Chapter 4
Introduction to Multiple Regression

Now that we have added a new tool to our statistical tool box, let's take a moment to review what we have.

1. **The Correlation Coefficient**: A single summary number that tells you whether a relationship exists between two variables, how strong that relationship is and whether the relationship is positive or negative.
2. **The Coefficient of Determination**: A single summary number that tells you how much variation in one variable is directly related to variation in another variable.
3. **Linear Regression**: A process that allows you to make predictions about variable "Y" based on knowledge you have about variable "X".
4. **The Standard Error of Estimate**: A single summary number that allows you to tell how accurate your predictions are likely to be when you perform Linear Regression.

I want to spend just a little more time dealing with correlation and regression. This chapter is only going to provide you with an introduction to what is called "Multiple Regression". Multiple regression is a very advanced statistical too and it is extremely powerful when you are trying to develop a "model" for predicting a wide variety of outcomes. We are not going to go too far into multiple regression, it will only be a solid introduction. If you go to graduate school you will probably have the opportunity to become much more acquainted with this powerful technique.

Quick Review

- You use correlation analysis to find out if there is a statistically significant relationship between TWO variables.
- You use linear regression analysis to make predictions based on the relationship that exists between two variables.

The main limitation that you have with correlation and linear regression as you have just learned how to do it is that it only works when you have TWO variables. The problem is that most things are way too complicated to "model" them with just two variables.

For example, suppose I asked you the following question, "Why does a person receive the compensation that they do?" What would you say? You might say something like the following:

- It could have something to do with how long a person has worked for the company.
- It could have something to do with how much experience a person has doing their specific kind of work.
- It could have something to do with their age (Age is a "proxy" for experience).
- It could have something to do with the type of work they do.
- It could have something to do with their performance ratings.
- It could have something to do with what part of the country they live in.

You probably get the idea. How much a person gets paid is really based on more than just a single piece of information. In reality, all of the above factors (and more besides) are likely to play some role in why a person gets paid what they do.

If you were going to use standard correlation to study why people receive the compensation they do, you would be limited to only looking at one of these things at a time. For example, you could use correlation to study the relationship between a person's current compensation and their time with the company (as we did in the chapter on linear regression). You could also use correlation to study the relationship between a person's current compensation and how many years of school they completed. However, you could not do both to find out how a person's current compensation is related to both their education and how long they have worked for the company. Remember, Pearson's correlation is a "bi-variate" tool meaning that it is designed to find relationships between only *two* variables.

And yet, we know that life is so complicated that it takes way more than two variables to even begin to explain/predict why things are the way they are.

What you need is a new tool—Multiple Regression.

Multiple Regression (R)

A statistical tool that allows you to examine how **multiple independent variables** are related to a dependent variable. Once you have identified how these multiple variables relate to your dependent variable, you can take information about all of the independent variables and use it to make much more powerful and accurate predictions about why things are the way they are. This latter process is called "Multiple Regression".

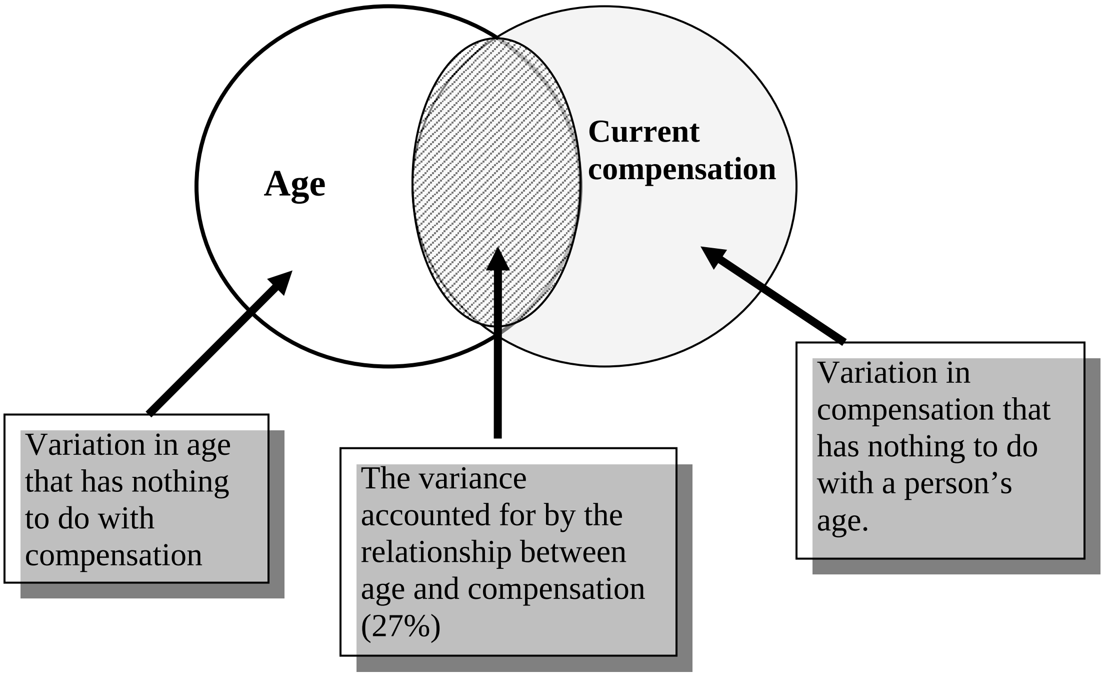
Let's take a look at a couple of examples that will hopefully make this concept a bit easier for you to grasp. I am going to use Venn Diagrams similar to what we used to try and get a handle on what the Coefficient of Determination means.

- Figure 4-1 presents a picture of how two variables are related to each other.
- Figure 4-2 presents a picture of how two independent variables are related to a dependent variable—while the two independent variables are NOT related to each other.

- Figure 4-3 presents a picture of how two independent variables are related to a dependent variable—while the two independent variables ARE related to each other.

You must understand these three figures in order to understand the concepts of multiple correlation and multiple regression.

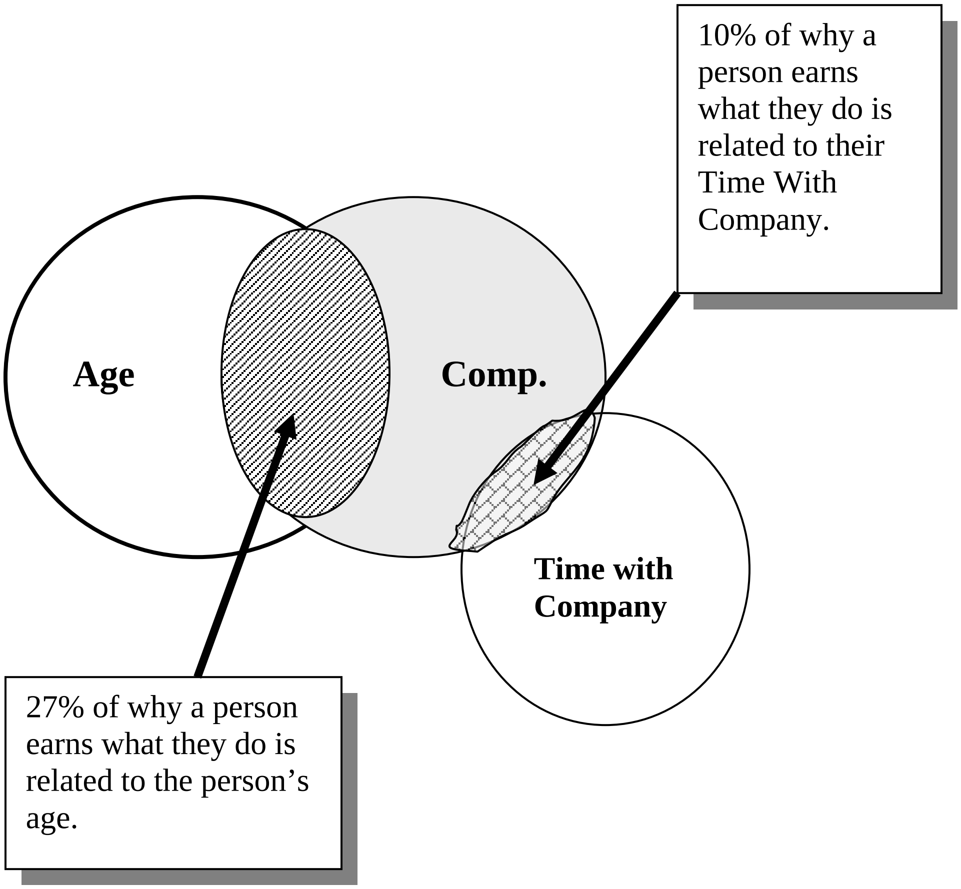
FIGURE 4-1
Example of the Relationship between Age and Current Compensation


Age
Current compensation
Variation in age that has nothing to do with compensation
The variance accounted for by the relationship between age and compensation (27%)
Variation in compensation that has nothing to do with a person's age.

In this example, 27% of what there is to know about a person's current compensation is accounted for by that person's age. In other words, if you know a person's age, you know about 27% of what you need to know to make an accurate prediction about what their compensation is.

There is nothing new in Figure 4-1. This is simply a re-statement of what you already read about in the chapter on correlation. If you are unclear about what Figure 4-1 means, please return to that chapter and review the "coefficient of determination". The main point is that the correlation between age and compensation tells us that a person's compensation seems to change as a person ages. This makes intuitive sense because one would expect that as a person ages, he or she works their way up at their job and gets paid more.

FIGURE 4-2
Example of the Relationship between Age (for those over 18 years of age) Time with Company and Compensation



10% of why a person earns what they do is related to their Time With Company.

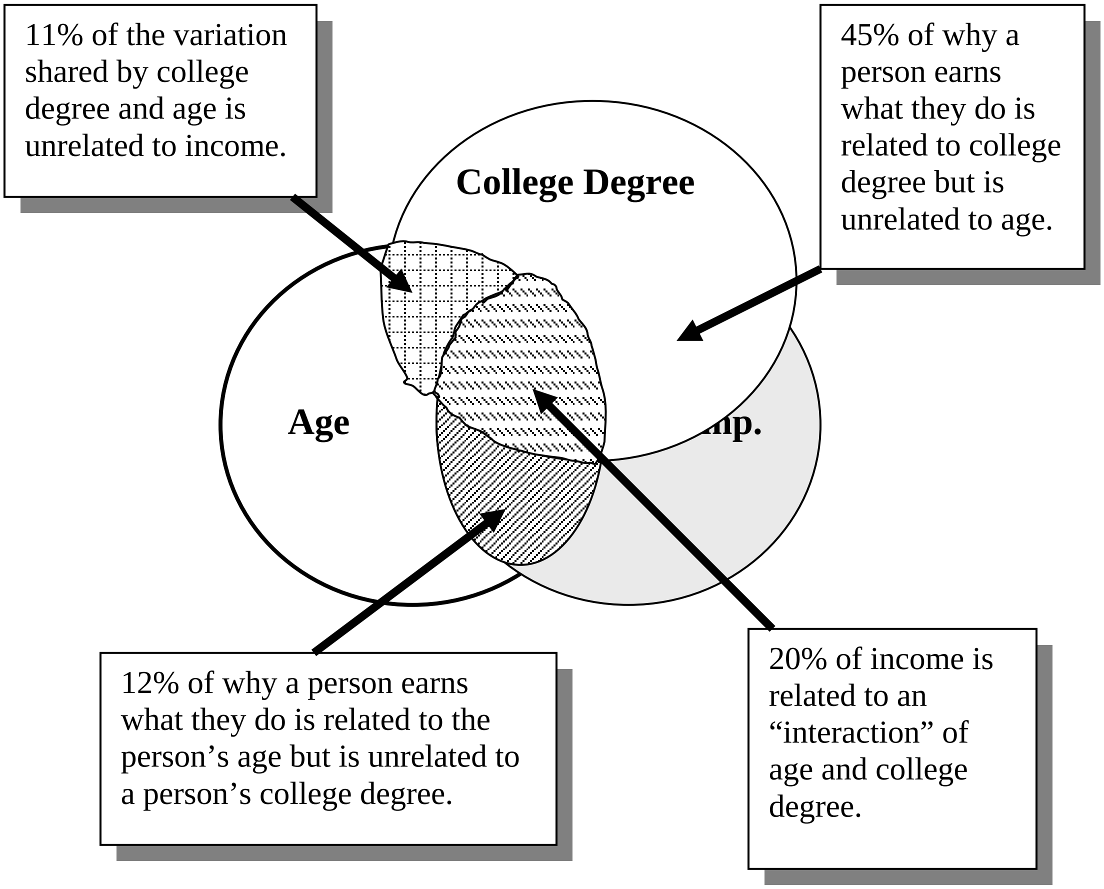
27% of why a person earns what they do is related to the person's age.

Important Point!

Notice that a person's time with company accounts for about 10% of why they earn what they do. By adding this variable to our study, we improved our understanding of why people earn the income they do from 27% to 37%. In other words, using two variables rather than one variable, we improved our ability to make accurate predictions about a person's salary.

Figure 4-2 is a good illustration about what multiple correlation and regression is designed to do. By having more than one "predictor" variable (age and time with company), we are able account for more of the variance in compensation. As a result, we can make much more accurate predictions. This is because the second variable adds additional important information about your dependent variable (compensation).

FIGURE 4-3
Example of the Relationship between Age (for those over 18 years of age), College Degree and Compensation



11% of the variation shared by college degree and age is unrelated to income.

45% of why a person earns what they do is related to college degree but is unrelated to age.

12% of why a person earns what they do is related to the person's age but is unrelated to a person's college degree.

20% of income is related to an "interaction" of age and college degree.

Important Point!

Notice that this is much more complicated! 12% of a person's compensation is related to age, 45% is related to a person's college degree and 20% is related to an interaction between age and college degree. In this case we have pumped up our ability to predict/explain compensation to nearly 77%!

Hopefully, after looking at Figures 4-1, 4-2 and 4-3 you understand the following facts:

- When you have more than one independent variable you may very well be able to make more accurate predictions about your dependent variable.

  Think of it like trying to buy a car. If you only ask one of your friends what kind of car they think is best, you will get some information but it will be limited. But…if you ask ten of your friends the same question, you will get more information and are probably more likely to identify a good car. The same thing is true with regression research. If you are trying to develop a way to predict whether a seriously depressed person is likely to commit suicide, your prediction will be much more accurate if you take multiple sources of information (independent variables) into account like age, medical history, history of prior attempts at suicide, score on a clinical depression survey, number of friends/family in the person's support network.

- Things are pretty straight forward when you have multiple independent variables AND they are not related to each other (Figure 4-2).

- Things get much more complicated when your multiple independent variables are related to with each other. In other words, when the independent variables "interact" with each other as well as with the dependent variable. In this case, in order to be able to make predictions you need to break all of the correlations down so that you can figure out the value of multiple R.

So, again, trying to keep this kind of simple ("Yea, right!" you say) here is what we want to do.

1. We want to calculate a single summary number that tells us how strong the relationship is between ALL the independent variables and the dependent variable. What we want is similar the correlation coefficient "r". Remember, however, that "r" is only used with 2 variables. The statistic you are going to learn is called "R" (which is a capitol "r"). Whenever you see a capitol "R" it is interpreted just like any regular correlation coefficient except it tells you the strength of the combined relationships between all the independent variables and the dependent variable.
2. We want a single summary number that tells us how much of the variability in the dependent variable is related to ALL of the independent variables. When we talked about regular old correlation, we learned about the "Coefficient of Determination" which is symbolized as "r^2". We can do the very same thing with "R" to get an understanding of how much variation in the dependent variable is accounted for by the independent variables. To get this statistic, all you need to do is square your "R" value which gives you "R^2".

3. Finally, you want to be able to actually make predictions about a dependent variable taking into account all of the information provided by all the independent variables.

How to Compute R (Which is also called the multiple correlation coefficient)

The best way to show you how to do this is to use and example. Let's assume that you are a personnel psychologist working for General Motors. The company wants to develop a new hiring process that will help them identify job applicants who will be the most productive car salespeople.

Here is how you used multiple regression to develop a way to predict who will make the best salespeople.

1. You went out and took a random sample of 5 currently employed General Motors salespeople. Really, you need to have a larger sample, but to make this a little easier to follow I am using a really small data set.

2. You collected the following information about each of the 5 salespeople:
   - Highest year of school completed
   - Motivation as measured by the Higgins Motivation Scale
   - How many dollars in sales the person made last year

3. You calculate the correlation between each possible pair of variables:
   - Correlate: **Highest year of school completed** with **Motivation**
   - Correlate: **Highest Year of School Completed** with **Dollars in Sales**
   - Correlate: **Motivation** with **Dollars in Sales**

4. Plug the correlations into the Multiple R formula

5. Do the math!

<u>STEP 1 – Select your random sample</u>

Remember, taking a random sample is critical if you want to be able to take your findings and use them to make decisions in the real world.

<u>STEP 2 – Collect your Data</u>

Suppose that we took our 5 randomly selected salespeople and collected the information you can see in Table 14-1.

TABLE 4-1
Data Collected From Random Sample of 5 General Motors Salespeople

Independent Variable 1 (X1)	Independent Variable 2 (X2)	Dependent Variable (Y)
Highest Year of School Completed	Motivation as Measured by Higgins Motivation Scale	Annual Sales in Dollars
12	32	$350,000
14	35	$399,765
15	45	$429,000
16	50	$435,000
18	65	$433,000

I am not going to take the time to work through calculating the correlation coefficients between these three variables. Just remember that you do it just like you did back in Chapter 12. Lets assume that you did all the math an you got the following information:

	Mean	**Standard Deviation**
Highest Year of School	15	2.236
Motivation	45.4	13.164
Annual Sales	$409,353	$36,116.693

Correlation between Highest Year of School and Motivation ($r_{x1,x2}$) = 0.968

Correlation between Highest Year of School and Annual Sales ($r_{x1,y}$) = 0.880

Correlation between Motivation and Annual Sales ($r_{x2,y}$) = 0.772

Using this information we are ready to use the correlation coefficients above to compute “R”.

The Formula for R

$$R = \sqrt{\frac{\left[\left(r_{y,x1}\right)^2 + \left(r_{y,x2}\right)^2\right] - (2r_{y,x1}r_{y,x2}r_{x1,x2})}{1 - \left(r_{x1,x2}\right)^2}}$$

Now all we need to do is plug in the numbers and do the math. I think once we have finished this, you will agree that this was the easy part!

STEP 1 – Plug in the Numbers

$$R = \sqrt{\frac{((.880)^2 + (.772)^2) - (2(.880)(.772)(.968))}{1 - (.968)^2}}$$

STEP 2 – Working the Math

$$R = \sqrt{\frac{(.7744 + .5960) - (1.3152)}{1 - .9370}}$$

Then…

$$R = \sqrt{\frac{1.3704 - 1.3152}{0.063}}$$

Then…

$$R = \sqrt{\frac{0.0552}{0.063}}$$

Then…

$$R = \sqrt{.8762}$$

Finally…

$$R = \sqrt{.8762}$$

Therefore…

R = .9360

So, what does this “R = .9360” mean? It is really pretty simple. It tells you, “The combined **correlation between Years of Education and Motivation** with a **salesperson’s Annual Sales** is .**9360**.”

Remember, all correlations—even multiple correlations must be between + or – 1.00. A Multiple Correlation, just like any other correlation, of 1.00 means that the two independent variables, when taken together have a perfect relationship with salesperson annual sales. If “R = 0.00” that would mean that there was no relationship at all between education, motivation and annual sales.

Since our Multiple Correlation is .9360, the two variables seem to have a very strong relationship with annual sales. In other words, we could make very accurate predictions about how much money a salesperson will bring in if we know nothing more about the person than their education and their score on a motivation assessment scale. If all of this were true data (rather than the made up data I have created to help you understand the process) we would have a VERY powerful way to select new salespeople and we would become very rich—very fast!

I don’t know if you have fully captured the vision on this, the let me say again…THIS IS VERY COOL, VERY POWERFUL, AND VERY IMPORTANT.

You could use this technique to do all kinds of things like:

- Predict/Explain a person’s current compensation based on a number of employee characteristics;
- Predict/Explain patient survival after surgery based on a number of personal characteristics;
- Predict the likelihood of a recently released convicted criminal re-offending based on a number of personal characteristics.
- Predict a graduate student’s likelihood of performing well in graduate school based on a number of personal characteristics
- I could go on and on and on…..

Making Predictions: Multiple Regression

Okay, so now we have a measure that allows us to establish whether or not our independent variables are effective predictors of our dependent variable. Now we can take the next step and actually use our knowledge to make predictions. This will be very similar to what was done in Chapter 13 but with an extra step.

Remember that with standard linear regression the algebraic formula for making predictions is:

$$Y' = a + bX$$

In the formula above:

Y' = A predicted value of Y (which is your dependent variable)

a = the value of Y when X is equal to zero. This is also called the “Y Intercept”.

b = The change in Y for each 1 increment change in X

X_1 = an X score on your first independent variable for which you are trying to predict a value of Y

X_2 = an X score on your second independent variable for which you are trying to predict a value of Y

<u>The Formula for Multiple Regression</u>

$$\mathbf{Y' = a + b_1X_1 + b_2X_2}$$

Y' = A predicted value of Y (which is your dependent variable)

a = The “Y Intercept”.

b_1 = The change in Y for each 1 increment change in X_1 (In our case, this is Highest Year of School Completed).

b_2 = The change in Y for each 1 increment change in X_2 (In our case, this is level of motivation as measured by the Higgins Motivation Scale.)

X = an X score (X is your Independent Variable) for which you are trying to predict a value of Y

How to Calculate b_1 and b_2

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - \left(r_{x1,x2} \right)^2} \right) \left(\frac{SD_y}{SD_{x1}} \right)$$

$$b_2 = \left(\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - \left(r_{x1,x2} \right)^2} \right) \left(\frac{SD_y}{SD_{x2}} \right)$$

$r_{y,x1}$ = Correlation between Highest Year of Education and Annual sales

$r_{y,x2}$ = Correlation between Motivation and Annual Sales

$r_{x1,x2}$ = Correlation between Highest Year of Education and Motivation

$(r_{x1,x2})^2$ = The coefficient of determination (r squared) for Highest Year of Education and Motivation)

SD_y = Standard Deviation for your Y (dependent) variable.

SD_{x1} = Standard Deviation for the first X variable (Education)

SD_{x2} = Standard Deviation for the second X variable (Motivation)

<u>Calculating the Regression Coefficients</u>

Highest Year of Education

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - \left(r_{x1,x2} \right)^2} \right) \left(\frac{SD_y}{SD_{x1}} \right)$$

$$b_1 = \left(\frac{(.880) - (.772)(.968)}{1 - (.968)^2} \right) \left(\frac{36{,}116.693}{2.236} \right)$$

Excerpted from The Radical Statistician by Jim Higgins, Ed.D.

$$b_1 = \left(\frac{(.880)-(.747)}{1-.937}\right)\left(\frac{36{,}116.693}{2.236}\right)$$

$$b_1 = \left(\frac{.134}{.063}\right)\left(\frac{36{,}116.693}{2.236}\right)$$

$$b_1 = (2.127)(16{,}152.367)$$

$$b_1 = 34{,}356.085$$

Motivation Score

$$b_2 = \left(\frac{r_{y,x2} - r_{y,x1}r_{x1,x2}}{1-\left(r_{x1,x2}\right)^2}\right)\left(\frac{SD_y}{SD_{x2}}\right)$$

$$b_2 = \left(\frac{(.772)-(.880)(.968)}{1-(.968)^2}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

$$b_2 = \left(\frac{.772-.852}{1-.937}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

$$b_2 = \left(\frac{-0.08}{0.06}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

$$b_2 = (-1.333)(2743.596)$$

$$b_2 = -3{,}657.213$$

How to Calculate "a"

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$\overline{Y}$ = The mean of Y (Your dependent Variable)

$b_1 \overline{X}_1$ = The value of b_1 multiplied by the Mean of your first independent variable (in this case, Highest Year of Education.

$b_2 \overline{X}_2$ = The value of b_2 multiplied by the mean of your second independent variable (in this case, Motivation score)

<u>Calculating "a"</u>

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$a = 409{,}353 - (34{,}356.085)(15) - (-3{,}657.213)(45.4)$$

$$a = 409{,}353 - 515{,}341.275 - (-166037.470)$$

$$a = 60{,}049.195$$

So…Let's Make a Prediction!

Okay, let's say you interviewed a potential salesperson and found that they had 13 years of education (they took 1 year of college after high school) and they scored 49 on the Higgins Motivation Scale. What would be your prediction of how much money in sales this person would bring in on an annual basis?

Years of School = 13

Motivation Score = 49

The Formula:

$$\mathbf{Y' = a + b_1X_1 + b_2X_2}$$

$$\mathbf{Y' = 60{,}049.195 + (34{,}356.085)X_1 + (-3{,}657.213)X_2}$$

Note that I did not plug in the numbers for X_1 and X_2. These are the places where you plug in your values that you are going to use to make a prediction. In this case, X_1 refers to the number of years of school (**13**) and X_2 is the motivation score (**49**).

So, if we plug in these final numbers, we can make our prediction. See below.

$$\mathbf{Y' = 60{,}049.195 + (34{,}356.085)(13) + (-3{,}657.213)(49)}$$

$$\mathbf{Y' = 60{,}049.195 + 446{,}629.105 + (-179203.437)}$$

$$\mathbf{Y' = 685{,}881.737}$$

So, given a job applicant with 13 years of education completed and who received a motivation score of 49 on the Higgins Motivation Scale, our single best prediction of how much this person will earn for our dealership is $685,881.74. Pretty cool, huh? Think a for a few minutes about how a tool like this could be used in whatever career field you are thinking about going in to!

Terms to Learn

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

Multiple Correlation
Multiple Regression