Calculating A Simple Linear Regression

Now that correlations are fresh in our minds, let's learn about simple linear regressions. Simple linear regressions are, in a sense, based on correlations. You'll soon see that the steps for calculating a simple linear regression are very similar to the steps for calculating a correlation coefficient. Regressions are powerful and go beyond correlations, however, in that they allow you to **make predictions** that go beyond the data you have. Let's go through an example to illustrate some of these ideas.

Let's say that you want to be able to make predictions about how good a person will be at the popular video game "Fortnite" on the basis of how many hours a day they play the game. It stands to reason that, the more someone plays the game, the better they will be. Let's measure play time in hours per day. We'll measure ability using the metric common to shooter video games, "kd," which is a person's kill-to-death ratio in the game. If a person as a kd of 2.00, for example, this means they get 2 kills for every 1 death. A person with a kd of 1.00 gets, on average, 1 kill each time they die. Obviously, a higher kd means a greater level of skill and ability in the video game.

Keeping this in mind, let's poll 12 different Fortnite players on (1) how many hours each day they play Fortnite, and (2) their kd. Here's the data we'll be working with:

Hours	kd
3	0.8
1	0.5
2	0.7
5	1.5
4	1.2
3	1.0
2	0.8
1	0.6
9	3.0
3	0.4
2	0.7
10	4.5

As a reference, I'll go ahead and tell you that the correlation between these two variables is extremely strong and positive: 0.95. This makes sense, right? Again, the more you play the game, the better you should be at it.

But, again, let's say we want to know more than just the simple correlation between the two variables. Here are two things a simple linear regression would allow us to do:

1. Make predictions about how good someone will be at the game on the basis of how many hours per day they play the game.
2. Know how much better you'd be at the game for each additional hour you play per day.

This second bit is really useful to know. Think about it—if you're interested in getting better at Fortnite, you might want to know how useful each additional hour of practice per day would be. If you find that playing an additional hour of Fortnite each day only improves your kd by 0.05 (almost nothing), you may feel that it's not quite worth it to put in this extra time. If instead, however, you find that each additional hour of play improves your kd by, say, 0.25 or 0.50, this might make it worth your time.

Let's see how we might go about calculating a simple linear regression on these data. First, the overall formula for the simple linear regression is as follows:

Simple linear regression

$$\hat{y} = bX + a$$

In this formula, $\hat{y}$ is the **predicted value**. In this example, $\hat{y}$ would be a person's predicted kd on the basis of X, which is how many hours a day they play Fortnite.

The other two terms, b and a, are the slope and the y-intercept, respectively. These values are important because calculating a simple linear regression is calculating a *"line of best fit"* (i.e., a line in a scatterplot of data that best describes the relationship between the two variables). Knowing the slope and the y-intercept allow you to know where the line of best fit exists in the x-y coordinate plane.

First, let's talk about the **slope**. The slope literally describes how steep (or not steep) the line is. You can interpret the slope as the effect that X has on Y. If the regression line has a large slope, this means that X has a large effect on Y (e.g., playing more Fortnite has a large effect on how good you are at Fortnite). If the regression line has a small slope, however, this means that X has a small effect on Y. A slope of 0 is a flat line, which means X has no effect on Y (i.e., Y does not vary on the basis of X). Here's the formula for the slope, b:

Slope

$$b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

Next, let's talk about the **y-intercept**. The y-intercept is the only other thing you need in order to know where this line exists on the x-y coordinate plane. The y-intercept tells you where the regression line hits the y-axis (i.e., when $X = 0$, this is what Y equals). The formula for the y-intercept, a, is as follows:

Intercept

$$a = \frac{\Sigma Y - b\Sigma X}{n}$$

These formulas might look ugly, but take a closer look: You'll notice that these formulas only require the same terms as the correlation coefficient! Specifically, you'll need to calculate the following:

$$n \qquad \Sigma X \qquad \Sigma Y \qquad \Sigma X^2 \qquad \Sigma XY$$

Let's go ahead and do that now. If you'd like more detail on how to calculate these values, take another look at the Correlations *Guide*!

Hours	kd	x_squared	xy
3	0.8	9	2.4
1	0.5	1	0.5
2	0.7	4	1.4
5	1.5	25	7.5
4	1.2	16	4.8
3	1.0	9	3.0
2	0.8	4	1.6
1	0.6	1	0.6
9	3.0	81	27.0
3	0.4	9	1.2
2	0.7	4	1.4
10	4.5	100	45.0

Now that we've squared the x values (Hours) and taken the product of our x (Hours) and y (kd) values, we're ready to take the sum of each column. This gives us the values we need for these formulas. In this case, $\Sigma X = 45$, $\Sigma Y = 15.7$, $\Sigma X^2 = 263$, and $\Sigma XY = 96.4$. We also know that $n = 12$, since we have 12 participants in this particular sample. We're ready to go! Let's start plugging into our formulas.

First, let's calculate the slope, b. It's important to start by calculating b, because you need to know b in order to find a; take a look at the formulas above if you're unsure why this is.

$$b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} = \frac{96.4 - \frac{(45)(15.7)}{12}}{263 - \frac{(45)^2}{12}} = \frac{96.4 - 58.9}{263 - 168.8} = 0.398$$

Great! Now that we have b, let's use it to calculate a.

$$a = \frac{\Sigma Y - b\Sigma X}{n} = \frac{15.7 - (0.398)(45)}{12} = -0.184$$

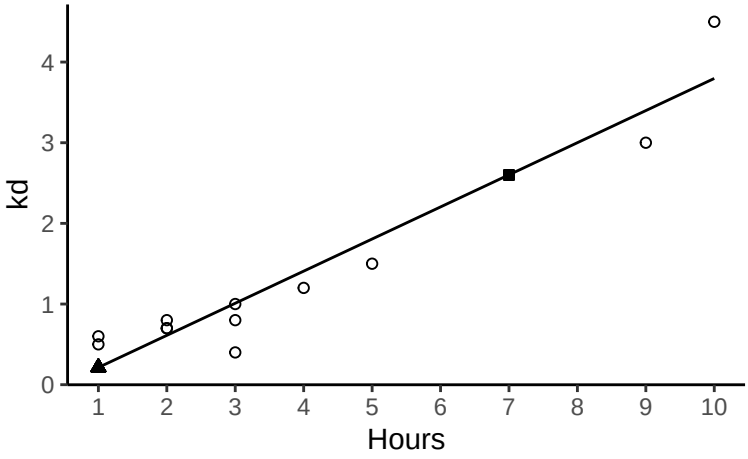
Okay, so if we plug b and a into our formula for the simple linear regression, we get the following:

$$\hat{y} = 0.398X - 0.184$$

That's our final answer! Let's take a moment to interpret what this formula for the regression line really means.

First of all, a slope of $b = 0.398$ means that, for each additional hour of Fortnite played per day, we can expect someone's kd to improve by 0.398.

Next, let's plug in various X values to really understand how this formula works. Let's say that someone plays 1 hour of Fortnite per day. Well, we can predict their kd as follows: $\hat{y} = 0.398(1) - 0.184 = 0.214$ (see the `triangle` shape on the graph below). That's not very good, as it means this person would get only 1 kill for every 5 deaths in the game! Now, let's say that someone plays 7 hours of Fortnite a day. Here we can predict their kd as $\hat{y} = (0.398)(7) - 0.184 = 2.6$ (see the `square` shape on the graph below). So we predict that this player would get (on average) 2.6 kills for each death. Much better. :)



And that's it! Simple linear regressions may be a bit challenging to understand at first, but they're so useful that they're worth the time and effort to master! Try to think about the potential applications here. Being able to make predictions about, really, anything, is a super useful super power to have.