Chapter 7

Correlation[1]

So far in this book, we have limited ourselves to looking at only one variable at a time, trying to learn as much as possible about that single variable. However, most of our data is made up of many variables, all interacting and having effects on each other. In this chapter you will explore relationships between two variables using graphical methods (scatterplots), computational methods (correlation), and algebraic methods (equations of functions).

- *As a result of this chapter, students will learn*
  - √ How to read and interpret a scatterplot
  - √ How correlation describes the relationship between two variables
  - √ The meanings of "positive" and "negative" relationships between two variables
  - √ About the slope and y-intercept of straight lines and how to compute these
- *As a result of this chapter, students will be able to*
  - √ Identify variables with a positive or negative relationship using the correlation coefficient
  - √ Construct a correlation table using StatPro to determine which variable relationships are most influential
  - √ Estimate the correlation coefficient of two variables based on a scatterplot
  - √ Set up a scatterplot according to conventions about axes, etc.
  - √ Add trendlines to a scatterplot

7.1 Picturing and Quantifying the Relationship Between Two Variables

In many of the previous examples in this book you have probably been tempted to go too far in your conclusions. For example, if you were to look at information about employees at a company and you learned that the salaries were negatively skewed and that the ages of your employees were also negatively skewed, you might be tempted to claim that one variable (for instance, age) influences the other variable (in this case, salary).

However, it would be dishonest to make such a claim with the tools we have discussed so far. In fact, the relationship between the two variables could be exactly the opposite of what you claim: it could be that the low salaries are all earned by employees who are older and that younger employees are making more money. It is even possible that the two variables are unrelated entirely. All of our tools up to now have been tools to analyze data one variable at a time. In order to speculate about relationships between two or more variables, we need new tools that include two variables at a time. A graphical tool for this analysis is the scatterplot. This is a two-dimensional graph made up of points where each point represents a pair of observations, one for each of the two variables you are comparing. In this way, you can quickly spot connections between variables. Such connections are called **correlations** and can also be computed numerically with a fairly simple formula based on z-scores.

Consider the employee salary example above. One could speculate that the points representing the salary and age of each employee would show that older employees tend to have higher salaries (after all, they have been working longer, have more experience and have had more opportunities for promotion). If the graph shows this, then there might be a connection between the two variables.

We want to emphasize this as strongly as possible. Simply because the correlation between two variables is high does not mean that one variable is causing the changes in the other. Consider the following situation: You are interested in the performance of your stock brokers at a large investment firm. If you looked at the amount of money each broker earned for the firm and compared this to the number of cups of coffee that broker drinks each day at work, what would it mean if there were a strong positive correlation? Would that mean that drinking more coffee makes you a better broker? Clearly, this is absurd. What it does mean is that brokers who make more money for the firm also tend to drink more coffee. That's all it means. Why might this be so? There are many reasons. It could simply be that the amount of coffee consumed is a surrogate for the number of hours the broker works. More hours worked might lead to more money for the broker. But more hours worked will probably involve drinking more coffee.

For the remainder of this book, we will be dealing with how to represent relationships among variables. Our goal is to develop these relationships into mathematical equations called **functions** that we can use in our decision-making.

7.1.1 Definitions and Formulas

Scatterplot A scatterplot is a graph that takes sets of observations of two variables and plots them as points on a graph. Each point corresponds to a single observation of both

variables. The points are identified by an ordered pair, with the horizontal variable listed first. These ordered pairs are written as (x, y). After each point in the data is plotted, the scatterplot can help determine if there is a relationship between the two variables.

Axis and axes All graphs have an axis that shows a scale and in which direction the variable being graphed is increasing. "Axes" is the plural form of the word axis.

Quadrants In a scatterplot, the horizontal and vertical axis cross at a point called the origin which has coordinates $(0, 0)$. This divides the Cartesian plane (all the possible points of the scatterplot) into four regions called quadrants. Each quadrant is numbered according to the graph in figure 7.1.

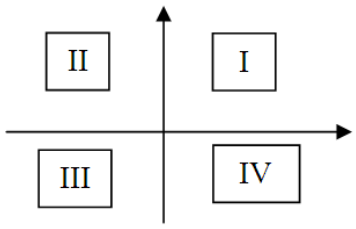

Figure 7.1: Diagram showing the labels for each of the four quadrants in an XY scatter plot. As usual, the x-axis runs left to right and the y-axis runs bottom to top.

Dependent Variable The dependent variable is usually graphed on the vertical axis. This is the variable that you suspect will be affected by a change in the other variable.

Independent Variable The independent variable is usually graphed on the horizontal axis. This is the variable that you suspect determines the value of the dependent variable. It is graphed on the horizontal axis because it is easier for the eye to scan left-to-right in picking a value for it and then scanning up the graph to determine the value of the dependent variable that corresponds to the value of the independent variable you picked.

Direct Relationship If the cloud of points on the scatterplot seems to move upward as the eye scans across the graph from left-to-right (as shown in figure 7.2), then the relationship between the two variables is said to be a direct relationship. This means that as the independent variable increases (gets larger in value), so does the dependent variable. Such a relationship is also referred to as a positive relationship or an increasing relationship. The graph in figure 7.2 shows a strong positive relationship between two variables.

Indirect Relationship If the cloud of points on the scatterplot seems to move downward as the eye scans across the graph from left-to-right (as shown in 7.3), then the relationship between the two variables is said to be an indirect relationship. This means that as the independent variable increases (gets larger in value), the dependent variable decreases.

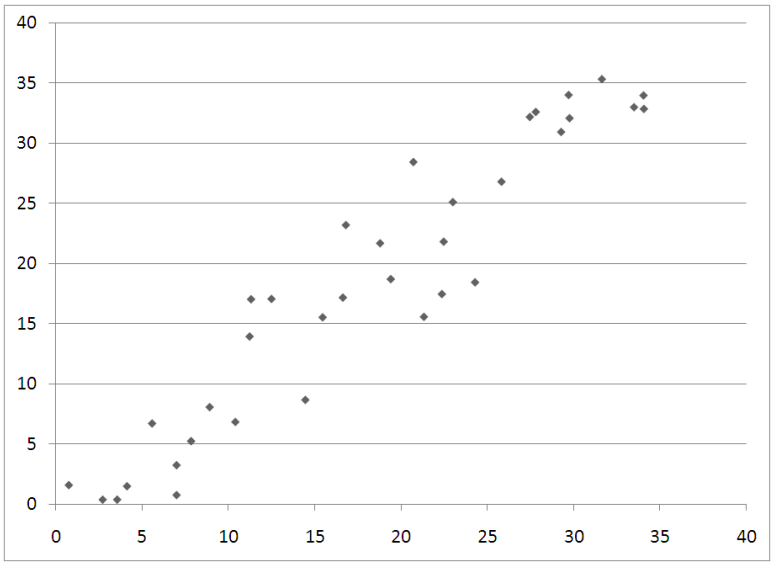

Figure 7.2: Illustration of a direct relationship between the dependent variable Y and the independent variable Y.

Such a relationship is also referred to as a negative relationship. The graph in figure 7.3 shows a strong negative relationship between the two variables graphed.

Correlation coefficient The correlation coefficient is a way of numerically determining two things:

1. Whether the relationship between two variables is direct, indirect or neither.
2. The strength of the linlear relationship between two variables.

Correlation is a number between -1 and +1 and is determined by the formula below, based on the z-scores of the two variables (the variables are called x and y in the formula).

$$\text{Correlation}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Notice that since this formula is based on the z-scores of the data, the overall correlation coefficient has no units. This makes it easier to interpret. Positive correlation means positive relationship, negative correlation means a negative relationship. Correlations close to +1 or -1 indicate strong relationships, while correlations close to zero indicate weak relationships, as shown in figure 7.4.

Correlation Matrix A correlation matrix (see table 7.1 for an example) shows the relationships among many variables at once in a table format. Each variable is listed twice - once along the top of the table and once along the side of the table. Each cell of the table contains the correlation between two variables (one from the row and one

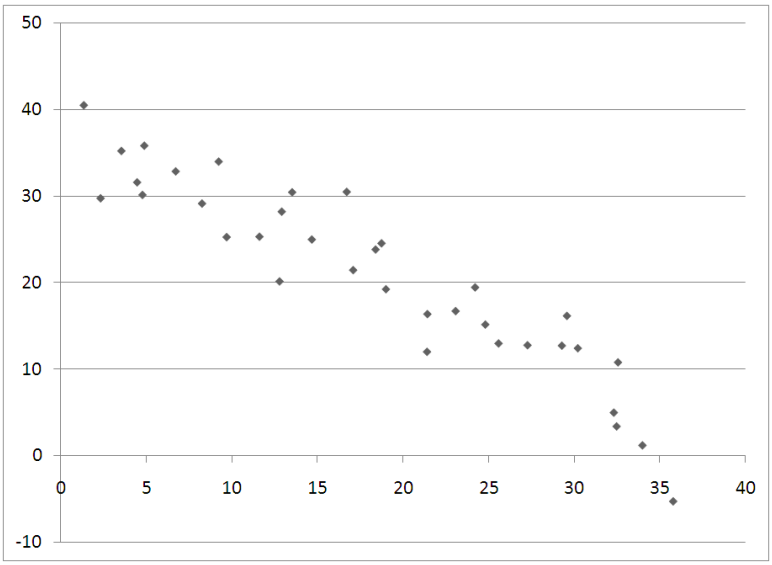

Figure 7.3: Illustration of an indirect relationship between the dependent variable Y, shown on the vertical axis as is standard, and the independent variable X on the horizontal axis.

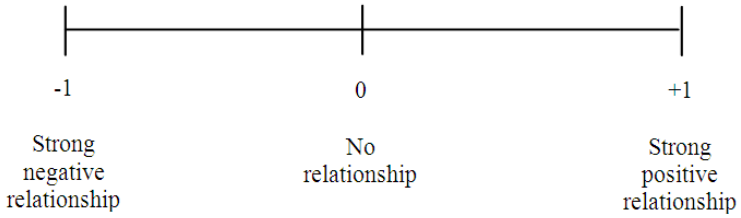

Figure 7.4: The scale of correlation, from -1 to +1.

from the column the cell is in). Usually such tables are only half filled in, since the correlation of x with y is the same as the correlation of y with x. Also, the diagonal entries are all +1, since a variable has a perfect correlation with itself.

Strong Relationship A strong relationship between two variables is seen in scatterplots with points that are tightly bunched together around some pattern (like a line or a curve). The graphs shown above under "Direct" and "Indirect" relationships are both strong relationships. Strong relationships have correlations close to +1 or -1.

Weak Relationship In a weak relationship, such as that shown in figure 7.5, there is almost no connection between the two variables. Figure 7.5 shows such a situation. This might result from graphing the two variables "grade on a test" and "amount of pizza consumed". Weak relationships have correlations close to zero.

Table of correlations	Age	Credits	WorkHours	SleepHours	GPA
Age	1.000				
Credits	0.221	1.000			
WorkHours	0.658	-0.439	1.000		
SleepHours	0.775	-0.886	-0.228	1.000	
GPA	0.342	0.669	-0.824	0.713	1.000

Table 7.1: Sample correlation matrix of relatinoships among the variables describing students at a large university.

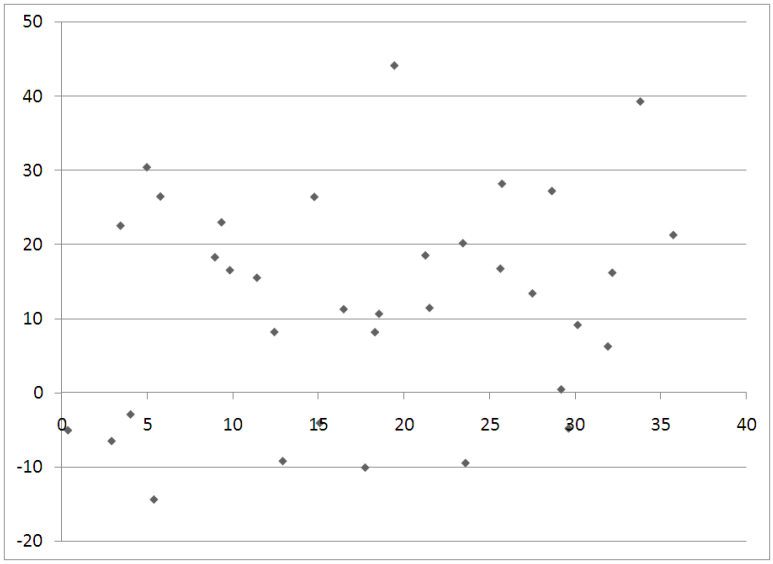

Figure 7.5: XY scatterplot showing a very weak relationship between the two variables.

7.1.2 Worked Examples

Example 7.1. Reading Variables and Relationships from a Graph
Suppose we have collected data on students taking the SAT shown in figure 7.6. If we have observations of the variables Study Time and Score, we might try to examine whether there is a relationship between the amount of time a particular student studies for the test and the score that this student receives on the test. We would then select Study Time as the independent variable, since we are guessing that study time predicts the test score. To create the scatterplot we then draw the axes and label them Study Time on the horizontal axis and SAT Score on the vertical axis. Next, we select a scale for each axis, based on the range for each variable. (Recall that the range is the difference in the maximum and minimum observations.) Finally, for each observation, we place a dot on the graph. The values of the two variables will determine where each dot is placed. For example, if one student studied 19 hours for the test and scored 741 (on a scale of 400-1600), the dot representing her score would be located along a line passing through the 19 hour mark on the horizontal axis, and

it would be lined up with the 741 mark on the vertical axis.

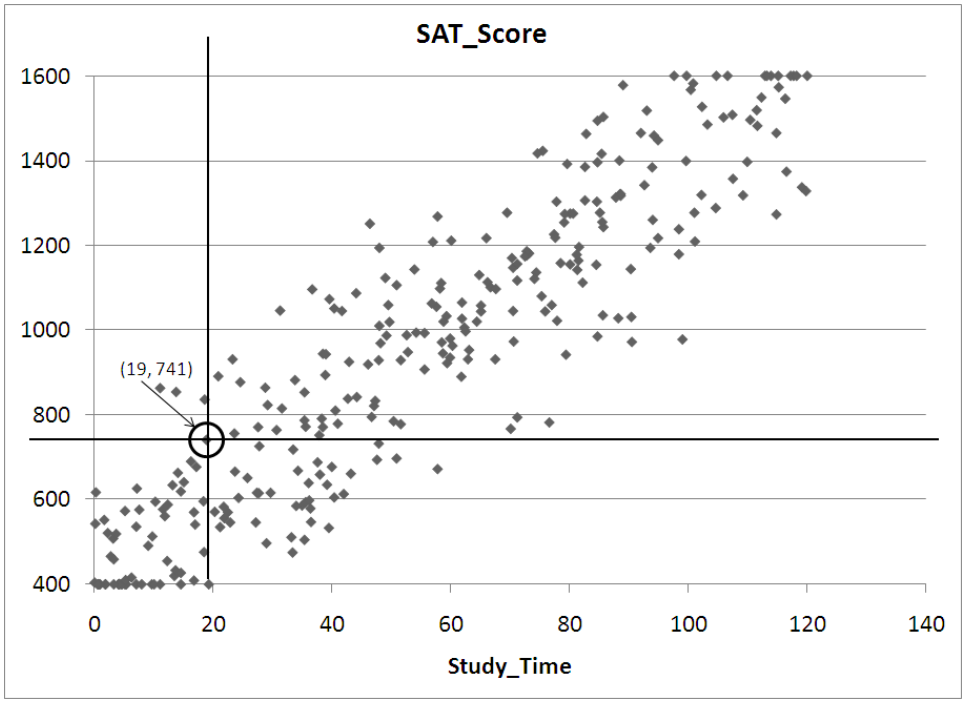

Figure 7.6: Scatterplot of SAT scores versus hours of study time.

After plotting all of the data on the graph above, it is clear that the variable Study Time has a strong influence on the final score a student receives on the SAT. The relationship looks quite strong and positive: as study time increases, students score higher on the test. Notice however, that the relationship is not perfect. There is a wide range of scores for students spending, for example, 20 hours studying for the test. In fact, all we can say for certain is that 20 hours of studying will probably get a score between 400 and 800 on the test. If we increase the amount of studying, though, the final score is quite likely to be higher. For example, 60 hours of studying seems to result in a score between 1000 and 1300.

Example 7.2. Reading a Correlation Matrix
Suppose we collect observations of several variables related to employees at Gamma Technologies: Age, Prior Experience (in years), Experience at Gamma (in years), Education (in years past high school), and Annual Salary. The matrix of correlations of such data might look like this:

Table of correlations	Age	Prior Experience	Gamma Experience	Education	Annual Sallary
Age	1.000				
Prior Experience	0.774	1.000			
Gamma Experience	0.871	0.443	1.000		
Education	0.490	0.362	0.308	1.000	
Annual Salary	0.909	0.669	0.818	0.650	1.000

To read the table, simply choose two variables and look up the intersection of those two variables in the table. If we choose Age and Gamma Experience, the correlation is 0.871. This number is quite high, indicating a strong positive relationship. Thus, we expect that older employees have been with the company longer. (This is not much of a discovery.) However, the strongest relationship between two variables in this study is between Age and Annual Salary. The correlation of 0.909 indicates that Age is an excellent indicator of salary: older employees make more money. Also, notice that the correlation between any variable and itself is always 1.000. You may also notice that the correlation of "Prior Experience" with Salary is slightly higher than the correlation of Education with salary. This means that this company places slightly more importance on experience over education. The last thing to notice is that part of the chart is blank. This is because the correlation of the variable Age to Prior Experience will be the same as the correlation between Prior Experience and Age. There is no need to duplicate the information.

Example 7.3. Strong and Weak Correlation Through Pictures
Note: Before reading this example, you may wish to review the material on z-scores in section 5.1 (page 134).

Consider the gas mileage for cars, a topic you may have spent some time thinking about recently. We have collected data on a sample of vehicles on the road in the file C07 AutoData.xls. The data include the gas mileage (measured in MPG or miles per gallon), the power of the engine (measured in horsepower) and the weight of the vehicle (measured in pounds). What general conclusions can we draw from the data, as represented in the graphs and charts below? As you can see from the graphs, all three variables are strongly correlated. However, two of the relationships are inverse relationships: As the weight of the vehicle increases, gas mileage decreases. As the power of the engine increases, the mileage also drops. However, the positive relationship shows us that larger cars (as measured by weight) tend to have more powerful engines (by horsepower). Three graphs illustrating various relationships among variables about automobiles in figures 7.7, 7.8, and 7.9.

Which of these relationships is the strongest? This is much harder to tell from the graphs. It appears that all three of the relationships have very similar correlations (in magnitude). To estimate the correlations, we need to know the means of the three variables.

Variable	MPG	Engine	Weight
Mean	31.50	90.84	2756.52

Now, we can draw in the means (this has been done in the above graphs) and use this to estimate the correlation between the variables in each graph. In the "Engine vs. Weight" graph, notice that most of the observations are in the upper-right and lower-left quadrants. This means that most of the observations will serve to increase the correlation coefficient. In the upper-right quadrant, $z_x > 0$ and $z_y > 0$ for each observation, so the product is also positive. In the lower-left quadrant, $z_x < 0$ and $z_y < 0$, so the product is also positive. However, there are a few observations in the upper-left quadrant which decrease the correlation (since the z_x scores of these observations is negative and the z_y scores are positive, this contributes a negative to the total correlation). There are quite a few observations in the lower-right quadrant which will also decrease the correlation ($z_x > 0$, but $z_y < 0$ for

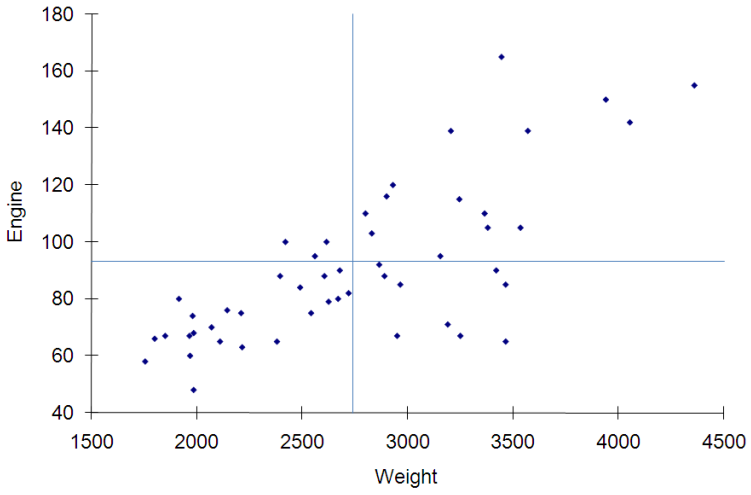

Figure 7.7: Engine power (in horsepower) versus car weight (pounds).

these). Based on this, we expect the correlation to be high and positive, but not perfect. A good estimate would be around 0.8.

Since the other graphs are similar in terms of spread, we expect their correlations to be the same magnitude as the first graph. Since they represent inverse relationships, though, these correlations must be negative. You could reasonably estimate the correlations to be about -0.8 for both graphs.

Example 7.4. How Correlation Works

Consider the data graphed on the scatterplot below. For each of the five data points, we can fill in the table below in order to estimate the effect of each point on the overall correlation of the data.

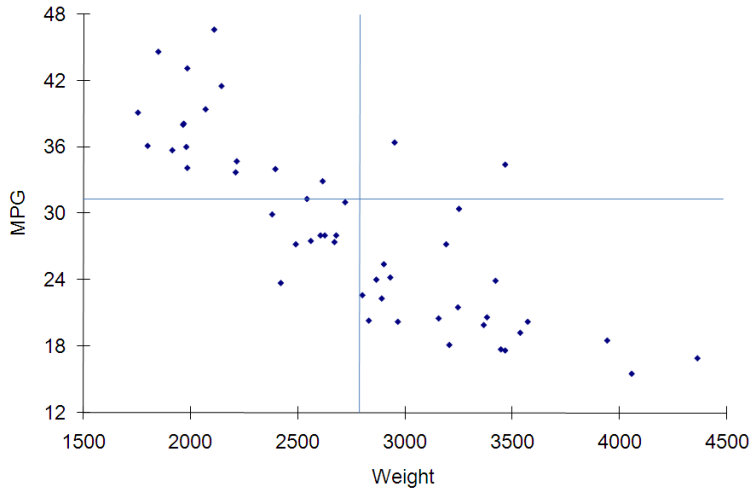

Figure 7.8: Gas mileage (miles per gallon) versus car weight (pounds).

POINT	Sign of Z score of Point's x	Sign of Z Score of Point's y	Sign of the products of the Z Scores	Increase or Decrease Correlation	Size of effect on correlation: No effect, a little, or a lot
A	Negative	Positive	Negative	Decreases	A lot
B	Negative	Negative	Positive	Increases	A little
C	Positive	Negative	Negative	Decreases	A little
D	Positive	Negative	Negative	Decreases	No effect
E	Positive	Negative	Negative	Decreases	A lot

So we see that four fo the five points contribute to a negative correlation, while one (B) increases the correlation. Point D has almost no effect on the correlation because the y-coordinate of D is almost equal to $\bar{y}$, makings its z-score basically zero. Overall, these data indicate a correlation of maybe 0.7 or so.

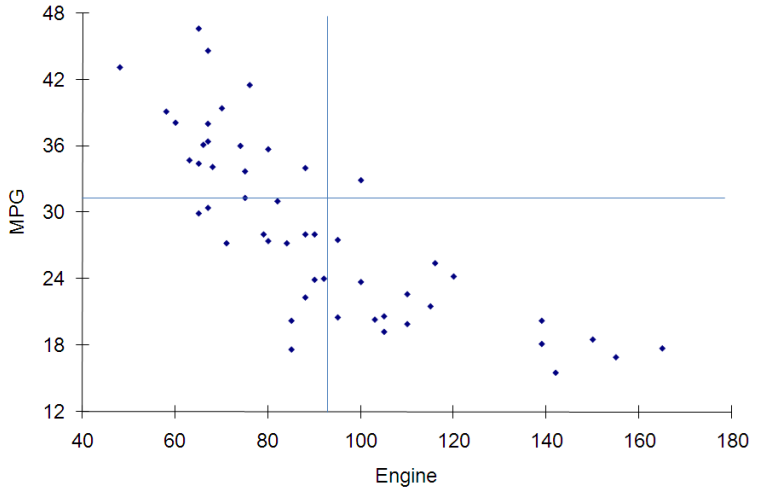

Figure 7.9: Gas mileage (miles per gallon) versus engine power (hp).

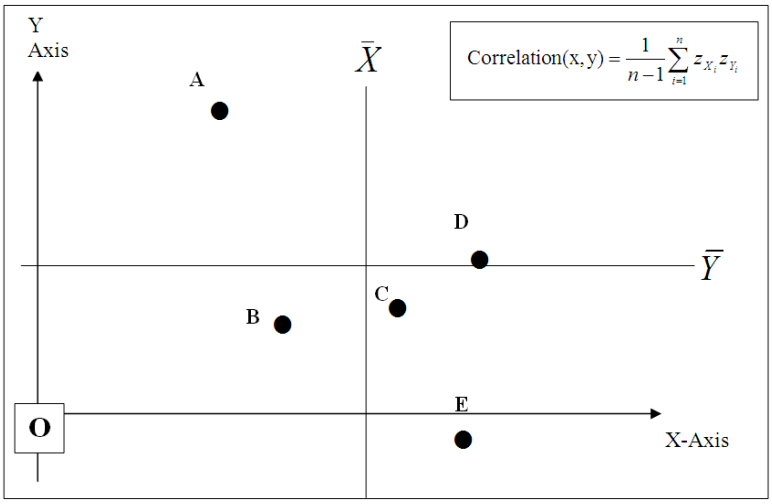

Figure 7.10: Scatterplot of points with means of X and Y shown.

7.1.3 Exploration 7A: Predicting the Price of a Home

Instructions: Using data file C02 HOMES.XLS, answer each question below.

1. Compute the mean and standard deviation for each of the following numerical variables:

	Taxes	Year	Acres	Size	Value	Price
Mean						
Standard deviation						

2. Using the mean as a model, how much would you say the *typical* single-family home costs in this market?

3. How reliable is your estimate?

4. Using a table of correlations, calculate the correlation coefficient (r) for the following pairs of variables:

	Taxes	Style	Bath	Bed	Rooms	Year	Acres	Size	Value
Price									

5. Based on the correlation coefficients, which of the above variables seems to have the MOST effect on the PRICE of a house? Which as the LEAST effect?

6. Generate a scatterplot that describes the relationship between PRICE and SIZE. Which variable is the independent variable (should be on the x-axis)? Which variable is the dependent variable (on the y-axis)? What does Excel report as the Correlation for this relationship? Your scatterplot should look something like figure 7.11.

7. Draw a vertical line on the above chart to represent the MEAN for SIZE

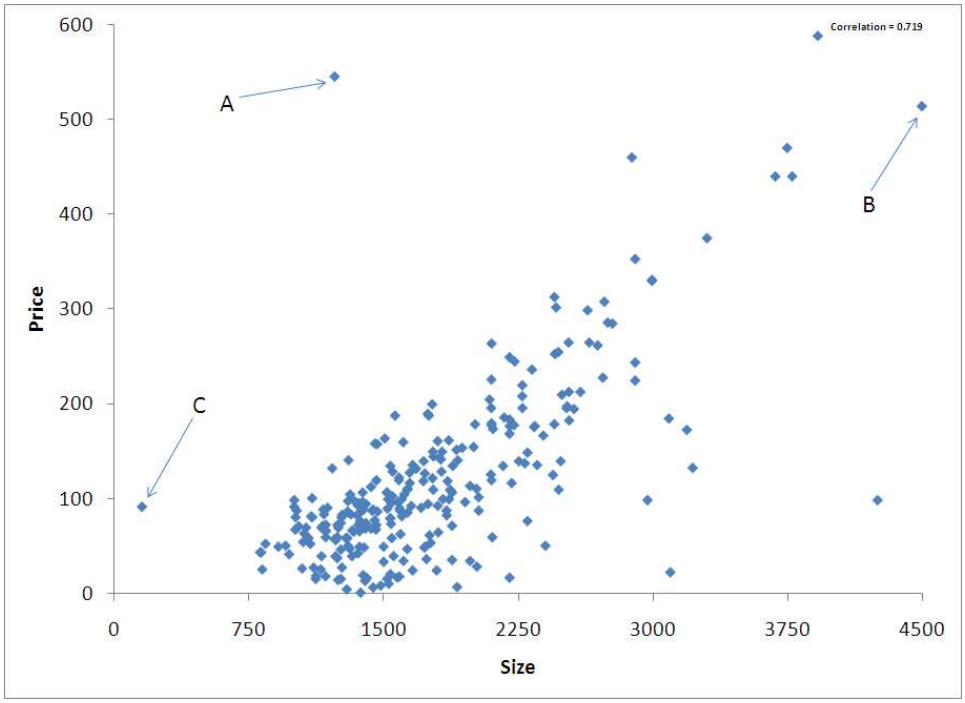

Figure 7.11: Scatterplot showing home price versus size.

8. Draw a horizontal line on the above chart to represent the MEAN for PRICE.

9. In Excel, hover your mouse over the points marked A, B and C on the chart to determine the values for PRICE and SIZE at each point. Then fill in the table below to estimate the correlation.

	SIZE	PRICE	z-score for SIZE (X)	z-score for PRICE (Y)	Total Contribution to the Numerator of Correlation
A			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	
B			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	
C			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	

7.1.4 How To Guide

Scatterplots with StatPro

StatPro makes scatterplots very easy. The procedure follows the same basic steps as all other StatPro routines:

1. Select the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data. This is located under "Charts/ Scatterplots".
3. Verify that the data region is correct.
4. Select the variables to which to apply apply the routine. For scatterplots, you can select as many variables as you like (hold down the control key "CTRL" and use the mouse to select several variables). Every possible combination of the variables selected will be graphed. This means that if you select two variables, you get one graph. Three variables: three graphs. Four variables: six graphs. If you selected ten variables, you would get forty-five graphs!
5. Fill in the details of the routine. For each possible combination of variables, StatPro will ask you which one you want to be on the vertical axis. Be sure to pick the proper variable so that the graph will show the relationship you are interested in seeing. The dialog box for this is shown in figure 7.12.

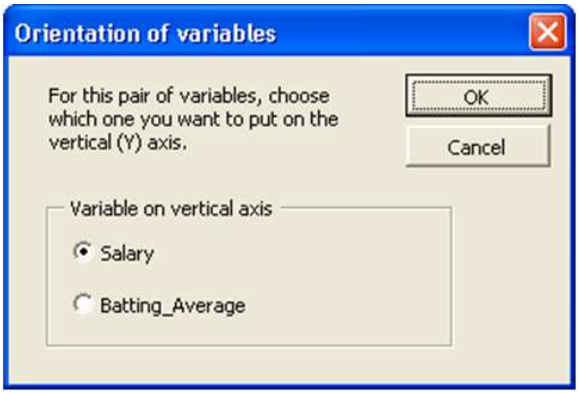

Figure 7.12: Selecting variables for making a scatterplot in StatPro.

6. Select the placement for the output of the routine. StatPro automatically puts each graph on a separate worksheet in the workbook. The sheet will be named in the form "Y-variable VS X-variable" with the names of your two variables filled in. The best part about using StatPro to make the scatterplots is that you can select any variables for the independent and dependent variables. If you use Excel's normal graphing tools, you are limited in your options as to how the graph can be made (unless you go to a lot of trouble).

Scatter Plots in EXCEL without StatPro

First, select the data you want. For Excel, this means that you must highlight all the data (and the variable names at the tops of the columns) that you want to graph. If the two variables are not right next to each other, highlight the first column of data, then hold down the control key (CTRL) and highlight the second column of data. Click the "Insert" ribbon and select *scatter* from the list of plot types. Then select the subtype of graph that you want to create. See figure 7.13.

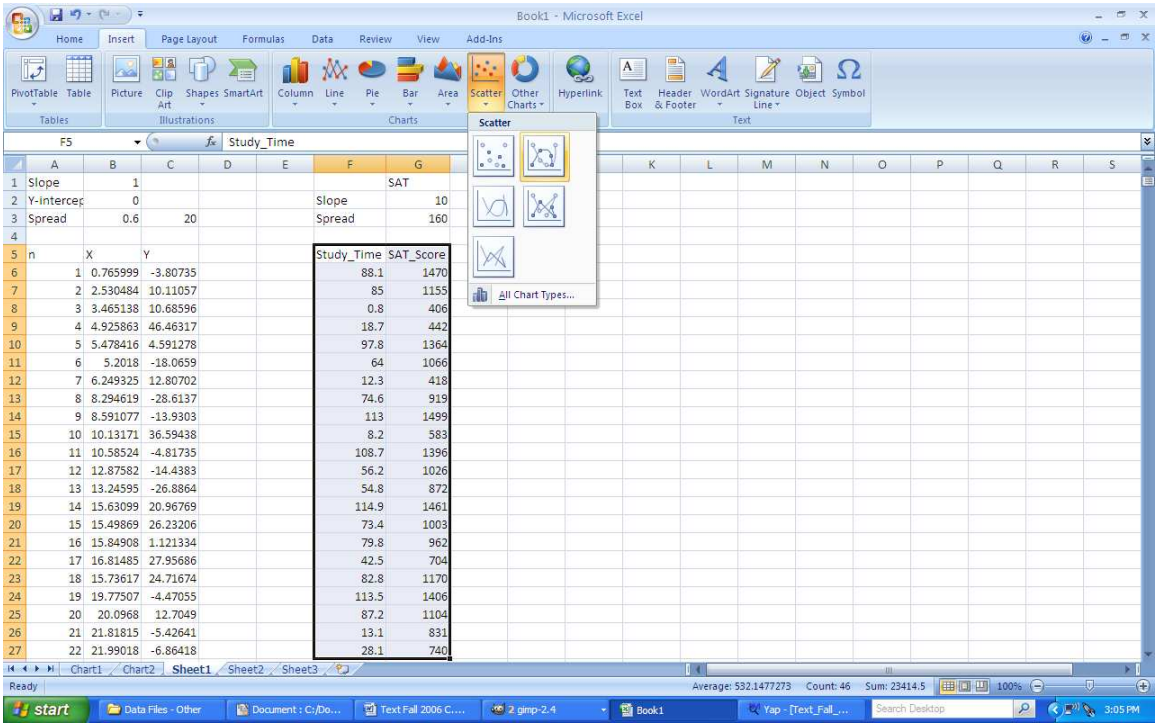

Figure 7.13: Screen image showing the selection of data (highlighted) and inserted a scatter plot.

Note that when making scatterplots in Excel without the use of StatPro, Excel will assume that the left column variable is the independent variable and the right-column variable is the dependent variable. To change this, you will have to first make the graph, then select the graph, and then click "Select Data" from the "Graph/Design" ribbon.

Moving a Graph

Any chart or graph, whether created in Excel or with StatPro, can be moved to either be a chart in an existing worksheet or a separate worksheet page by itself. To do this, click on the chart and select "Move Chart" from the "Graph/Design" ribbon. Then select the option you want from the dialog box shown in figure 7.14

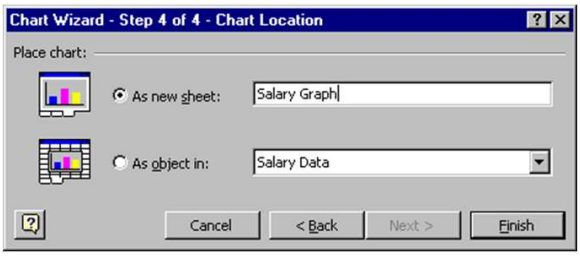

Figure 7.14: Dialog box for moving a chart to a new location.

Correlation Tables using StatPro

To create a table of correlations, follow the usual steps in activating a StatPro routine. In step 2, select "StatPro/Summary Stats/Correlations and Covariances". In step 5, you will see a screen like the in figure 7.15. Usually, you can just click "OK" at this dialog box to move on, because none of these settings will need to be changed. We suggest placing the resulting calculations on a new worksheet and naming the worksheet something like "Correlations."

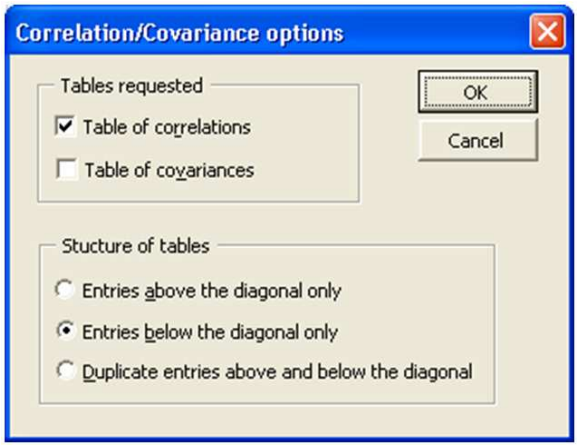

Figure 7.15: Creating a correlation matrix in StatPro.

Covariances are related to correlations, but are much harder to interpret since they have units and may range in size from negative infinity to positive infinity. The second set of options, "Structure of tables" doesn't really matter; it simply allows you to calculate the correlations above the diagonal, below the diagonal, or both. But remember, the values are the same above and below the diagonal.

Correlation with Excel

To calculate the correlation between two variables (with the same number of observations) using Excel, type

```
=CORREL(X values, Y values)
```

With this formula, it is critically important that you have the same number of observations of both variables, or you will get an error message.

7.2 Fitting a Line to Data

The easiest relationship between two variables to model is a linear relationship. Straight lines are easy to picture, they have simple equations, and each part of a straight line equation can be easily interpreted into real-world terms. Consider the data shown in figure 7.16. The independent variable is the size of a home in hundreds of square feet and the dependent variable is the price of the home in thousands of dollars. The data were taken from a sample of fifteen homes in a single neighborhood that all sold within one year. The graph clearly indicates a strong linear relationship between the two variables: larger homes tend to have higher prices.

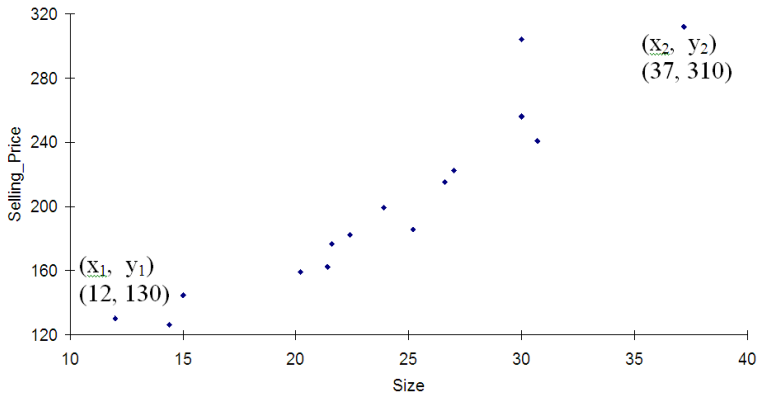

Figure 7.16: Plot of home selling prices (dollars) versus home size (square feet).

We can easily draw a straight line through this data that does a reasonable job representing the data. But what do we really mean by "representing the data"? Clearly we cannot draw a single straight line which passes through all of the data points. How then do we decide what the best line is? Each line is characterized by two numbers, **slope** and **y-intercept**. By carefully choosing these numbers we can make the line fit the data better. But how? Slope is basically the tilt of the line: larger values make the line more tilted, positive values tilt up, and negative values tilt down. The line for this data must have a positive slope. Furthermore, since the two extreme data points are about (37, 310) and (12, 130) we see that an increase in size of $37 - 12 = 25$ hundred square feet results in a price increase of $310 - 130 = 180$ thousand dollars. Thus, the slope of the line is approximately $180/25 = 7.2$ thousand dollars per hundred square feet of size.

Now that we have an estimate of the slope for this line, we can compute the y-intercept. Since the equation of the line is $y = A + Bx$ where B is the slope we just found (7.2), and since the line must pass through one of the points we used, we can plug all the known information into the equation and use algebra to find the value of A that makes the line with that slope pass through that point. So, we have $310 = A + 7.2 * 37$. We want to solve for the unknown A. We find that $A = 43.6$. Thus, we might estimate the line as $y = 43.6 + 7.2x$.

In this section, we will explore the equations of straight lines and use them to model relationships between two variables. We will also see how these equations can be used to make predictions about data that is not part of the data set. This involves specifying a value

of the independent variable and calculating the dependent variable from the equation. We will also see how to determine values of the independent variable that give rise to specified values of the dependent variable. This is usually referred to as "solving an equation."

7.2.1 Definitions and Formulas

Slope The slope of a straight line is a number that tells you exactly how much the dependent variable will increase for a given increase in the independent variable. Usually it is represented as a decimal number or a fraction and it is calculated from looking at the "rise" of the straight line between two points (this is the vertical distance between them) and comparing this to the "run" (the horizontal distance separating the two points). If the two points are labeled (x_1, y_1) and (x_2, y_2) then the slope is the change in y divided by the change in x. (Note that the Greek symbol delta, Δ, represents the phrase "change in".)

$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Y-intercept The y-intercept is the position on the vertical axis (possibly not shown on the graph) where a straight line crosses.

Equation of a straight line The most common way to represent the equation of a straight line is in **slope-intercept form**:

$$y = A + Bx.$$

In this equation, A is the y-intercept and B is the slope. The two other letters represent the variables: x is the independent variable, y is the dependent variable.

The equation can also be represented in **point-slope form**:

$$y - y_1 = B(x - x_1).$$

where B is again the slope and (x_1, y_1) is a point on the line. Both forms are equivalent; they are simply written in a different form to make it easier to use one or the other, depending on which two pieces of information you have. For example, if you re-arrange the point-slope form, you can produce $y = Bx + (y_1 - Bx_1)$, showing that the y-intercept $A = y_1 - Bx_1$.

Trendline A trendline is a line drawn on a graph to represent the relationship between two variables. These trendlines can take many forms. In Excel, there are five basic trendline options: linear, exponential, logarithmic, power, and polynomial. Trendlines are also called lines of best fit, even though trendlines are not always straight lines. Perhaps they should be called curves of best fit or trendcurves?

Linear relationship A linear relationship between two variables is characterized by a constant slope. A scatterplot of the two variables looks like a straight line. The graph in figure 7.17 shows a linear relationship, a linear trendline for it, and the slope and y-intercept of that trendline.

Function A relationship between two variables (called the independent and dependent variables) in which every value of the independent variable is associated with one and only one value of the dependent variable. Functions can be represented graphically (as lines or curves on a set of axes), as a table showing sample values, by an equation, or by a verbal description in words. On a graph, the test of whether a relationship is represented with a function is called the vertical line test and consists of drawing vertical lines on the graph. If any line crosses the graph more than once, the relationship is not a function.

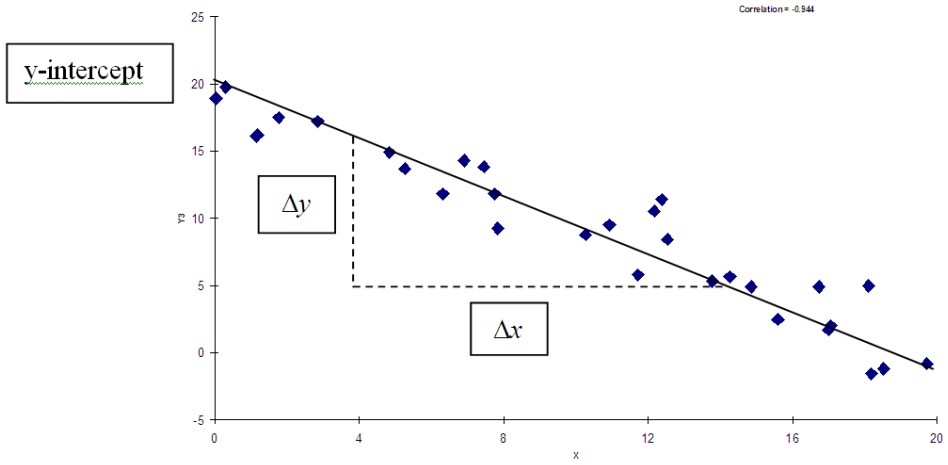

Figure 7.17: Sample linear relationship showing y-intercept and slope.

7.2.2 Worked Examples

Example 7.5. Estimating slope and y-intercept from a scatterplot
In the graph above (figure 7.17) we can easily make estimates of the slope and y-intercept of the trendline and use these to write down its equation. This equation could then be used to make predictions of other values.

The y-intercept appears to be about 21. It might be a little smaller, but clearly the trendline hits the y-axis above the tick mark for 20.

The slope is a little harder. We need two points on the line. Fortunately, this trendline seems to pass through several of the points on the scatterplot. (This is not always the case. The procedure for finding trendlines does not guarantee that the trendline will pass through any of the data points.) This line seems to pass through the points $(3, 17)$ and $(14, 5)$. Thus, when the run is $(14 - 3) = 11$ the line has a rise of $(5 - 17) = -12$ (notice the negative

sign; it means that the relationship is indirect or decreasing). Thus, the slope of the line is approximately rise over run $= -12/11$ which is about -1.091.

Putting this together, we get the equation of the line to be $y = 21 - 1.091x$.

Example 7.6. Using data to find the equation of a line
Suppose we have data that consists of only two points. This means that we have two ordered pairs: one for each point. The ordered pair is another way to give data. Rather than listing the variables in columns, as we have done it in EXCEL, we list the data like this: $(1, 2)$ and $(3, 6)$. These ordered pairs are given so that the first number is the value of the independent variable that is associated with the number after the comma, the dependent variable. For example, in the ordered pair $(1, 2)$, the 1 is the independent variable that gives 2 for the dependent variable. The ordered pairs listed above would be identical to the table below:

X	Y
1	2
3	6

How many straight lines are there that are a "best fit" to the data above? Do you think this will be true for any two data points? If you play around with this for a little while, you'll discover that only one line can be drawn that passes through both points. What would the slope of this "best fit" be for the two point data set listed above? What about the y-intercept?

If we use the formulas above, the slope should be (6-2)/(3-1) = 4/2 = 2. This means that for every one unit we move to the right along this line, we also move two units up. Finding the y-intercept is a little trickier. Let's use the slope-intercept form of the equation of a line. We already know the slope, so the equation must be $y = A + 2x$. To find A, just remember that we also know the point (1,2) is on the line, so $2 = A + 2(1)$. If we work with this expression, we find that $2 = A + 2$, and the only number A which works in this equation is 0, so the y-intercept must be 0. This means that the equation of the line is $y = 2x$.

Note that we could also use the point $(3, 6)$ to find the y-intercept, A. We should get the same equation for the line using either of the two points.

Example 7.7. Calculating Values from Trendlines (Making Predictions)
In August 1997 Consumer Reports printed an article on different makes of backpacks. They measured three variables for each backpack: average price, total volume (in cubic inches), and the number of standard 5" by 7" books it could hold. A sample of the data is shown in table 7.2. (The full data set C07 Backpacks.xls includes 30 different backpacks.)

After plotting the price of the data versus the number of books the bags hold, Excel computes the following trendline (constants have been rounded to two decimal places):

Price = -30.68 + 1.46*Number of Books

The equation tells us that we can expect the price of a backpack to increase about $1.46 for each additional 5" x 7" book it holds. Thus, if a backpack were designed to hold 60 books, we could expect the price to be about

Price	Volume	Number of Books
48	2200	59
45	1670	49
50	2200	48
42	1700	52
29	1875	52
50	1500	49
35	1950	49

Table 7.2: Data on backpacks from *Consumer Reports*.

Price = -30.68 + 1.46*(60) = $56.92.

We can also ask the question another way: How many 5" x 7" books would you expect to fit into a backpack that you paid $45 for? To deal with this question, we can either set it up in a spreadsheet and try using GOAL SEEK to find the answer (see the how to guide for this section), or we can solve it with a little algebra:

$45 = -30.68 + 1.46*Number of Books
$45 + $30.68 = 1.46*Number of Books
$75.68 = 1.46*Number of Books
Number of Books = 75.68/1.46 = 51.84 which is about 52 books.

7.2.3 Exploration 7B: Adding Trendlines

Part I. Using data file C02 HOMES.XLS, answer each question below.

1. Create a scatterplot of SIZE and PRICE, as you did in the earlier exploration in this chapter. Add a trendline to it. Sketch the trendline here.

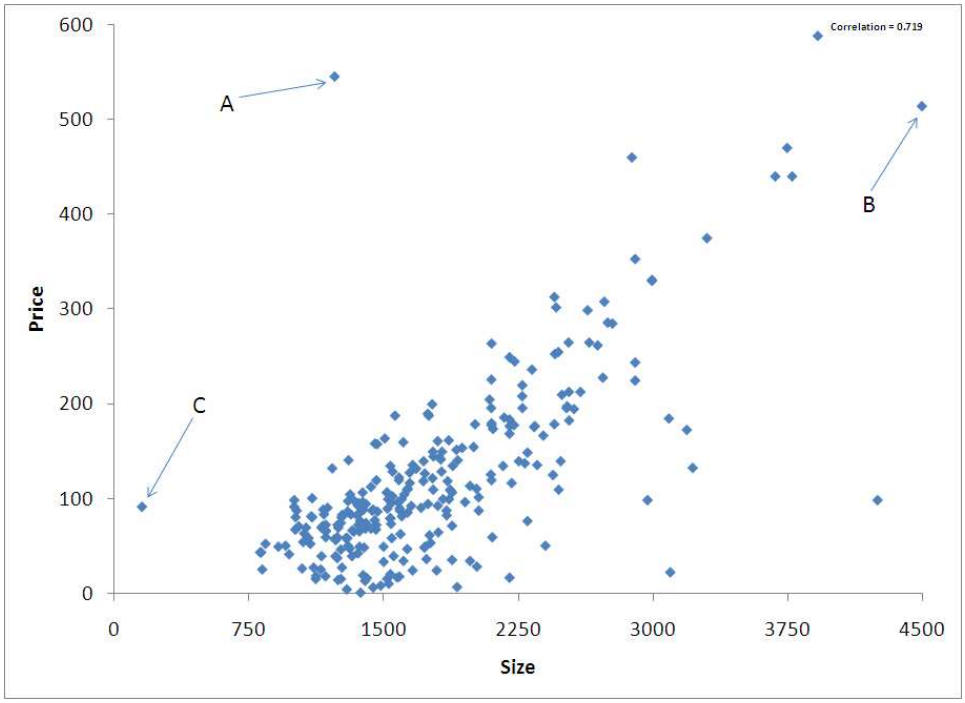

Figure 7.18: Scatterplot showing home price versus size.

2. What is the correlation coefficient (r) for this relationship?
3. What is the equation of the best-fit trendline for this relationship?

Part II. Create a new scatterplot between PRICE and TAXES. Be sure to do the following:

- Modify the font size
- Add units to both axis labels
- State the trendline equation in terms of the Model Variables
- Add a trendline

1. What is the correlation coefficient (r) for this relationship?
2. What is the equation of the best-fit trendline for this relationship?

7.2.4 How To Guide

Adding Trend Lines to a Scatter Plot

Now we will use EXCEL's capabilities to explore the relationship between the two variables by creating a "Trend line".

1. Position your pointer over one of the points on the scatter plot and right-click your mouse. Select "Add Trendline..." from the menu that appears.

2. You will now have a window (see figure 7.19) that shows several different types of functions that EXCEL can graph on top of your data. Let's select "Linear", which is the default choice. Don't click on "OK" yet, as we have some options to set in order to really take advantage of EXCEL.

3. Make sure you select "Display Equation on Chart" and "Display R-squared value on chart". This will help us in the future.

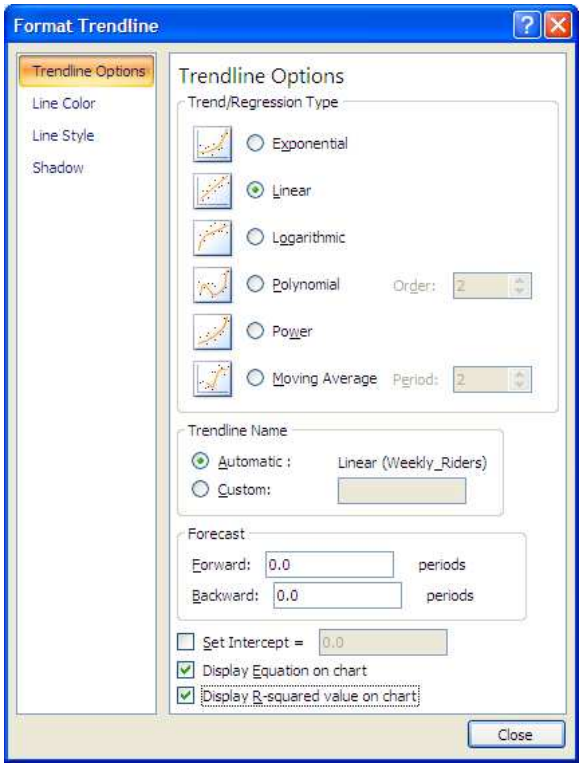

Figure 7.19: Trendline tool in Excel (access by right-clicking on a data point, and selecting "Add trendline...").

4. When you have finished setting the options, click on "Close". You should now see your scatter plot with two new things added. One of these will be a solid line, the other will be a little text box that displays the equation of the line and the R-squared value. For right now, think of R-squared as a measure of how closely the line resembles the data.

The closer this number is to "1", the better the line describes the data. We can also get R^2 values and equations for the other types of trend lines that EXCEL will graph.

5. Try out some other trend lines with this graph. EXCEL can put as many onto the same graph as you want. Simply repeat all the steps above for making a trend line, except choose a different function in step 2 each time.

A note about the other shapes for trendlines: In later chapters, we'll explore the other types of trendlines and what they are good for. For right now, just be aware that straight lines aren't the only option. Also, as we'll find out, some trendlines simply can't be used with certain data. If this is the case, Excel will automatically "grey out" those choices from the list.

A note about the Polynomial choice for trend lines: Polynomials come in different degrees. You can control the degree of the polynomial that Excel uses by adjusting the number in the box next to the polynomial trendline. Excel allows degree 2 through 6 polynomials.

Entering an Algebraic Model into EXCEL

In order to take a mathematical model and input it into Excel to make predictions, we need to complete three steps: Enter values for the parameters, create the data table for the x and y variables, and plot the data. These steps are outlined below for the linear model

$$y = A + Bx.$$

1. Enter the parameters. Notice that our model has two parameters, A and B. (It also has two variables, X and Y.) We need to tell EXCEL what numbers we want to use for A and B. Let's try $A = 0.5$ and $B = 3$ Now we will enter this into EXCEL, along with some labels so that we can read the spreadsheet when we are done. Enter the labels "A" into cell A3 and "B" into cell A4, and enterthe values of these two parameters into cells B3 and B4, respectively. This is shown in the screen illustration in figure 7.20.

2. Create the data table. Remember: EXCEL works best with data, and we have been working with data that is organized with variables listed across in columns and observations of those variables listed as rows. Our variables in the linear model are X and Y. So we need two columns, labeled X and Y. I'll enter these labels in cells D1 and E1. Next, in the column under the "X" we need to enter some values for the independent variable, X. We can pick anything we like, but it's easiest if we pick a nice pattern like 1, 2, 3, 4, 5... or 0.1, 0.2, 0.3, 0.4, 0.5... We'll need a lot of values, so we'll let EXCEL get the pattern going. I'd like to use 1, 2, 3, 4, 5... so I'll enter 1 in cell D2 and 2 in cell D3. Then I'll highlight those two cells. Notice that if you position the cursor over the little box in the lower right corner of the highlighted box, it turns into a "+". Click the left mouse button, hold it down, and drag straight down the screen as far as you want to go. Notice what happens: all the cells you dragged through have a dashed box around them and there is a little yellow box with a number in it. As soon as you release the mouse button, EXCEL will fill in the values according to the pattern that you started: 1, 2, 3, 4, 5... as far as you dragged down to. This setup is illustrated in figure 7.20 with the results shown in figure 7.21

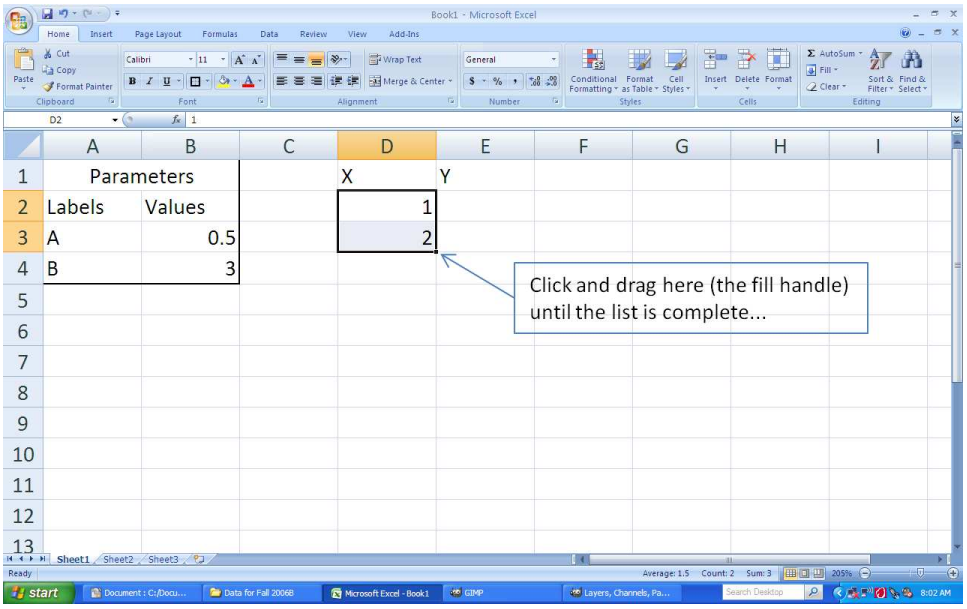

Figure 7.20: Setting up parameters and data table for creating a linear function.

We have now created half of the data table, the X values. We need to use a formula to get the Y values. Notice that our model says we can get Y by computing the value of $A * X + B$. A and B will always be the numbers we put in cells B1 and B2, but there are lots of X values that we just created.

To enter the formula for the Y variable, we need to first click on cell E2. All formulas in EXCEL begin with an equal sign, so type =. Now we need to tell EXCEL to look up the value of the parameter A. We've put this in cell B3 so we'll enter \$B\$1. Now we multiply this by X; the first value of X is in cell D2, so type D2. (The "*" stands for multiplication.) Finally, we need to add the value of B to this. So type +\$B\$2 since the parameter B is stored in cell B4. You should have typed the formula below into cell E2:

```
= $B$3 * D2 + $B$4
```

Why the dollar signs for the parameter cells (B3 and B4) but not the variable (D2)? Remember, there is only one value for A and one for B. We need to make sure that EXCEL always uses cell B3 for the value of A. The dollar signs tell EXCEL "No matter what, do not change the cell reference from B3." This is the way we force Excel to use an absolute cell reference.

Now, we just need to copy this formula to all the other cells in column E so that we get one Y value for each and every X value. Click on cell E2, position the cursor over the little box in the lower right corner of the cell, click the left mouse button, hold the button down and drag the cursor so that you highlight each cell in column E that has an X value next to it in column A. Now release the mouse button and EXCEL will fill the formula in.

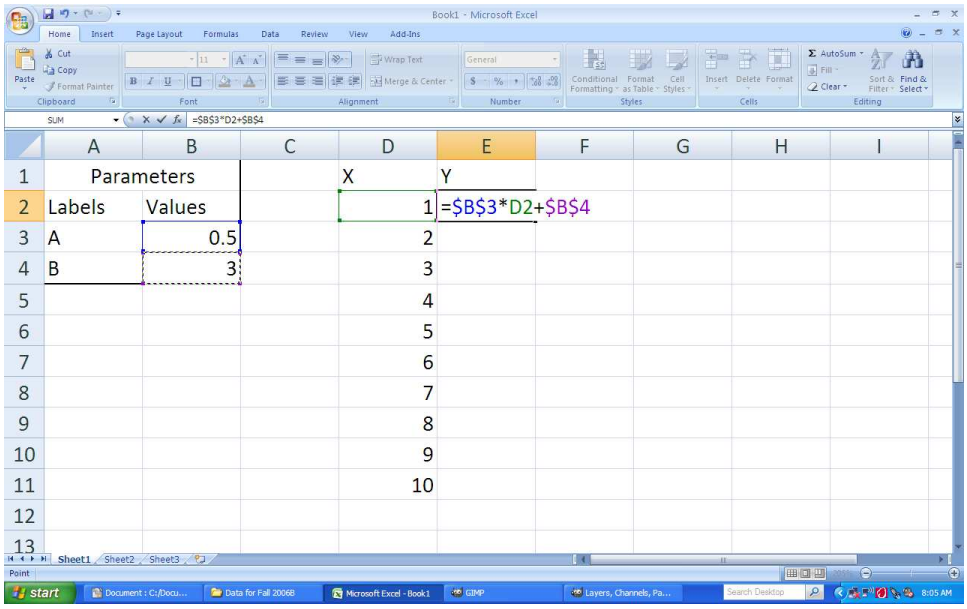

Figure 7.21: Entering the formula for a line.

To see more about the dollar signs for B3 and B4, click on cell E7 and then look up at the formula bar. Notice that when EXCEL copied the formula, the parameter references (B3 and B4) didn't change, but it is looking for the X variable in cell D7 instead of D2! This is because X is a variable - each value of X corresponds to a different value of Y - the A and the B are parameters; they never change once we set them at the top. Each value of Y uses the same values for A and B, those that we typed into the top of the spreadsheet.

3. Create the plot of the model. This is the easy part; it's just like making a scatter plot, which we've discussed in previous sections. First highlight all the data; in this example, it's cells D1 to E11. Now click on the chart wizard icon on the tool bar. Select "XY (scatter)" and pick a subtype; we've chosen the one in the second row and second column. After you're done, it should look like the figure.

   Now you can try changing the parameters and observing what happens to the graph. You could also try different models by changing the formula that you type in step 2. Explore! That's the best way to learn.

Using Goal Seek to Solve an Equation

Goal seek is a way to have Excel find approximate solutions to equations. To set it up, you need to set up your spread sheet so that there are two cells with information. The first cell contains a guess for the solution to the equation. The second cell contains a formula to calculate the actual result, based on the cell containing your guess. So, in the backpack example above, you might guess that the backpack will hold 40 books. Place this information in cell B1 and label it with "Books" in cell A1. Then, in A2, put the label "Price" and in

B2 enter the formula to calculate the price: = -30.68 + 1.46*B1. Select cell B2 (containing the price calculation) and activate the Data Ribbon. From there, select "What if analysis" and choose Goal Seek from menu. This is shown in figure 7.22.

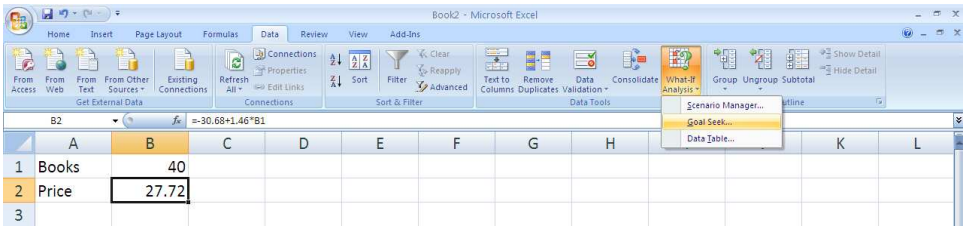

Figure 7.22: Setting up for using Goal Seek.

Fill in the values shown in figure 7.23 and hit "OK". Excel will place the results in the cells on the spreadsheet. (Note: Goal Seek needs a good guess in order to work!)

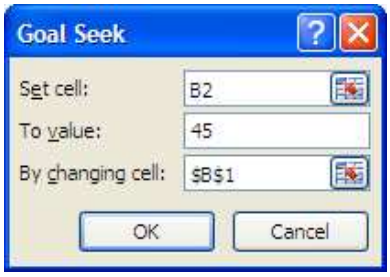

Figure 7.23: Entering values to determine how many books a $45 backpack will hold.

7.3 Homework

7.3.1 Mechanics and Techniques Problems

7.1. Look at the data on home prices in the Rochester, NY area in 2000 found in the data file "C07 Homes.xls".

1. If you were to use this data to predict the sales price of a home, which variables would you use? Based on your intuition about homes, rank the top five most important variables in determining the price of the home in order from most influential to least influential.

2. Use the graphical and numerical tools of this chapter to determine the five variables that most influence the price of a home. Rank them in order. Compare these results with your estimates in part (a). Provide evidence for all conclusions.

3. If some of the independent variables in a data set are related to each other, you may have a problem called "co-linearity". Are there any variables in the home data that you would expect to be related? Based on the numerical calculations (and possibly graphs) are any of the independent variables co-linear? Which ones? To what degree?

7.2. Consider the data in "C07 Electricity.xls" which contains observations of total monthly electric power usage compared to the size of the home (in square feet).

1. Create a scatterplot of this data. Do you expect that a simple linear model will be a good fit to this data? Why or why not? Use the features you see in the graph to explain your answer.

2. Add a linear trendline (along with its equation) to the graph. What is the best-fit simple linear model for predicting monthly electricity usage as a function of home size? What do the slope and y-intercept mean? Do these numbers make sense? Why or why not?

3. Use the model to predict the electricity usage for the following two homes: Home #1 is 2050 square feet. Home #2 is 3200 square feet.

7.3. Suppose you have two different phone plans to select from when you make long distance calls. Plan #1 costs a flat rate of 7 cents each minute (or fraction of a minute) that the call lasts. Plan #2 costs only 3 cents per minute, but has a 39 cent connection charge for all calls, no matter how long. Which calling plan would you use for a 3 minute call? Which would you use for a 45 minute call? How can you decide ahead of time which plan to use when making a call? Explain all of your answers using trendlines and scatterplots to help. Be sure your explanation uses terms like *slope* and *y-intercept* and includes information about the units of the variables involved.

7.3.2 Application and Reasoning Problems

7.4. Consider two airports that are located near each other, such as the Buffalo International Airport (in Buffalo, NY) and the Rochester Airport (in Rochester, NY). Suppose you were to collect data from each airline at each airport as to what percentage of their flights arrive on time. Your data might look something like that in the data file "C07 Airports.xls".

1. Would you expect the two variables to be strongly or weakly correlated? Explain your answers based on an analysis of the situation, not on the actual data.

2. If you said the correlation is strong, would it be positive or negative? Explain your answer. Is this relationship causal? In other words, do more on time arrivals at one airport cause more on-time arrivals at the other airport, or is it merely a coincidence that more on-time arrivals at one airport tend to be associated with more on-time arrivals at the other airport?

3. If you said they are weakly correlated, what other variable might you measure between the two airports that would be strongly correlated?

4. How do your predictions compare with the results from the actual data?

7.3.3 Memo Problem

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: Truck maintenance data

Our services have been retained by Metro Area Trucking to analyze the records they have maintained on the trucks in their fleet. The company has locations around the Rochester area, some inside the city limits and some outside the city limits. The director of operations, Ms. Mini V. Driver, at the company has asked that we determine how the different locations affect the maintenance costs on the trucks.

She has provided data on each of the trucks in the fleet. The data includes information on last year's maintenance expenses for the truck, the mileage of the truck, the age, the type of truck, and whether it is based at one of the in-city or out-of-city locations. As a first look at the data, you should separate the data into trucks that are based in the city and trucks outside the city. Mini Driver suspects that mileage and age are the most important factors, so use everything at your disposal to explain how these two quantities affect the maintenance costs. I need a full report, including graphs, tables, and formulas, as well as an analysis and explanation of what each piece of information means.

Attachment: Data file "C07 TruckData.XLS"