Chapter 3
Introduction to Linear Regression

Now we are moving on up to the big time! You are going to learn how to do something that is very remarkable—you are about to learn how to predict the future without having to call the Psychic Friends Network. What's more—your predictions are likely to be much more accurate!

If *that* is not worth the time and trouble of taking the time to learn how to use statistics, I don't know what is!

TABLE 3-1
What is Linear Regression and What Does it Tell You

1. Linear regression uses the fact that there is a statistically significant correlation between two variables to allow you to make predictions about one variable based on your knowledge of the other.
2. You should not do linear regression unless your correlation coefficient is statistically significant (See Chapter 2 for details related to determining statistical significance).
3. For linear regression to work there needs to be a **linear** relationship between the variables.

In chapter 2 you learned what a correlation coefficient is, how to calculate it, and how to determine whether it is statistically significant. You also learned how to tell how strong the relationship is with the help of the coefficient of determination. So what do you do with that information? How does it help you do anything practical?

One word—prediction! After you have found a statistically significant correlation coefficient, there is one more thing that you can do—and it is one of the coolest things in statistics—you can make predictions about one variable based on your knowledge of the other variable.

The stronger the relationship is between the two variables (larger correlation coefficients which also mean larger coefficients of determination) the more accurate any predictions you make are likely to be. Remember, the coefficient of determination tells you the amount of variation in one variable that is directly related to—or accounted for—by the other variable.

Look at Figure 3-1, to get a better idea about what I am describing. Notice that, in addition to being a scatterplot showing the relationship between Time With Company and Compensation, you now see a line drawn through the middle of the group of dots. This line, called the regression line (you will learn more about his later), was made possible by your friend the correlation coefficient.

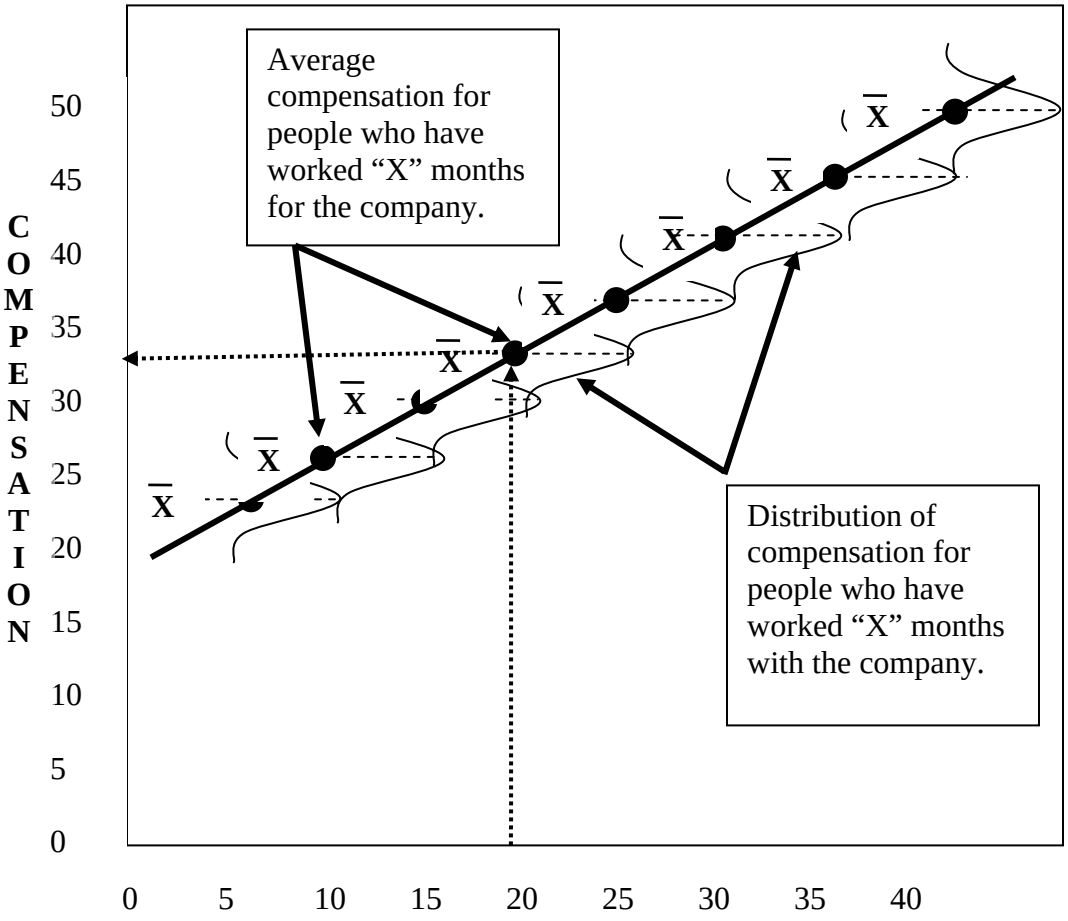
The regression line is a kind of "moving average" that is drawn through the balancing point between the dots at each point on your X-axis. This line is the one an only line that

could be drawn in a manner so that the difference between it and every individual dot is the smallest. See Figure 3-1, below.

Regression Line:

The regression line for Y (Hourly Compensation) on X (Time With Company) is a kind of "moving average". In other words, if you took the average of people's compensation at each point on the Time With Company scale, the regression line is drawn through those averages (See Figure 3-1 below). Actually, it is a little more complicated that this but it gives you an idea about what is happening.

FIGURE 3-1
Scatterplot showing distribution of Y (hourly Compensation) for each value of X (Time With Company)



Whenever you randomly select a sample of n=1 from a distribution of scores and you have to guess the value of that person's score, your best single guess—having nothing else to go by—is the mean of the distribution.

The same thing holds true in linear regression. If you are trying to predict how the compensation of an employee who has worked for the company for, say, 20 months, then your best single guess is the average compensation paid to people who have worked for 20 months with the company. Looking at Figure 3-1, above, you can see that the average compensation score for people who have worked for the company for 20 months is around 32 dollars per hour. So, if you knew that an employee had worked for the company for 20 months—and knew nothing else about the employee—your best guess about the compensation that employee receives is around 32 dollars per hour. Now, that's not too difficult, is it?

Okay, now consider this. Even though your best guess for this employee is that he or she receives around 32 dollars per hour, notice that this is only the mean of those who have worked for the company for 20 months. Even though it is your best guess, there are going to be people who have worked 20 months for the employer but day who receive more than the average and there will be those who receive compensation that is lower than the mean. In other words, there is variation around the mean—just like in any distribution of scores.

For those people who are above or below the mean, our prediction will not be accurate. The farther an individual's score is above or below the mean, the less accurate our prediction is going to be about that individual person. See Figure 3-2 for a graphical illustration of this concept.

The important point here is that in every prediction there is going to be *some* error. Still—even though your prediction may not be perfect—it is going to be much more accurate than simply guessing or trusting your intuition.

However—and this is a key point—**the larger your correlation coefficient is between the two variables, in this case Time With Company and Hourly Compensation, the stronger the relationship that exists between them. The stronger the relationship, the more accurate your prediction will be!**

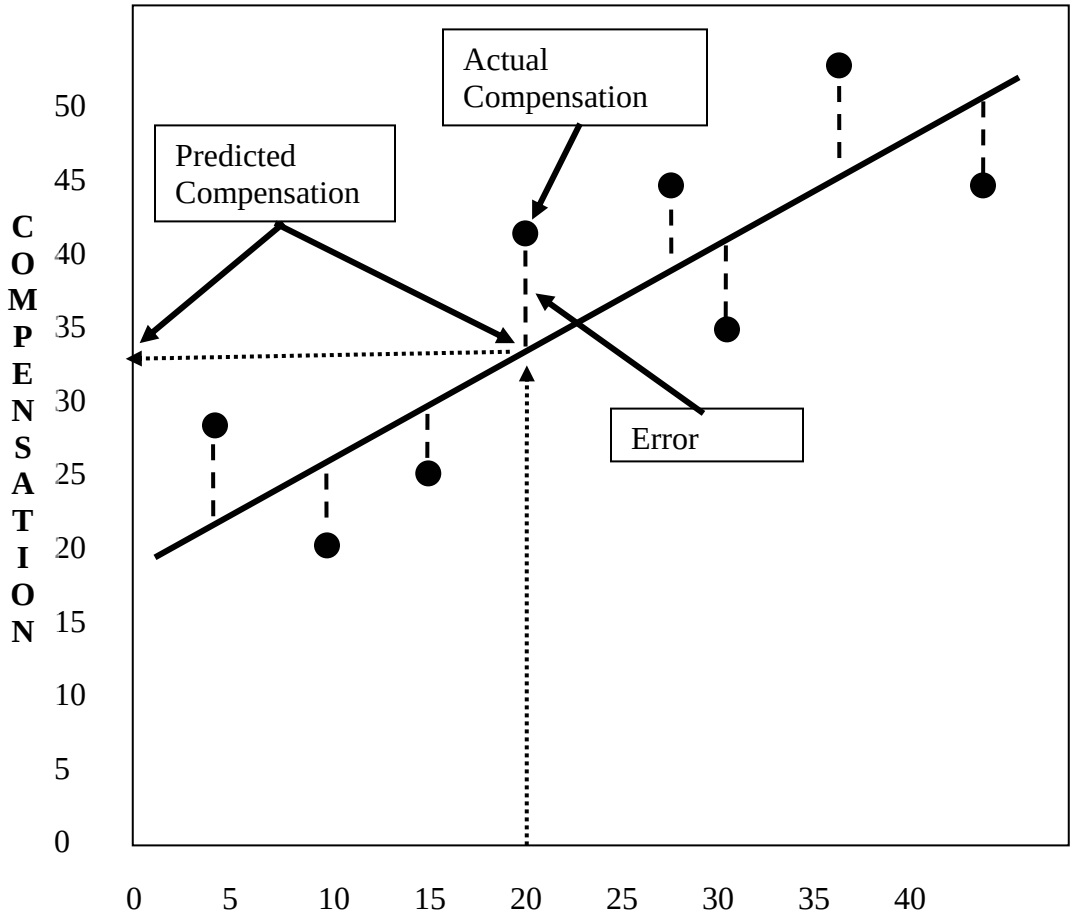
Remember that r = 0.00 means there is no correlation between the variables. This means there is no relationship and the best prediction you can make is simply a guess. However a correlation coefficient of plus or minus 1.00 tells you that there is a *perfect* relationship between the two variables. If there is a perfect relationship between Time With Company and Hourly Compensation, you can predict with 100% accuracy what a person's compensation will be based on nothing more than a knowledge of an employee's Time With Company—with no error.

Key Point

The larger your statistically significant correlation coefficient, the more accurate any predictions you make are going to be.

In the real world you are very unlikely to find relationships with a correlation coefficient as high as 1.00. Instead, it will be somewhere between 0.00 and 1.00 (and probably closer to 0.00!) Therefore, while any predictions that you make when you have a correlation of less than 1.00 will not be perfect (in other words there will be some error) it will still be a whole lot more accurate than guessing!

FIGURE 3-2
Scatterplot of Hourly Compensation by Time With Company with regression line included



Think About It!

- Suppose you are the OFCCP and you are exploring a contractor's compensation practices. If you knew that there was a correlation coefficient of .89 between the years of education and hourly compensation, you could make a scientific prediction about how likely compensation of any individual based solely on how many years of education he or she has.

So, try to master the skills that this chapter will cover. Don't just skim the material and hope for the best. Try to really understand what is happening and you will come away with a skill that will help you clearly understand how to apply statistical analyses to evaluating compensation discrimination as well as potentially answer questions raised by the OFCCP regarding a company's compensation practices.

So…How Do You Do This Linear Regression Thing?

The main thing you need to do in linear regression is calculate the **regression line**. Once you identify the regression line, it is pretty easy to make predictions about Y for any value of X (and the reverse is also true).

There are many ways to calculate the regression line. I am going to show you a method that uses information about the mean and standard deviations of each variable as well as the correlation between them.

Before we jump into the formulas, however, I want to make it clear to you that the general idea of what we covered at the beginning of this chapter is close enough, even though it is a bit simplistic, to give you the idea about what is happening. If you understand that, then you will be okay with the concept of linear regression. As I go into the formulae just remember that all we are trying to do is take all the information we can into account in order to try and make the most accurate predictions that we can.

The Raw Score Formula For Linear Regression

The formula I am going to show you looks kind of nasty to some, but trust me, it is not. All you have to do is take it slow and you will see that you are not having to do anything complicated at all. Like cooking, just add the ingredients according to the recipe and do the math and you will be home free.

The Raw Score Formula for Linear Regression

$$Y' = r\left(\frac{SD_y}{SD_x}\right)\left(X - \overline{X}\right) + \overline{Y}$$

Below is a list of the "ingredients" that you need in order to "cook" yourself up a predicted score on the Y variable for a given value of X using the raw score formula.

The List of Ingredients for Raw Score Linear Regression

r = The correlation coefficient between X and Y
SD_y = Standard Deviation on the Y variable
SD_x = Standard Deviation on the X variable
X = A raw score for which you want to predict Y
$\overline{X}$ = The mean of your X variable (e.g., time with company)
$\overline{Y}$ = The mean of your Y variable (some measure of compensation)

Continuing with the example of Hourly Compensation and Time With Company that we have been dealing with, I am going to assume that you know how to calculate the mean and standard deviation for Hourly Compensation and Time With Company. I also will assume that we have calculated the correlation between Hourly Compensation and Time With Company. I am re-writing these below:

r = -.940 **(The correlation between Hourly Compensation and Time With Company)**

SD_x = 12.040 **(Standard deviation for Time With Company)**

SD_y = 8.050 **(Standard Deviation for Hourly Compensation)**

$\overline{X}$ = 26.00 **(The mean of the Time With Company variable "X")**

$\overline{Y}$ = 28.70 **(The mean of the Hourly Compensation variable "Y")**

Once you have the above information, all you need to do is plug the "ingredients" into the formula. Below I have re-written the formula.

The formula

$$Y' = r\left(\frac{SD_y}{SD_x}\right)\left(X - \overline{X}\right) + \overline{Y}$$

The formula with all the numbers entered except the X score you want to predict a Y score for

$$Y' = -.940\left(\frac{8.050}{12.040}\right)\left(X - 26.00\right) + 28.70$$

STEP 1 – Do the division that is inside the parentheses

$$Y' = -.940\left(.669\right)\left(X - 26.00\right) + 28.70$$

STEP 2 – Beginning at the left side of the equation, begin doing the multiplication

$$Y' = (-.629)\left(X - 26.00\right) + 28.70$$

STEP 3 – **Chose an X score that you want to predict Y for** (in this case I am going to chose "15") and continue with the math.

$$Y' = (-.629)\left(15 - 26.00\right) + 28.70$$

so…

$$Y' = (-.629)\left(-11\right) + 28.70$$

STEP 4 - Do the multiplication (remember you multiply before adding or subtracting!)

$$Y' = 6.919 + 28.70$$

STEP 5 – Do the final bit of addition

$$Y' = 35.619$$

There you go! An employee who has worked for the company for 15 months is predicted to have an hourly compensation of $35.619! That is about all there is to it!

One last thing we need to know. Since we know that every prediction has error associated with it, we know that there is a chance our prediction will be wrong. So, once we have made a prediction, it would be wonderful if we could tell how accurate our

prediction is likely to be—if we knew what our chance was of being wrong. There is a statistic that tells us just that. It is called the *Standard Error of Estimate*.

The Standard Error of Estimate

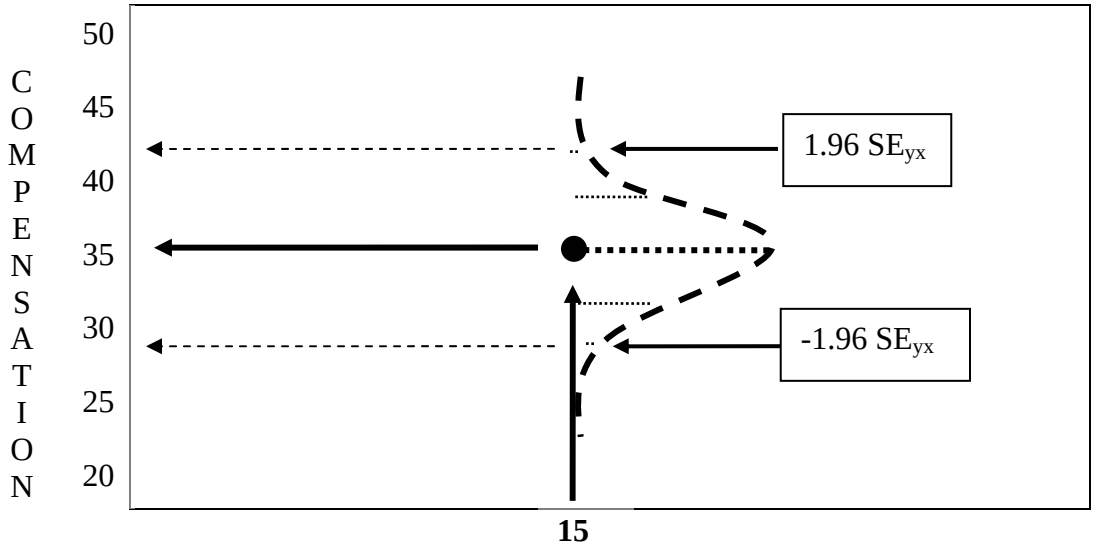
As you may recall from earlier in this chapter, when you make a prediction using linear regression, it is kind of like choosing the mean of the distribution of Y scores for a value of X (see Figure 3-1 if you are not sure what I am talking about). Remember that while the mean is your best single prediction, the actual scores for a given value of X "spread out" around that mean.

The Standard Error of Estimate (SE_{yx}) is the standard deviation of that distribution of scores. It happens to be true that 95% of all scores in a normal distribution fall between -1.96 and +1.96 standard deviations around the mean.

If we assume that the distribution of scores around the mean on the regression line is "normal", and if the Standard Error of Estimate is the standard deviation of that distribution, then once you calculate the Standard Error of Estimate, you will be able to tell how accurate your prediction based on linear regression is.

This may sound a little confusing to you, so let me give you a graphical illustration for a new mother who exercises 15 minutes a day.

FIGURE 3-3
Example of how the Standard Error of Estimate helps you tell how accurate your predictions based on linear regression are



Notice in Figure 3-3 that your predicted compensation for an employee who has worked for the employer for 15 months is around 35 dollars per hour. Similarly, while the actual hourly compensation for most of the people who have worked for the company for 15 months tend to "cluster" or "group" pretty closely around the mean, some people have higher compensation while others have lower compensation.

Notice also from Figure 3-3 that, similar to any standard deviation, 95% of employees who have worked for the company for 15 months will have an actual compensation that falls somewhere between approximately $28 and $43 per hour.

This is a VERY POWERFUL thing! If you calculate the Standard Error of Estimate, it will be very easy to say "If you tell me a person's time with the company, I can predict their compensation. Furthermore I can say with 95% confidence that, even if my prediction is not *exactly* accurate, the person for whom you are predicting their compensation will in fact have a compensation of between 1.96 SE_{yx} above and 1.96 SE_{yx} below my prediction." In other words, you can get an idea about just how accurate your prediction is!

If your correlation coefficient is large (which means that the relationship is strong) the standard error of estimate will be small which means, while your prediction may not be perfect you are likely to be very *close*. On the other hand, if the correlation is small (meaning that the relationship is weak) your standard error of estimate will be large meaning that there is greater error in your prediction. That is why larger correlation coefficients are a really good thing! They ensure pretty accurate predictions.

So how do you calculate the standard error of estimate? Believe it or not, after all of these big formulae that we have been working with, this one is really easy.

Formula for Standard Error of Estimate When You Predict Y from X

$$SE_{yx} = SD_y \sqrt{1 - r^2}$$

In English, this formula is saying "The Standard Error of Estimate for predicting Y based on X is equal to the Standard Deviation for the Y variable, multiplied by the square root of 1 minus r squared." So, to work this out, you take the following steps:

STEP 1 – Square r (multiply r by itself)
STEP 2 – Subtract the number you got in Step 1, above, from 1
STEP 3 – Take the square root of the number you got in Step 2, above
STEP 4 – Multiply the number you got in Step 3 by the standard deviation of Y

Again, sticking with the example of exercise and post-partum depression that we have been working with all along, we need the following information:

SD_y = 8.050

r = - .940

The first thing you need to do is plug the numbers into the formula as shown below.

$$SE_{yx} = 8.050\sqrt{1-(-.940)^2}$$

STEP 1 is to square the correlation coefficient as shown below.

$$SE_{yx} = 8.050\sqrt{1-.884}$$

STEP 2 requires you to subtract the value of r^2 (in this case it is .884) from 1.

$$SE_{yx} = 8.050\sqrt{.116}$$

STEP 3 is to take the square root of the value you got in Step 2, above.

$$SE_{yx} = (8.050)(.341)$$

STEP 4 is to finally multiply the result of STEP 3 by the standard deviation of the Y variable.

$$SE_{yx} = 2.745$$

So, the Standard Error of Estimate, which is a kind of standard deviation for your predictions, is 2.745. Since 95% of the actual scores will fall within -1.96 and +1.96 standard deviations around your predicted Y score, you need to multiply the standard error of estimate by -1.96 (which will tell you the number associated with 1.96 standard errors of estimate below) and by +1.96 (which will tell you the number associated with 1.96 standard errors of estimate above). See below for an example.

$-1.96\ SE_{yx}$ = (-1.96)(2.745) = -5.380

$1.96SE_{yx}$ = (1.96)(2.745) = +5.380

If our predicted Y score was 35.619, then we can say with pretty good accuracy that **95% of the time, employees who have worked for 15 months with the company will have an hourly compensation of between $30.239** (which is $35.619 minus $5.380) **and $40.999** (which is $35.619 plus $5.380) per hour.

There you have it! It may take some practice, but you have just learned a VERY powerful tool for making predictions about the future. This tool can be used in a wide variety of ways as you will discover when doing your homework! As you work through each homework problem, take the time to carefully read the scenario associated with each question. It will help you understand how linear regression can be used to solve problems in the real world.

Terms to Learn

You should be able to define the following terms based on what you have learned in this chapter.

Linear Regression
Predicted Score (Also called Y')
Regression Coefficient (also called "b")
Regression Line
Slope (also called "a")
Standard Error of Estimate (also called S_{yx})