

Introduction to Regression

*Dr. Tom Pierce
Radford University*

In the chapter on correlational techniques we focused on the Pearson R as a tool for learning about the relationship between two variables. As a descriptive statistic the Pearson R contains information about the direction of the relationship, as well as its strength. A value for R can also be tested for statistical significance to help the researcher decide whether a relationship between the two variables exists out there in the real world (i.e., in the population as a whole).

Determining the nature of relationships among variables is an important step in developing and testing theories about human behavior. However, this isn't the only way to put a correlation coefficient to work. Remember, the presence of a significant correlation means that the two variables overlap in terms of the information they provide. This means that one variable can be used to give you information about what a person's score is most likely to be on the other variable. Once researchers have established that a correlation exists between two variables they can use it to make a best guess about a person's score on one variable when the only thing they know about them is their score on the other variable. And that's what regression is all about.

Regression has at least two important applications:

1. First, there are many situations where a person might want to use regression for the express purpose of predicting a person's scores on one variable when the only things they know about a person are their scores on one or more other variables. For example, an industrial/organization psychologist might want to use information they get from a job applicant to predict their score for a measure of job performance obtained a year after they've started work. A clinical psychologist working in a prison may want to be able to predict which prisoners are most likely to commit an act of violence while incarcerated. In these cases the person using regression doesn't necessarily care what the predictor variables are as long as they result in accurate guesses about the behavior in question. If my job is to predict who's going to be a good employee and who isn't then I don't care if the best predictor of future job performance is the amount of cheese the applicant ate last week. I'll use the amount of cheese to predict job performance and I'll like it. Theory-schmeery.
2. A second use of regression is to test different ideas about the relationships among the variables that are relevant to a given area of study. The goal of this line of work is to develop a theoretical understanding of how the variables in a set of variables are related to each other and, when possible, to make some shrewd guesses about the chain of events that leads to some people having higher or lower scores on the behavior of interest. In the long run, the more we understand about *why* people behave in the way they do, the better we'll

be in to help them in the future. For example, the more we understand about the factors that lead to job burnout the more likely it is that a person in human resource management could intervene effectively to prevent it.

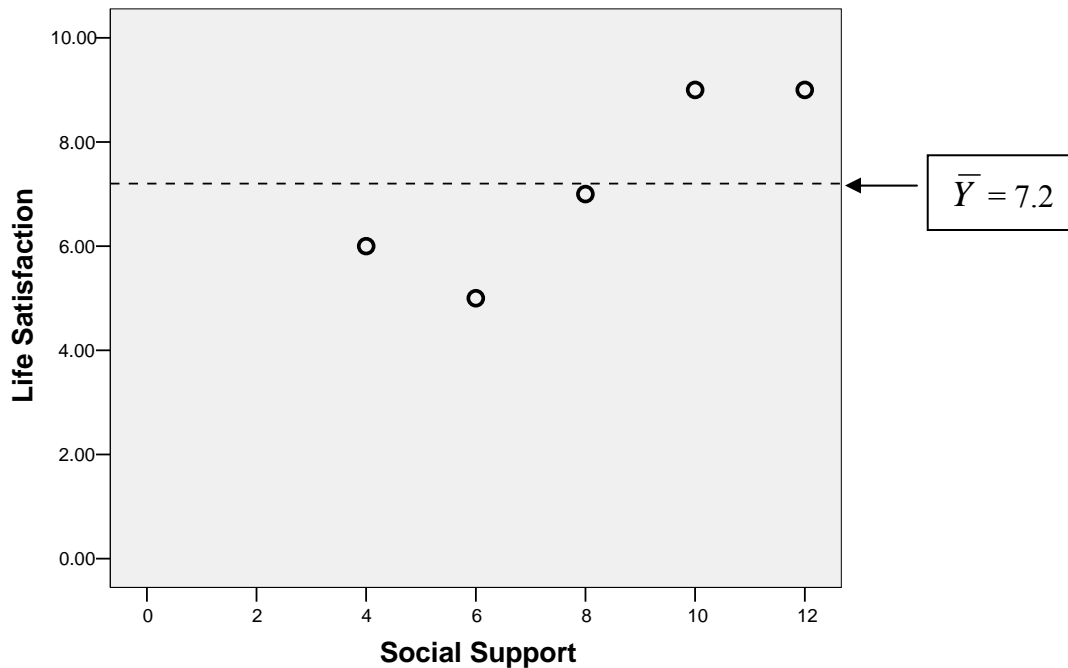
Okay. Let's start by talking about regression as a tool for predicting a person's score on a particular variable. We're going to work with the same variables we had in the correlation chapter. In that data set, we had scores for measures of life satisfaction and social support from five Alzheimer's caregivers. We found that the correlation between the two variables (.884) was significantly different from zero. Let's say that the researcher wants to predict scores for life satisfaction from scores for social support. By convention, the variable doing the predicting is referred to as variable X. It is also known as the **independent variable** or the **predictor variable**. The variable being predicted is referred to as variable Y. It is also known as the **dependent variable** or the **criterion variable**. In this example, social support is the predictor variable (X) and life satisfaction is the criterion variable (Y).

The data are presented below:

	X	Y
Participant	Social Support	Life Satisfaction
1	4	6
2	6	5
3	8	7
4	10	9
5	12	9
<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> $\bar{X} = 8.0$ $S_X = 3.16$ </div> <div style="text-align: center;"> $\bar{Y} = 7.2$ $S_Y = 1.79$ </div> </div> <p style="text-align: center;">$r_{xy} = .884$</p>		

The goal of regression is pretty simple. You're going to use what you know about a person's score on variable X to make a best guess about that person's score on variable Y. For example, suppose you're asked to make a best guess about a person's score for life satisfaction, but the only thing you know about them is that their score for social support is 4.0. Because you don't know their actual score for life satisfaction you're going to have to estimate what it would have been if you'd actually measured it. So what do we mean by a *best guess*? Well, you know that it's too much to expect that your predicted score is going to be exactly equal to the person's actual score. But it is reasonable to expect that we've done everything we can to **make the difference between our best guess and their actual score as small as possible**. I'm going to use the symbol Y' to represent a predicted score for variable Y. So, I could say that what I want to do is to make the difference between Y and Y' as small as possible.

Figure X.1 The scatterplot between social support and life satisfaction.



Take a look at the scatterplot between social support and life satisfaction. The points are higher on the right side of the graph than on the left side. This makes sense because we already knew that there was a positive relationship between the two variables. You should also notice that, logically, the predictor variable *social support* (variable X) is on the X-axis and that the criterion variable of *life satisfaction* (variable Y) is plotted on the Y-axis.

So what does the graph tell you about making a best guess for someone's score for life satisfaction when the only thing you know about them is their score for social support? One place to start would be to ask "*What would your best guess for life satisfaction have to be if you **didn't** know their score for social support?*" In other words, all you know about the person is that they're one of the five people in the study. At this point the best you can do is to pick the number that's as in the middle of the raw scores for life satisfaction that you can get – that's the mean score for life satisfaction of 7.2.

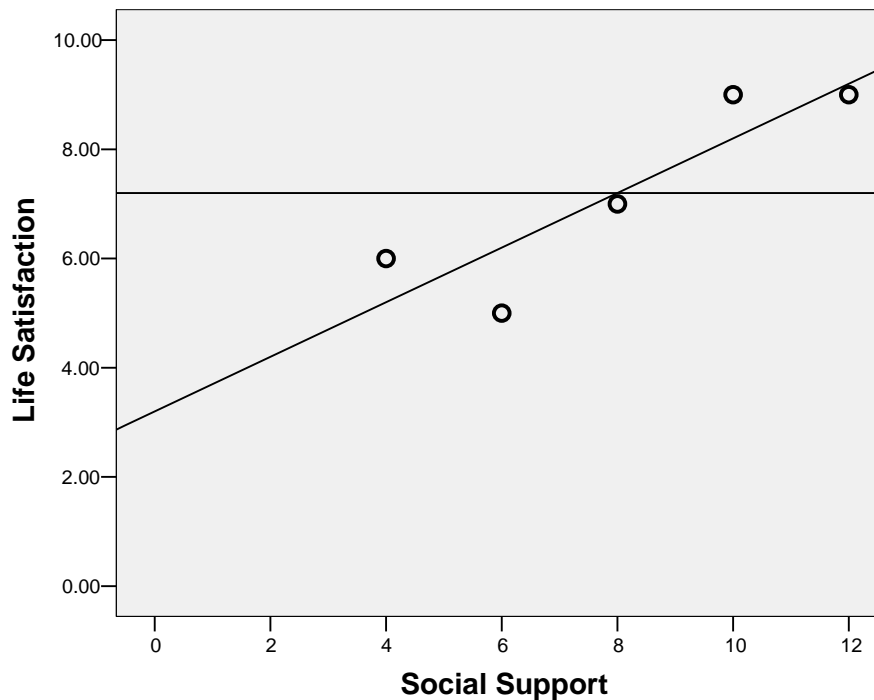
On the graph, you'll notice that the mean for variable Y of 7.2 is shown as the flat line. *And if you use the mean of 7.2 how far off can you expect this best going to be?* Well, the standard deviation of raw scores for variable Y is 1.79, so on average raw scores for Y deviate from their mean of 7.2 by 1.79 points. Therefore, ***on average, if you use the mean of Y as your best guess you're going to be off by 1.79 points.***

Okay. Now, let's say you get one additional piece of information to help you to make your best guess. You know that the person's score for social support is 4.0. Look at the scatterplot. *What do scores for variable Y tend to look like when scores for X are down around 4.0? Do they look very close to 7.2?* No! The scatterplot shows you that ***when people have low scores on variable X they tend to have scores for variable Y that are lower than the mean of 7.2.*** It also shows you that when people have high scores on variable X they tend to have scores on variable Y that are higher than the mean of Y. Without going numbers crazy at this point, we may not know exactly what number to pick as our best guess for variable Y, but we know already that using the mean of Y as our best guess for a score on Y when someone has a score of 4.0 on X would be a pretty bad idea! ***The presence of a significant positive relationship between X and Y tells us that our best guess for Y ought to change – it ought to get larger – as scores for X get larger.***

Okay, the scatterplot tells you that when someone has a score of 2.0 for variable X your best guess for that person's score on variable Y ought to be below the mean for variable Y. *How far below the mean for Y should this best guess be?* Again, we can start by looking at the scatterplot. The scatterplot contains all the information we have about the relationship between X and Y. The thing we need to do now is to find a way to capture or describe this relationship in terms of a simple pattern.

You could reasonably ask at this point “*Well, what pattern?*” The answer to that question is based on the fact that the relationship between most pairs of variables in psychology is best captured by the simplest pattern you could have – the pattern of a straight line. We're going to describe the relationship between the two variables we're working with by drawing a straight line through the five points in the scatterplot. We want the line to run as close to the points in the scatterplot as we can get. After we've drawn the best line we can we're in a position to use it to make a best guess about a person's score for variable Y when all we know about them is their score for X. When we start from a particular point along the scale for variable X, each point on the line represents how far up on the scale for variable Y we have to go to get to the best guess we can make about a person's score for Y. Below is the same scatterplot with a line drawn through the points in the scatterplot.

Figure X.2



We want to see what someone's predicted score for Y is when someone has a score of 2.0 on X. Go to the location on the X-axis that corresponds to a score of 2.0. Now go straight up from there to the line we just drew through the scatterplot. The person's predicted score for Y is however far up you have to go on the scale for variable Y. If you want to know what this number is, start at the point on the line we just identified and go straight across to the Y axis. The value on the Y-axis looks like it's right around 4.0.

And that's pretty much it. This strategy works the same way for any value of X that you'd like to work with. And it gives us predicted scores for Y that are pretty close to the most accurate ones we could get. However, this method is based on visually "guesstimating" how to draw the best fitting line through the points in the scatterplot. If anyone were to ask you if you had the best line possible – and, consequently, the most accurate predicted scores possible – you'd have to say "probably not". To get the best guesses possible we need to do more than guesstimate where the best-fitting line should be. As a statistical tool, regression provides a set of quantitative methods that can tell you how to draw the regression line that gives you the most accurate predicted scores for Y possible.

So how will a statistician know that they've got the absolutissimus bestust regression line possible? There must be some rule or criterion they use for knowing that they've got the best one. And there is. The nice thing is that we've already talked about what that criterion is! We talked about it when we described what a statistician would find really spiffy-neato-radically-awesome about the mean. If you think back a few chapters, we said that the mean, unlike the mode or the median, took the data from every subject into

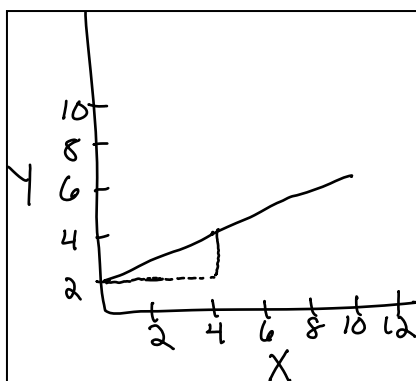
account. We also said that a statistician would say that the mean is as “*in the middle*” of a set of raw scores as you can get. What was the criterion – the rule – for knowing this? Our friendly statistician smugly told us that “*the mean is the one number that minimizes the sum of squared deviations around it*”. The mean was the one number we could subtract from each of the raw scores that would make the sum of squares as small as it could possibly be. The equation for the mean represented the “least-squares solution” for making the average deviation of raw scores from the mean as small a number as possible

The rule for knowing you’ve got the best regression line possible is that it makes the average difference between a person’s actual score on variable Y as close to their predicted score for Y (the point on the line) as possible. We need to make differences between Y and Y’ as small as we can, which means that we want to minimize the variability of point around the line. This is the same kind of thing as minimizing the variability of a set of raw scores around their mean. ***We’ll know we’ve got the best regression line possible when it minimizes the sum of squared deviations around it.*** In other words, when we calculate $Y - Y'$ for each person, we’ll have the deviation from the line for each person. If we take each of these deviations, then square them, and then add them up, we’ll have the sum of those squared deviations. We need the equation for the straight line that makes this sum of squares as small as it can be.

The equation for a straight line

To draw any straight line you need to know two things. First, you have to know where the line runs into the Y-axis. In other words, when X is equal to zero, what’s the value for Y. This number is known as the **Y-intercept** and in the behavioral sciences it generally goes by the symbol of a small-case letter a. For example, look at the line below. The y-intercept of this line is 2.0.

Figure X.X



Okay, now you know the location of one point on the line. You know that it only takes two points to draw any straight line. To find a second point you need an additional piece of information. You need to know the angle at which the line moves away from the y-intercept. Does it go up at a really steep rate, is it flat and move straight across from left to right, or does it go down as you move from left to right? This information is contained

in the **slope** of the line. The slope is a ratio of however much the line moves up or down for a particular distance that it moves from left to right. It's the amount of change in variable Y that you need to have in order to reach a second point on the line divided by the change in variable X that you need in order to reach that second point. An easier way to remember this might be to remember the slope of a straight line as "the rise over the run". The symbol for the slope is a small-case letter b. Take a look at the line again. The figure shows that when you start at the y-intercept you have to go up 2 points and over 4 points to get to the second point on the line. This gives us a ratio of a change in Y of +2.0 to a change in X of +4.0, or 2 divided by 4, or +.5.

$$\text{Slope} = \frac{\text{Change in Y}}{\text{Change in X}} = \frac{2}{4} = +.5$$

Okay, now we know both the y-intercept of the line (2.0) and the slope (+.5). To know the value for Y for a point on the line all you have to do is start with the y-intercept and then add the number you get when you multiply the slope of the line by the score on X. Expressed in the form of an equation you get:

$$Y = a + (b)(X)$$

or, in this case...

$$Y = 2 + .5X$$

So, if someone has a score on X of 8...

$$Y = 2 + (.5)(8) = 2 + 4 = 6.$$

The equation for this line tells us that when X is 8.0 Y is equal to 6. Obviously, if the slope of the line is zero the line doesn't move up or down as you go from left to right. It's a flat line. If the slope of the line is a negative number the line is going down as you go from left to right. This is the case because the change in Y is a negative number.

Equations for the slope and y-intercept of the best-fitting regression line

We know that we want to be able to write the equation for the best-fitting regression line. This will be the one line we could draw through the points in the scatterplot that run as close to those points as we can get, on average. This means that we need to know how to calculate the slope and y-intercept for this line. Below I'm going to give you the equations for both the slope and the y-intercept. They're easy to work with, so calculating these values won't be a problem. The thing that I want you to remember is that these equations represent the solution to a problem – the problem of how to draw the line that runs as close as possible to those points in the scatterplot. In the language a statistician might use, these are the equations for the slope and y-intercept of the line that minimizes the sum of squared deviations around it.

Here's the equation for the slope of the regression line:

$$b = \frac{S_Y}{S_X} (r_{xy}) = \text{Slope of the regression line}$$

Using the data from our example we end up with...

$$b = \frac{1.79}{3.16} (.884) = .566(.884) = .50$$

...a value for the slope of the line of .5.

Okay, now here's the equation for the y-intercept of the regression line.

$$a = \bar{Y} - (b)(\bar{X}) = \text{Y-intercept of the regression line.}$$

Again, using the data from our example we end up with...

$$a = 7.2 - (.5)(8.0) = 7.2 - 4.0 = 3.2$$

a value for the y-intercept of 3.2!

So, given that the slope of the best-fitting line is .5 and the slope of that line is 3.2, the equation for how to draw this line in the scatterplot is $Y' = 3.2 + .5X$.

Putting the regression equation to work

One common use for a regression equation is to use it to get predicted scores for the criterion variable (Y) when the only thing you know about a person is their score on the predictor variable (X). We now know that the equation for the regression line is $Y' = 3.2 + .5X$. So, let's say we know that a person has a score of 2.0 for social support (X). Plug 2.0 in for X, multiply it by the slope of the line, and then add the y-intercept.

$$Y' = 3.2 + (.5)(2.0) = 3.2 + 1.0 = 4.2$$

The regression equation says that when a person has a score of 2.0 for social support the best guess we can make about that person's score for life satisfaction is 4.2. And that's really all there is to it.

But how good is this best guess? Is having this best guess really worth all of the trouble it took to get it? Those are questions that the next couple of sections of this chapter will deal with.

The regression equation when there is no relationship between the predictor variable and the criterion

Here's a strange question. What would the regression equation look like if the correlation between X and Y were zero. This would be a situation where variable X isn't able to tell you anything about a person's score for Y. What would your best guesses for Y be when this is the case?

If the correlation between X and Y were zero then that's the number we'd need to plug into the equation for the slope of the line.

$$b = \frac{S_Y}{S_X} (r_{xy}) = \frac{1.79}{3.16} (0.0) = 0$$

Working the numbers through tells me that when the correlation between X and Y is zero the slope of the regression line will also be zero. In other words, the regression line will be perfectly flat.

When the slope of the line is zero what happens to the y-intercept? Plug a value of zero into the equation for the y-intercept.

$$a = \bar{Y} - (b)(\bar{X}) = 7.2 - (0)(8.0) = 7.2 - 0 = 7.2$$

When the correlation between X and Y is zero the y-intercept of the regression line will end up being the mean value for Y. This means that when the correlation between X and Y is zero the regression equation will end up as...

$$Y' = \bar{Y}$$

...which means that no matter what a person's score on X turns out to be their predicted score for Y will always be equal to the mean of Y. This is exactly the same number we said we'd use as our best guess if we'd never had X to help us! ***You can use regression when there's no relationship between the predictor variable and the criterion, but the predicted scores you end up with won't be one bit more accurate, on average, than never bothering with regression in the first place.***

Evaluating the regression line: the Standard Error of Estimate

Now we know how to get a predicted score for Y when the only thing we know about them is their score on X. Let's say we think about the predicted score for subject 1. They had a score for variable X of 4.0. If you plug that value for X into the regression equation you end up with a predicted score for Y of 5.2. For subject 1 $Y' = 5.2$. But we also know that the actual score on variable Y for subject 1 was 6.0. This means that the regression equation gave us a best guess that was .8 points too low. The difference between the

person's actual score (Y) and the predicted score using the regression line (Y') is +.8. In other words...

$$Y - Y' = 6.0 - 5.2 = +.8$$

This value of .8 represents the amount that the regression equation is off by for that person. This value is sometimes referred to as an **error of prediction**. Other authors refer to it as a **residual**. The term *error of prediction* seem pretty self-explanatory. The term "residual" makes sense as well if you look at it like this: Y' is the best guess we can make about a person's score for Y, given the information we have available to help us. But this best guess isn't perfect. Even after we've made the best guess we can make there is still something *left over* that variable X can't help us with. This left over bit is the amount that the regression equation is off by. It's the difference between the person's actual score and the predicted score, $Y - Y'$.

Okay, +.8 is the amount that the regression equation is off by for subject 1. At this point we can go ahead and figure out the amount that the regression equation is off by for *all five subjects*. In other words, plug the value of X in for each subject, get the predicted score for Y for each subject, and then calculate the difference between their actual score and their predicted score. These values are displayed in the table below.

Subject	Social Support X	Life Satisfaction Y	Y'	Residual Y - Y'
1	4	6	5.2	+0.8
2	6	5	6.2	-1.2
3	8	7	7.2	-0.2
4	10	9	8.2	+0.8
5	12	9	9.2	-0.2

$$\begin{aligned} \bar{X} &= 8.0 & \bar{Y} &= 7.2 \\ S_X &= 3.16 & S_Y &= 1.79 \\ r_{xy} &= .884 \end{aligned}$$

In the column at the far right you can see the residuals (errors of prediction). One very useful piece of information would be an average of these errors of prediction. In other words, **what's the average amount that the regression line is off by?** What would happen if you were to calculate the mean of the five errors of prediction? Well, the sum of these errors of prediction turns out to be zero! And that will be the case for any set of residuals we might run into. We're dealing with deviation scores and **the mean of any set of deviation scores will always turn out to be zero**. So calculating the mean of these deviations – these errors of prediction – will always turn out to be zero. This is not helpful. So what do we do?

One way of getting rid of the negative signs in the deviation scores is to square them. Okay, now we've got five squared errors of prediction. If we add them up we'll the sum of squared errors of prediction (or the sum of squared residuals). In the table below, we've done just that and ended up with a sum of squared deviations of 2.8.

Subject	Social Support X	Life Satisfaction Y	Y'	Residual Y - Y'	Squared Residuals (Y - Y') ²
1	4	6	5.2	+0.8	.64
2	6	5	6.2	-1.2	1.44
3	8	7	7.2	-0.2	.04
4	10	9	8.2	+0.8	.64
5	12	9	9.2	-0.2	.04
					2.8
$\bar{X} = 8.0$		$\bar{Y} = 7.2$			
$S_X = 3.16$		$S_Y = 1.79$			
$r_{xy} = .884$					

The equation for calculating this sum of squares is...

$$\sum(Y - Y')^2 = 2.8$$

The sum of squared errors of prediction is one way of measuring the variability of the points in the scatterplot (the location of actual scores for Y) from the regression line (the location of predicted scores for Y). But we wanted to get an average of our errors of prediction. What do we have to do now? We've got the sum of squared errors of prediction. If you take a sum of squares and divide by the number of degrees of freedom you get the mean of those squared deviations. You get a variance. In this case, because we're dealing with two variables, the number of degrees of freedom is going to be $N - 2$. So if we take our sum of squares of 2.8 and divide it by 3 ($N - 2 = 5 - 2 = 3$) we'll get the average squared deviation from the regression line. Here it is.

$$\frac{\sum(Y - Y')^2}{N - 2} = \frac{2.8}{3} = .93$$

The mean of the five squared errors of prediction is .93. But who wants to think in terms of squared errors of prediction? *We want an average of the original errors of prediction. What do we need to do to get that?* Well, we squared the errors of prediction originally to get rid of the negative signs. That gave us numbers in squared deviation units. We can go back to the original units of regular old deviations by starting with the variance of .93 we just calculated and taking the square root of that number! The square root of .93 is .97. ***This value of .97 represents an average of the original amounts that the regression line is off by in giving you predicted scores for Y.*** It's an average of the five original errors of

prediction. The name for this statistic is the **standard error of estimate**. The symbol for the standard error of estimate is $S_{Y'}$.

$$S_{Y'} = \sqrt{\frac{\sum(Y - Y')^2}{N - 2}} = \sqrt{\frac{2.8}{3}} = \sqrt{.93} = .97$$

Okay, the standard error of estimate is .97. Is that good or bad or what? We need something to compare this number to. One thing you can do is to compare the standard error of estimate to the standard deviation for Y.

The standard deviation for Y (1.79 in our example) represents the largest possible value that the standard error of estimate could take. This is so because the mean for Y (7.2 in this example) is the best guess you can make for someone's score for Y when the correlation between X and Y is zero. The average deviation from this best guess – the mean of Y – is the standard deviation for Y. So, when X is completely worthless as a predictor predicted scores for Y will be wrong, on average, by whatever number we've got for the standard deviation for Y. In our example, the worst case scenario would be to have an average error of prediction of 1.79. Our regression equation allows us to reduce our average error of prediction to a noticeably smaller number – .97. The regression equation has allowed us to chop our average error of prediction by almost half. This seems pretty good.

Largest and smallest values for the standard error of estimate

What are the upper and lower limits for the standard error of estimate? In other words, what are the largest and smallest values that the standard error of estimate could take? To make it a little easier to think about this question I'm going to give you a second equation for the standard error of estimate.

$$S_{Y'} = S_Y \sqrt{1 - r^2}$$

What's the lowest number that the standard error of estimate could possibly be? This is really the same question as asking when the predicted values for Y would be as accurate as they could be. The answer to that is when the relationship between X and Y is as strong as it could be. This would be when the correlation between X and Y is perfect -- +1.0 or -1.0. *Okay, so what would the scatterplot look like if the relationship between X and Y were perfect?* It would mean that the points in the scatterplot would fall on a perfectly straight line. This means that the regression equation would tell us to draw a line right through all of the points in the scatterplot. If the line goes right through every point in the scatterplot then there will never be any difference between a person's actual score for Y and the predicted score that falls on the line. If the difference between actual

scores and predicted scores is always zero then the average error of prediction will also be zero. So, **when the correlation between X and Y is +1.0 or -1.0 the standard of estimate will be zero.** If you plug a value of 1.0 in for correlation in the equation above, you'll end up with a value of zero for the standard error of estimate.

Okay, *what's the largest number you could have for the standard error of estimate?* When would the regression equation give you the least accurate predicted values for Y – when X tells you absolutely nothing about scores for Y. We just talked about this in the last section. This would be the situation when there is no relationship at all between X and Y, when the correlation between X and Y is zero. When the correlation between X and Y is zero X is completely worthless as a predictor for Y. A completely worthless predictor isn't any better than never having X as a predictor in the first place. *What would your best guess for Y be if you didn't have X to help you?* It would be the mean score for Y. *When your best guess for Y is the mean of Y how far off, on average, will your best guesses be?* It would be the standard deviation of raw scores for Y. **When the correlation between X and Y is zero the standard error of estimate will be the same number as the standard deviation for the raw scores for Y.**

Significance of the regression equation

The standard error of estimate can show you how far off your predicted scores for Y are going to be off by, on average. But it doesn't give you any information about whether your best guesses using regression (values for Y') are significantly better than best guesses when you can't use regression (the mean of Y). In a sense, this question is asking whether predicted scores for Y that we get from the regression equation are significantly better than worthless. That doesn't sound like a terribly high standard to have to meet, but if you can't even be assured of that then generating a regression equation is pretty much a waste of your time.

Are predicted scores for Y significantly better than worthless? This is really the same question as asking whether the predictor variable (X) is able to account for a significant amount of variability in scores for the criterion (Y). Let's think about it in those terms.

If you want to know whether X accounts for a significantly amount of variability in Y you first have to know how much variability there is to account for. We already know how to measure the amount of variability there is in the scores for a single variable. We can calculate a sum of squared deviations for that variable. Basically, all we need to do is to calculate the deviation of each person's raw score for Y from the mean of Y. Then we square each of these deviations. Then we add all of these squared deviations up. Here's what it looks like for our example.

Subject	Social Support X	Life Satisfaction Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
1	4	6 - 7.2	-1.2	1.44
2	6	5 - 7.2	-2.2	4.84
3	8	7 - 7.2	-0.2	0.04
4	10	9 - 7.2	+1.8	3.24
5	12	9 - 7.2	+1.8	3.24

	$\bar{X} = 8.0$	$\bar{Y} = 7.2$		12.8
	$S_X = 3.16$	$S_Y = 1.79$		
		$r_{xy} = .884$		

The sum of squared deviations above of 12.8 represents all of the variability there is for variable Y that needs to be accounted for. It's based on deviations of raw scores for Y from their mean. This is a **sum of squares total** in the same sense that we talked about it in the One-Way ANOVA chapter. Here's the equation for this sum of squares:

$$\sum(Y - \bar{Y})^2 = 12.8$$

Okay, **12.8 is all the variability there is for variable Y, in sum of squares units, that needs to be accounted for.** Out of that total amount of variability of 12.8 that needs to be accounted for, how much can the predictor variable X actually account for? We need a sum of squares accounted-for. Every sum of squared deviations is based on a deviation. So what deviation should the sum of squares accounted for be based on? Here's one way of thinking about it.

Let's take a look at subject 2. Subject 2 has a score on X of 6 and an actual score for Y of 5. If, for some reason you didn't have access to this person's score for X what would your best guess for Y have to be. We've already said that it would have to be the mean score for Y of 7.2. Okay, now you know that subject 2 has a score on X of 6 and you plug that value into the regression equation ($Y' = 3.2 + .5X$). When X is equal to 6, Y' turns out to be 6.2. In other words, on average, when people have a score of 6 on X people have a score of 6.2 on Y. Here's the thing. What's your best guess when you don't have X to help you – 7.2. For subject 2 what's your best guess when you **do** have X to help you – 6.2. How much better is your best guess when you have X to help you compared to when you don't? That this difference between the predicted score for Y (Y') and the mean of Y (M_Y). For subject 2...

$$Y' - \bar{Y} = 6.2 - 7.2 = -1.0$$

This tells us that when we've got variable X to help us, **when a person has a score of 6 on X our best guesses are going to be one point closer, on average, to their actual scores on Y.** This is a deviation that we can attribute to variable X. The regression equation tells us that when a person has **a score of 6 for X** that our best guess for Y ought

to be one point lower than the mean for Y. *The deviation between a person's predicted score for Y and the mean of Y represents how much better our best guess is using the regression equation compared to when we can't use it.* So the deviation that X is able to account for in a person's score for Y is equal to $Y' - \bar{Y}$.

All right, let's compute this deviation account-for for each person. Then all we need to do is to square these deviations and then add them up. Here it is.

Subject	Social Support X	Life Satisfaction Y	Y'	Y' - \bar{Y}	(Y' - \bar{Y}) ²
1	4	6	5.2 - 7.2	-2.0	4
2	6	5	6.2 - 7.2	-1.0	1
3	8	7	7.2 - 7.2	0.0	0
4	10	9	8.2 - 7.2	+1.0	1
5	12	9	9.2 - 7.2	+2.0	4
				$\bar{X} = 8.0$	$\bar{Y} = 7.2$
				$S_X = 3.16$	$S_Y = 1.79$
				$r_{xy} = .884$	
					10

The table above shows us that the sum of squared deviations that are accounted-for is equal to 10. The equation for the sum of squares accounted-for is presented below. By the way, instead of referring to this sum of squares as the sum of squares accounted-for, in the context of regression statisticians often refer to this sum of squares as the **sum of squares regression**. They use this term because this represents the amount of variability accounted for *by the regression line*.

$$\text{Sum of Squares Regression} = \sum(Y' - \bar{Y})^2 = 10$$

The previous paragraph indicates that our predictor variable is able to account for 10 units out of the total of 12.8. Based on this information we can calculate the proportion of variability accounted for by X. This proportion will come from taking the amount of variability we're able to account for (sum of squares of 10) and dividing it by the total amount we needed to account for (sum of squares of 12.8). When we do this we get .78. Multiplying this number by 100 tells us that variable X accounts for 78% of the variability in variable Y. This seems like a pretty useful piece of information to have. What do you think it's called. We've actually already calculated this number in Chapter 5. Remember, .78 is the proportion of variability in Y accounted for by X. This means that variable X overlaps with variable Y by 78%. This is the same kind of information that a squared correlation gives us – and that's exactly what it is. We can use the sum of squares regression and the sum of squares total to calculate the squared correlation between X and Y. Here are the calculations...

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = \frac{10}{12.8} = .78$$

Okay. We've got a sum of squares total and a sum of squares accounted-for. What do you think the sum of squares not-accounted-for is going to turn out to be? If the pattern holds the same as when we did this same kind of thing in the context of ANOVA we ought to get a sum of squares not-accounted-for of 2.8. Obviously, this is what we ought to get if the SS_{Total} is equal to the $SS_{\text{Accounted-For}}$ plus the $SS_{\text{Not-Accounted-For}}$. In other words, 12.8 will be equal to $10 + 2.8$.

We've actually already calculated this sum of squares not-accounted-for. A deviation not-accounted-for is that part of a person's actual score for Y that a score for X is not able to get at. In other words, knowing a person's score for X can get you closer to a person's actual score for Y, but there is probably still going to be something left over that X cannot help you with. This bit that's left over is the residual we've already talked about. It's the difference between a person's actual score for Y and their predicted score for Y or $Y - Y'$. The deviation between the actual score for Y and the predicted score for Y is the deviation that cannot be accounted for by the predictor variable X.

Below, I've shown what happens when we calculate $Y - Y'$ for each subject, square these deviations, and then add these squared deviations up.

Subject	Social Support X	Life Satisfaction Y	Residual Y'	Y - Y'	Squared Residuals (Y - Y') ²
1	4	6	5.2	+0.8	.64
2	6	5	6.2	-1.2	1.44
3	8	7	7.2	-0.2	.04
4	10	9	8.2	+0.8	.64
5	12	9	9.2	-0.2	.04
		$\bar{X} = 8.0$	$\bar{Y} = 7.2$		
		$S_X = 3.16$	$S_Y = 1.79$		
		$r_{xy} = .884$			
					2.8

Again, the equation for calculating this sum of squares is presented below. This sum of squares not-accounted-for of 2.8 is also known as the **sum of squares residual** because it's the amount of variability that's left over after the predictor variable has accounted for as much as it can.

$$SS_{\text{Residual}} = \sum(Y - Y')^2 = 2.8$$

Okay, we've got three sums of squares flying around here: SS_{Total} , $SS_{\text{Regression}}$, and the SS_{Residual} . We know that sum of squares total is equal to the sum of squares accounted-for plus the sum of squares not-accounted-for.

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}$$

$$\sum(Y - \bar{Y})^2 = \sum(Y' - \bar{Y})^2 + \sum(Y - Y')^2$$

$$12.8 = 10 + 2.8$$

We still want to conduct a test of whether X accounts for a significant amount of variability in Y. We've got the same information here that we had when we tested the overall effect of an independent variable in the context of one-way ANOVA. We've got a sum of squares accounted-for and a sum of squares not-accounted-for. We can put them in the same kind of table we generated for ANOVA and calculate an F-ratio. This F-ratio will tell us how many times larger the variance accounted-for by the predictor variable is than the variance not-accounted-for by the predictor variable. Here's a start:

Source	SS	df	MS	F
Regression	10			
Residual	2.8			

We need to take the sum of squares in each row and divide it by the appropriate number of degrees of freedom. After we've done that we'll take the Mean Square Regression and divide it by the Mean Square Residual to get an F-ratio. Then we'll compare this observed value for F to the critical value for F that corresponds to the degrees of freedom used in the numerator and the denominator. No problem. But we've first got to talk about how to figure out the degrees of freedom for each row.

The number of degrees of freedom for the regression row is equal to the number of predictor variables minus one. I'm going to use the symbol k to represent the number of predictor variables. Using this scheme the $df_{\text{Regression}}$ is equal to $k - 1$. In this example we've got one predictor variable so we've got one degree of freedom in the Regression row.

The number of degrees of freedom for the residual row is equal to the number of subjects minus the number of predictor variables minus 1. In symbol form the equation for calculating the df_{Residual} is $N - k - 1$. In this example, we've got five subjects so we end up with $5 - 1 - 1$ or 3 degrees of freedom for the residual row.

The remainder of the calculations for calculating our F-ratio are shown below.

Source	SS	df	MS	F	F _{Critical}
Regression	10	1	10	10.75	10.13
Residual	2.8	3	.93		

$df_{\text{Regression}} = k = 1$
 $df_{\text{Residual}} = N - k - 1 = 5 - 1 - 1 = 3$

The observed value for F is 10.75 which indicates that the variance accounted-for by X is 10.75 times larger than the variance not-accounted-for by variable X. The critical value with 1 degree of freedom in the numerator and 3 degrees of freedom in the denominator is 10.13. The observed value for F is greater than the critical value for F so we can conclude that ‘Scores for Social Support account for a significant amount of variability in scores for Life Satisfaction, $F(1, 3) = 10.75, p < .05$.’

Standardized regression coefficients

On last thing about regression with one predictor variable. What would the regression equation be if the scores we were given for X and Y were in standard score units? Some researchers prefer to think about the regression equation in standard score units rather than in raw score units. If we were given Z-scores for X and Z-scores for Y and asked to produce the regression equation used to generate predicted standard scores for Y, what would we get for the slope of the line? For the Y-intercept? Let’s look at the equations.

First, let’s start with the slope of the line. Here’s the equation

$$b = \frac{S_Y}{S_X} (r_{xy}) = \text{Slope of the regression line}$$

Remember, we’re dealing with standard scores. What’s the standard deviation of any set of standard scores? 1.0. This means that the standard deviation of the scores we’re dealing with for variable Y is going to be 1.0. It also means that the standard deviation of the scores we’re dealing with for variable X is 1.0. Plug these numbers into the equation.

$$b = \frac{1.0}{1.0} (r_{xy}) = \text{Slope of the regression line}$$

The value for the slope of the line ends up as the same number as the correlation between the two variables!

Now we'll look at the equation for the Y-intercept:

$$a = \bar{Y} - (b)(\bar{X}) = \text{Y-intercept of the regression line.}$$

What's the mean of any set of standard scores? Zero! So the means for the scores for both variables X and Y are going to be zero. Plug these numbers into the equation and see what happens.

$$a = 0 - (b)(0) = \text{Y-intercept of the regression line.}$$

Zero multiplied by zero equals zero. This tells us that if we're working in standard score units the Y-intercept of the regression line will always be zero! Essentially, the Y-intercept drops out of the equation.

What we're left with is...

$$\text{Predicted standard score for Y} = r_{xy}(Z_X)$$

Working in standard score units means that there is only one regression coefficient instead of two. It also means that the one regression coefficient – the slope of the line – is in units that researchers are already very familiar with – the units of a correlation coefficient.