely suggest putting the output on its own sheet. Name the sheet so that you can tell which model it contains (in case there are several different explanatory variables to choose from).

   You will also notice that the simple regression routine will add two new columns of data to your worksheet. One column contains the predicted y-values (referred to as "fitted values") and the other column contains the residuals for each observation.

   Now that you have the regression coefficients, you can easily create Excel formulas to compute the predicted y-values, based on given x-values. This will let you explore the model's predictive power.

8.2 Using and Comparing the Usefulness of a Proportional Model

Now, it's one thing to have an equation to model data. We can always get regression equations for any data. However, the "best fit line" may not be a very good model for the data. We need a way to know not only the equation of the model, but also how good the model is. We will learn about two ways to measure "how good a model is". The first is a direct test for whether the two variables in the model are even linearly related. This is called the coefficient of determination (R^2) and is related to the correlation between the two variables. The second measure tells us how close predictions from our model will be to the actual data. This number is called standard error of estimate (S_e) and is sort of a standard deviation, indicating how spread out the data is from the model.

These two quantities relate to the entire regression model, reducing some characteristic "error" in the model down to single numbers. There are other ways to check on the quality of the regression model, however. StatPro (and most other statistical packages) provide diagnostic graphs for checking the regression model out. Two of the most important of these graphs are the graphs of the predicted values (also called fitted values) versus the actual response variable data and the graph of the residuals (the error in the model) versus the fitted values. A quick look at these two scatterplots can often tell you a lot about the quality of the model. Taken together with the coefficient of determination and the standard error of estimate, these are very powerful tools for determining the quality of the regression models you produce. After all, it is easy to simply point and click to produce more and more regression models; what is difficult is learning which ones are useful and to what extent they are useful.

8.2.1 Definitions and Formulas

Predicted values (fitted values) These are the predictions of the y-data from using the model equation and the values of the explanatory variables. They are denoted by the symbol $\hat{y}_i$.

Observed values These are the actual y-values from the data. They are denoted by the symbol y_i.

Residuals This is the part that is left over after you use the explanatory variables to predict the y-variable. Each observation has a residual that is not explained by the model equation. Residuals are denoted by e_i and are computed by

$$e_i = y_i - \hat{y}_i$$

Since these are computed from the y values, it should be clear that the residuals have the same units as the y, or response, variable.

Total Variation (Total Sum of Squares, SST) The total variation in a variable is the sum of the squares of the deviations from the mean. Thus, the total variation in y is

$$\text{SST} = \sum (y_i - \bar{y})^2$$

Unexplained variation (Sum of Squares of Residuals, SSR) The variation in y that is unexplained is the sum of the squares of the residuals:

$$\text{SSR} = \sum (y_i - \hat{y}_i)^2$$

Explained variation (Sum of Squares Explained, SSE) The total variation in y is composed of two parts: the part that can be explained by the model, and the part that cannot be explained by the model. The amount of variation that is explained is

$$\text{SSE} = \text{Total Variation} - \text{Unexplained Variation} = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$$

Regression Identity One will note that the Total Variation is equal to the sum of the Unexplained Variation and the Explained Variation.

$$\text{SST} = \text{SSR} + \text{SSE}$$

Coefficient of Determination (R^2) This is a measure of the "goodness of fit" for a regression equation. It is also referred to as R-squared (R^2) and for simple regression models it is the square of the correlation between the x- and y-variables. R^2 is really the percentage of the total variation in the y-variable that is explained by the x-variable. You can compute R^2 yourself with the formula

$$\begin{aligned} R^2 &= \frac{\text{Total Variation} - \text{Sum of Squares of Residuals}}{\text{Total Variation}} \\ R^2 &= \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \\ &= \frac{\text{SST} - \text{SSR}}{\text{SST}} = \frac{\text{SSE}}{\text{SST}} \end{aligned}$$

R^2 is always a number between 0 and 1. The closer to 1 the number is, the more confident you can be that the data really does follow a linear pattern. For data that falls exactly on a straight line, the residuals are all zero, so you are left with $R^2 = 1$.

Degrees of Freedom for a linear model The degrees of freedom for any calculation are the number of data points left over after you account for the fact that you are estimating certain quantities of the population based on the sample data. You start with one degree of freedom for each observation. Then you loose one for each population parameter you estimate. Thus, in the sample standard deviation, one degree of freedom is lost for estimating the mean. This leaves you with $n - 1$. For a linear model, we estimate the slope and y-intercept, so we loose two degrees of freedom, leaving $n - 2$.

Standard Error of Estimate (S_e) This is a measure of the accuracy of the model for making predictions. Essentially, it is the standard deviation of the residuals, except that there are two population parameters estimated in the model (the slope and y-intercept of the regression equation), so the number of degrees of freedom is $n-2$, rather than the normal $n-1$ for standard deviation.

$$S_e = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\text{SSR}}{n-2}}$$

The standard error of estimate can be interpreted as a standard deviation. This means that roughly 68% of the predictions will fall within one S_e of the actual data, 95% within two, and 99.7% within three. And since the standard error is basically the standard deviation of the residuals, it has the same units as the residuals, whcih are the same as the units of the response variable, y.

Fitted values vs. Actual values This is one of the most useful of the diagnostic graphs that most statistical packages produce when you perform regression. This graph plots the points $(y_i, \hat{y}_i)$. If the model is perfect ($R^2 = 1$) then you will have $y_1 = \hat{y}_1$, $y_2 = \hat{y}_2$, and so on, so that the graph will be a set of points on a perfectly straight line with a slope of 1 and a y-intercept of 0. The further the points on the fitted vs. actual graph are from a slope of 1, the worse the model is and the lower the value of R^2 for the model.

Residuals vs. Fitted values This graph is also useful in determining the quality of the model. It is a scatterplot of the points $(\hat{y}_i, e_i) = (\hat{y}_i, \hat{y}_i - y_i)$ and shows the errors (the residuals) in the model graphed against the predicted values. For a good model, this graph should show a random scattering of points that is normally distributed around zero. If you draw horizontal lines indicating one standard error from zero, two standard errors from zero and so forth, you should be able to get roughly 68% of the points in the first group, 95% in the first two groups, and so forth.

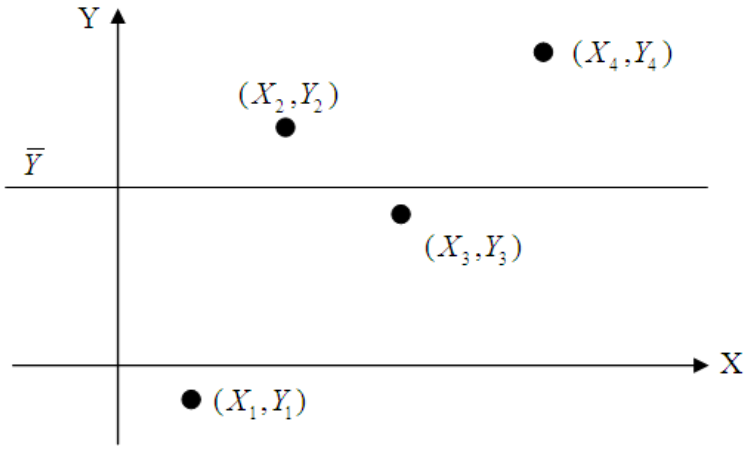

Figure 8.4: Sample initial data from which a regression line can be computed.

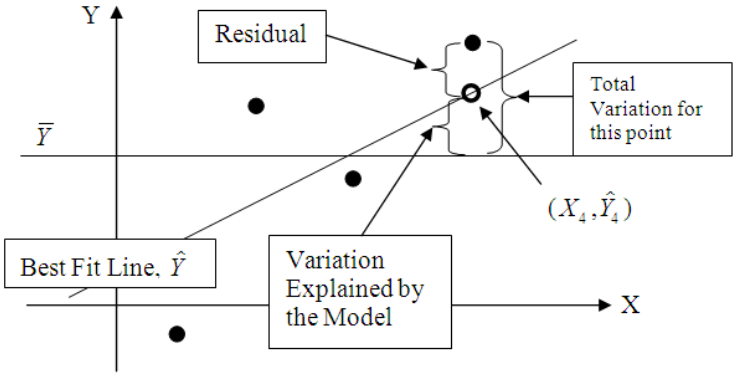

Figure 8.5: The various quantities involved in regression that are discussed above.

8.2.2 Worked Examples

Example 8.4. Interpreting the quality of a model
In the examples from section 8.1.2 (page 234), we developed and explored a model for predicting the price of a backpack based on the number of 5" by 7" books that it can hold inside. Using the data graphed in figure 8.6, StatPro produced the regression output below.

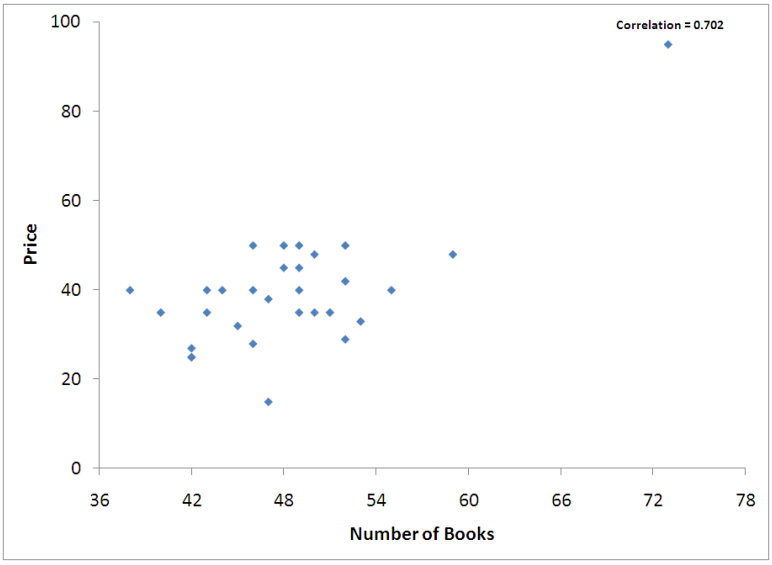

Figure 8.6: The various quantities involved in regression that are discussed above.

Results of simple regression for Price

Summary measures

R-Square	0.4931
StErr of Est	9.8456

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	2640.4857	2640.4857	27.2397	0.0000
Unexplained	28	2714.1810	96.9350		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-30.6751	13.5969	-2.2560	0.0321	-58.5272	-2.8231
Number of Books	1.4553	0.2788	5.2192	0.0000	0.8842	2.0265

Now we want to ask the question: "How good is this model at predicting the price of a backpack?" We'll start by examining the summary measures section of the output.

The R^2 value for this model is not terrible, but it is pretty low, being only 0.4931. This means that the number of books can only explain 49.31% of the total variation in price. That leaves over 50% of the variation in price unexplained. This could be for many reasons:

1. Some of the data could be entered incorrectly. In this case, look for outliers in the data and try double-checking all the data.

2. The number of books is not a good predictor for the price of a backpack. Try making simple regression models using the other explanatory variables to see if one of them does a better job.

3. There are additional variables that should be included in order to determine the price. Check out chapter 9 (page 257).

4. A linear model is not the best choice for modeling the price of a backpack as a function of the number of books it can carry. Check out chapter 11 (page 311).

Many times, the real reason is either #3 or #4. For example, our data and model do not include any variables to quantify the style of the backpack or its comfort when wearing it. Perhaps durability or materials are important variables. Maybe the name brand is important. Perhaps certain colors sell better. Perhaps certain extra features, additional pockets or straps for keys, are desired. Our data ignores these features. Essentially, our data tries to make all backpacks that are the same size cost the same amount of money. Clearly, this is not realistic. Given all these problems, though, an R^2 of 49% is not too bad. So, we might be able to convince ourselves that the model is useful.

The standard error of estimate for this model is about \$9.85. We can interpret this to mean that our equation will predict prices of backpacks to within \$9.85 about 68% of the time, or to within $2 \times \$9.85 = \$19.70$ about 95% of the time. That wide range in the predicted prices is probably due to other variables that are important (see point #3 above). Notice that S_e is always measured in the same units as the y-variable. This means that there is no "hard and fast" rule for what constitutes a good S_e. Our advice is to always compare S_e with the standard deviation of the response variable, since this tells us how accurate our simplest model, the mean, is for the data and we want to do better than that with our regression model. For these data, the standard deviation of Price is \$13.59. This means that if we just used the average price (\$39.67) as our model for backpack price, we would be less accurate than if we used the regression equation which at least accounts for the size of the backpack. In general, a good model will have S_e much less than the standard deviation of the y-variable.

Example 8.5. Computing R^2 from the ANOVA Table
What does the ANOVA table tell us? It actually tells us quite a lot, but for this text, we will only examine the first two columns of the ANOVA table. These are marked *df* and *SS*. These stand for **degrees of freedom** and **sum of squares**. Notice that the degrees of freedom that are explained is 1. This is the number of explanatory variables used in the model. *Degrees of freedom that are unexplained* is the number of observations, n, minus the explained degrees of freedom, minus one more for the y-intercept. Thus,

$$\text{Df (explained)} + \text{Df (unexplained)} + 1 = n$$

The sum of squares tells you how much of the total variation is explained and how much is unexplained. In this model, the amount of variation explained by "number of books" is given by SSE which is 2640.4857. The unexplained variation, SSR, is 2714.1810. This means that the total variation (SST) in y is 2640.4857 + 2714.1810 = 5354.6667. Thus, to calculate R^2, the fraction of the total variation in y that is explained by x, we simply compute a ratio:

$$R^2 = \frac{2640.4857}{2640.4957 + 2714.1810} = \frac{2640.4857}{5354.6667} \approx 0.4931.$$

Note that R^2 is also given by (SST - SSR)/SST. And since SST = SSE + SSR, we can rewrite this as ((SSR + SSE) - SSR)/(SSR + SSE) = SSE/(SSR + SSE). So you have several ways to estimate R^2 from the ANOVA table.

Example 8.6. Reading the Diagnostic Graphs
Let's examine the diagnostic graphs for the backpack model. There are two of these that are important. The first is the "Fitted versus Price". This graph is a scatterplot of all the fitted or predicted data ($\hat{y}_i$) versus the corresponding actual data (y_i). If the model predicted 100% of the variation in backpack prices ($R^2 = 1$) then each predicted value would equal the corresponding actual value, and the scatterplot would show a perfectly straight line of points with a slope of 1 and y-intercept of 0. The further the scatterplot is from such a line, the worse a fit the regression model is for the data. In this case, we get an interesting graph. The graphs in figure 8.7 show a rough trend like this, but there is a lot of spread around the "perfect line".

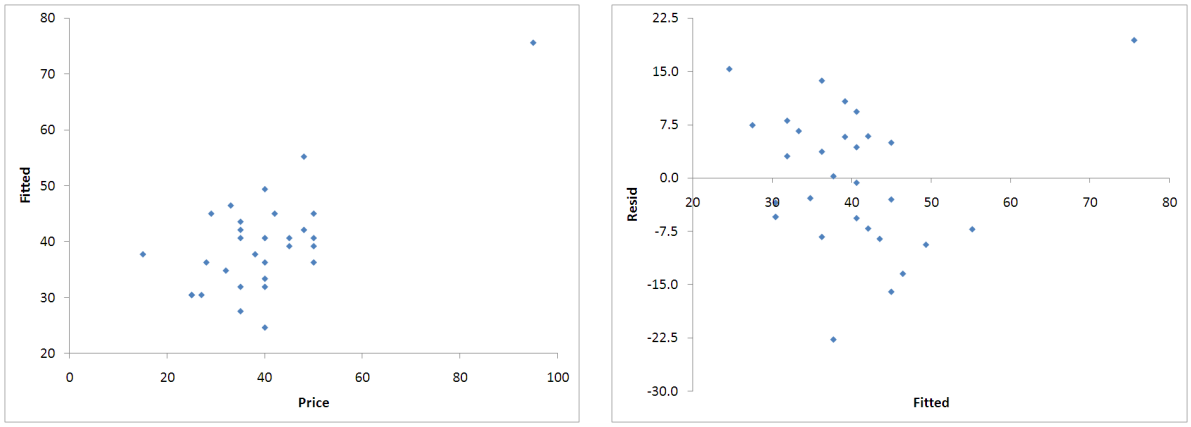

Figure 8.7: Diagnostic graphs for the backpack example. The graph at the left shows the fitted values versus the actual values. The graph at the right shows the residuals versus the fitted values.

RULE OF THUMB: When looking at the graph of fitted values versus actual values, there is no perfect explanation for how close to a straight line the graph needs to appear for the model to be "good." Notice the graph shown on the left in figure 8.7 does not really look much like a straight line. In fact, if it were not for the lone data point far to the upper right,

it might almost not look linear at all. The most important thing to do is not to make any absolute judgements about whether the model is good or bad; instead, focus on using the graph to explain potential problems with the model. Use the graph to describe the model's features. For example, in the graph above, we see that the model is not very accurate - for the most part, the data is randomly spread around the model results, with only one point really making the model behave linearly. This one data point's effects are called *leveraging* and this will be explored in the exploration.

The other diagnostic graph that is important is the graph of the residuals versus the fitted values. For this graph, we want to see the points randomly splattered around with no pattern. If there is a pattern to the points on this graph, we may need to try another kind of regression model, some sort of nonlinear equation. This will all be explained in more detail in later chapters, but for now, you should at least look at these graphs and try to determine whether they indicate that the model is a decent fit or not. We want the following to hold:

1. The fitted vs. actual graph should be close to a straight line with a slope of 1 and y-intercept of 0. This is because a perfect model would exactly predict the values of the response variable.

2. The residual plot should look like a bunch of randomly scattered points with no pattern. If there is a pattern, then we can alter the model to predict the pattern, producing a better model.

8.2.3 Exploration 8B: How Outliers Influence Regression

For this exploration we are going to investigate the relationship between the appraised values of homes and the actual sale prices of the homes. (All of the homes in the data file "C08 Homes.xls" were sold during a three-month period of time in the Rochester, NY region in 2000.)

First, construct a linear regression model for predicting the price of the home from the appraised value. Be sure that you have the routine construct the diagnostic graphs (Fitted vs. Actual and Residuals vs. Fitted). Also, make sure that you have the routine calculate the fitted values and residuals on the data worksheet. You will need all of these graphs and figures to explore the data.

1. What is the equation of your regression model? Is this model any good?

2. What does this model mean?

3. Now, on the residuals vs. fitted graph, draw horizontal lines (see the "How To Guide") to mark one standard error of estimate above and one below the horizontal axis (residuals = 0 along the axis). Draw similar lines to mark two, three and four standard errors. How many of the observations fall within 1, 2, 3, and 4 standard errors of the predicted values? What proportion of the observations fall in these ranges? (There are 275 observations total.)

4. Are there any outliers in the data - observations more than 4 standard errors from zero? How many are there? How do you think these outliers influence the quality of the regression model? How would the regression model change if you removed these outliers and re-ran the regression routine?

5. Next, we are going to identify the outliers and remove them from the data. To do this, we need to look at the actual data and sort it from smallest to largest residuals. Before we can do this, however, we need to delete the empty column between the data and the fitted values (this should appear in column N). To delete the column, place your cursor on the column header (N), right click, and select "Delete". Now, sort the data from smallest to largest residuals (see the "How To Guide" for section 5B.) Locate any observations with residuals more than 4 standard errors from zero. Delete these

observations from the data by deleting the rows the data are in (right click on the row, select "Delete"). Now, create a new regression model to predict the price of a home from its appraised value. What is the equation of this model?

6. Compare the two models, both their equations and their quality.

8.2.4 How To Guide

Linear Regression in Excel (without StatPro)

If you are ever trying to do linear regression and you do not have StatPro available, there are several options available. One would be to use the actual formulas for the regression coefficients and the summary measures to compute the slope, y-intercept, coefficient of determination and standard error of estimate. This would be a little tough. Fortunately, there are formulas already built in to Excel. The LINEST formula is used to ESTimate LINear trendlines. Let's use the backpack data from the examples and perform a linear regression on it to compute "Price" as a function of "Number of Books." This file has the x-data (number of books) in cells C2:C31 and the y-data in cells A2:A31. The LINEST formula has the following syntax:

```
=LINEST(known y values, known x values, const, stats)
```

Const refers to whether you want to calculate the y-intercept (the constant) from the regression (make it TRUE) or whether to force it to be zero (FALSE). We'll usually use TRUE.

Stats is another true/false variable. It indicates whether to calculate and output the summary measures. We'll almost always want it to be TRUE.

However, before you type in the formula, you should know that the output of it will have ten (10) pieces of information. Obviously, we can't put ten different numbers in a single cell, so we have to enter the formula as an array calculation.

First, highlight a block of cells that is two columns wide by five columns high.

Now, type the formula

```
=LINEST(A2:A31, C2:C31, TRUE, TRUE)
```

and hit CTRL+SHIFT+ENTER. (If you hit enter, you will only get the first of the ten numbers; then you have to start over!) The output will then appear in a 5 row by 2 column grid with the information shown below. The most important information is shown in bold.

Slope	**Y-intercept**	Regression coefficients
S_e for slope	S_e for y-intercept	
R^2 **for model**	S_e **for model**	Summary measures
F	Df	ANOVA information
SS (reg)	SS (resid)	

For more information about the LINEST function, type "regression" into the help system. If you check the "See also" portion of the help information, you will find out about the TREND function which helps you calculate other values, based on a set of known x and y values. There is also a separate SLOPE function which computes just the slope of the regression line. It has the syntax SLOPE(known y values, known x values). Used with INTERCEPT(known y values, known x values), you can get both coefficients.

How the Fill Handle Works to Complete a sequence of numbers

In Excel, you may have used the fill handle to copy a formula down a column or across a row. Remember, the fill handle is the little dot in the lower right corner of the active cell or active cell region. The fill handle can also be used to fill in patterns in a sequence of numbers that you enter.

For example, suppose you want to generate a column of numbers 10, 20, 30, 40, on up to 300. It would be tedious to type these by hand. Excel can help! Start by typing 10 in cell A1, 20 in cell A2 and 30 in cell A3. Now highlight the cells (A1:A3). Click and drag the fill handle all the way down the column until the little floating box that follows the cursor says "300". Release the mouse button and your list of numbers is filled in!

How does Excel know what you want? It uses simple linear regression to get the answer! When you highlight the first part of the pattern, it takes those numbers and treats them as the "known y values". Then, it assumes that the x-values are 1, 2, 3, 4... and fills in the linear regression.

Using the drawing tools in Excel (or Word)

Microsoft Office has many tools designed to help you add graphics to your work in order to enhance its appearance and improve productivity. The Insert ribbon has most of these drawing tools easily accessible. Most of the items that you might wish to draw - lines, arrows, circles, rectangles, etc. - are available from the "Shapes" menu. To draw lines, simply select the line tool. Once the line tool is selected simply click on the page where you want the first end-point of the line to be. Then drag the mouse across the screen and release it where you want the line to stop. To draw perfect horizontal or vertical lines, hold the shift key down while you sketch out the line; this will constrain the line to draw at forty-five degree angles.

The other drawing tools work very similarly to the line tool. For rectangles, holding the shift key while drawing will force the rectangle to be a perfect square. For ovals, holding the shift key will make them into perfect circles.

You can also add a textbox using that menu option. If you select this, and drag out an area on the drawing, you will create a region in which you can type text. This is useful for labeling your drawings and for pointing out important features of the drawings.

Once a drawing object is added to the page (or "canvas") you can make changes to it. Select the object and right click to see available options. You can layer objects, placing one in front of the other, by controlling the "order" of the objects. You can group objects together so that they can be moved as a whole. If you double-click (or right click and select "format drawing object") you can control the color of the lines, whether the object is filled in (and with what color) and its layout on the screen.

The best way to learn about the drawing tools is to experiment with using them.

8.3 Homework

8.3.1 Mechanics and Techniques Problems

8.1. Suppose you know statistics for X and Y shown below. You also know that the correlation of X to Y is 0.56. Use these to determine the equations of the least-squares best fit regression model to predict Y as a function of X. Produce a graph of this regression equation. Show all work.

Statistic	X-Variable	Y-Variable
Mean	15.27	107.93
Standard Deviation	7.82	38.77
First Quartile	5.3	47.1
Median	15.2	105.4
Third Quartile	22.6	160.3

8.2. The regression output below was developed from data relating the monthly usage of electricity (MonthlyUsage, measured in kilowatt-hours) to the size of homes (HomeSize, measured in square feet). One-variable statistics for each of these variables is also given below.

1. Use this information to write down the equation of the regression model. Explain what each part of the regression model means, paying particular attention to the unit of the coefficients in the regression equation.

2. Analyze the quality of the regression model you wrote down, based on the summary statistics in the regression output and the statistics on the X and Y variables.

3. Based on your regression model, what is the relationship between home size and monthly usage? Does this seem realistic? (Hint: What does the model predict for bigger and bigger homes? What about smaller homes? Are there any homes for which the model predicts a monthly usage of zero?)

Results of simple regression for Monthly Usage

Summary measures

Multiple R	0.9120
R-Square	0.8317
StErr of Est	133.4377

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	1	703957.1781	703957.1781	39.5357	0.0002
Unexplained	8	142444.9219	17805.6152		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	578.9277	166.9681	3.4673	0.0085	193.8984	963.9570
HomeSize	0.5403	0.0859	6.2877	0.0002	0.3421	0.7385

Summary measures for selected variables

	HomeSize	MonthlyUsage
Mean	1880.000	1594.700
Median	1775.000	1641.000
Standard deviation	517.623	306.667
Minimum	1290.000	1172.000
Maximum	2930.000	1956.000
Variance	267933.333	94044.678
First quartile	1502.500	1321.250
Third quartile	2167.500	1831.000
Interquartile range	665.000	509.750
Skewness	0.893	-0.308
Kurtosis	0.340	-1.565

8.3. Pie in the Sky, Inc. runs a chain of pizza eateries (See data file "C08 Pizza.xls".) The manager has collected data from each of the stores in the chain regarding the number of pizzas sold in one month, the average price of the pizzas, the amount the store spent on advertising that month, and the average disposable income of families in the area near the store.

1. The manager wants to know how these variables are related. Specifically, he wants to know which variable is the best to use for predicting the number of pizzas that a given store will sell in a month. Develop regression models to predict the quantity sold based on each variable in the data. Use the three models you develop to determine which variable is the most influential.

2. Use your best model to determine how many pizzas will be sold if a store has an average pizza cost of $11.00, spends $51,000 on advertising, and is in a region with an average disposable income of $40,000.

3. Based on the models that you developed, if a store wanted to sell 80,000 pizzas, what should the store do?

8.3.2 Application and Reasoning Problems

8.4. The file "C08 Hospitals.xls" contains hospital and physician data on a number of metropolitan areas. Use this data to develop a regression model for predicting the number of general hospitals that a metropolitan region can support based on the number of general physicians in the region. Once you have the model, use the residuals and the summary measures to identify the regions which are outliers (for these purposes, an outlier is more than four standard errors away from the fitted values). You should find a total of seven outliers in the data (including both too high and too low).

8.3.3 Memo Problem

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: Commuter Rail Analysis

Ms. Carrie Allover, the manager of the commuter rail transportation system of our fair city has contracted us to analyze how various factors affect the number of riders who use the rail system. Her Supervisory Board wants this information for long-range planning. Accordingly, she has sent along some data on the weekly ridership (number of people who use the train during a week) of commuters taking the train into the city, as well as some data on various factors thought to have an influence on the ridership. These data contain the following variables: Weekly riders, Price per ride, Population, Income, and Parking rate. The latter variable, Parking rate, refers to the cost of downtown parking.

To deal with Ms. Allover's requests, you will have to build several regression models with Weekly riders as the response variable, but before you proceed with this I want some common sense predictions on whether the coefficients of each of these explanatory variables will have a positive or negative sign; that is, whether the variable will have a positive or negative effect on the weekly ridership. Of course, you have to provide an explanation for your prediction. Some of these will be clear cut, but there may be a couple that are not so easy to predict and you won't know the answer until you actually run the model. But don't change your analysis if you prove to be wrong. Ms. Allover needs this kind of verbal, up-front analysis (whether right or wrong) so that she will be prepared to deal with possible responses, as well as misunderstandings, on the part of the Board.

After you have explained how you anticipate each variable will affect the number of weekly riders, go ahead and formulate the different models, one for each possible explanatory variable. Explain what each of the models means, using the coefficients in the regression output. In particular, describe how each explanatory variable actually affects the response variable, Weekly riders, including all appropriate units. This is extremely valuable information, Ms. Allover insists. Let's provide her with a brief analysis of how well the models fit the data as well as how accurate we can anticipate the predictions of the models will be.

Attachment: Data file "C08 Rail System.XLS"