Hi. This tutorial covers residuals. Let's take a look at a data set.

So this is a bivariate data set. And what it's measuring is several locations around the world, the altitude and then the mean air temperature. So the altitudes are measured in meters. And the mean air temperature, those are measured in degrees Celsius.

So let's take a look at that data graphed. So let's see, first of all, on the x-axis, we have altitude. And on the y-axis, we have mean air temperature in degrees Celsius. We can see that there is a very strong negative association. So as altitude increases, temperature decreases. So very, very strong.

And what I've done here is put a line of best fit here. And this is the equation for that line of best fit. This is actually rounded. If you look at it here, I've reproduced it here with less rounding. So the equation is y-hat equals negative 0.0064997x plus 14.99724. OK. So now, recall that this best fit line can be used to predict a mean air temperature for a given altitude. So let's determine the mean air temperature for an altitude of 500 meters.

So what you can do is you can estimate it from the graphed line. So you could actually mark where 5,000 is, trace that down on the graph and get an estimate. But since we have the equation, we can use the equation, by substituting 5,000 in for x here.

So I'm going to do that on the calculator. So negative 0.0064997 now times x. x I'm going to to let be 5,000, my given altitude. Then, plus 14.99724. And that gives me about negative 17.501. So negative 17.501. So at an altitude of 5,000 meters, the line of best fit predicts a mean air temperature of about negative 17.5 degrees C.

So that was a prediction for an x value that wasn't part of an observed value. So let's now test how well the line makes predictions. So this point, 920, 9.00, that represented the altitude and mean air temperature of Arthur's Pass.

So this now is my x value. This is my y value. These are both observed values. So now what y-hat value does the line give for x equals 920? When we're talking about y-hat, again, we're talking about a predicted y value. And that's going to, again, come from the equation.

So what I'm going to do is the same thing, but I'm going to put a 920 in there. So I'm just going to type in my equation. Now, I'm going to multiply by 920 plus 14.99724. Hit Enter there. And it gives me about 9.018, 1 if I round it.

So my predicted point is going to be 920, 9.018. So I can see that my prediction is pretty accurate. It was off by a little bit, but not by a lot. Difference of about 0.018 degrees Celsius. So what that error is it's called the residual. So the residual is the error associated with a predicted y value.

Now, the formula is y minus y-hat. So your residual equals y minus y-hat. The observed y value minus the predicted y value. If a residual is positive, the line under-predicted y. If the residual is negative, the line over-predicted y. So if I think about the residual for the point I just calculated, the residual is going to equal my observed y value, which was 9, minus my predicted y value, which was 9.018. So that residual is negative 0.018. So in this case, my residual's negative. So the line over-predicted y. And we can see that. Our predicted y is bigger than our observed y.

Now, another thing that residuals can be used for is to make what's called a residual plot. And a residual plot can be used to evaluate the fit of the best fit line. An effective best fit line should have a residual plot that shows no pattern. So let's look at the residual plot for the altitude and temperature data.

So basically what a residual plot does is you're going to make your graph and on the x-axis, you're just going to have your x values. But on your y-axis you're going to have your residuals. So instead of this just being the observed y values, it's going to be the y minus y-hat values.

So in order to make a residual plot though, we need residuals. So I've calculated all of these ahead of time. Notice, for Arthur's Pass, the 920, 9.00. Notice here was our y-hat value. And here was the residual associated with that. So basically, I just did all of that for all of the places here.

And if we think about each of these, those are our x and y values. Our predicted y values, these are our y-hat values and the residuals are our y minus y-hat values. So what I'm going to do now on my residual plot is compare the x values to the residual values. And I made that graph ahead of time also. Here is the residual plot.

Notice altitude is on the x-axis. Residuals are on the y-axis. And again, if the line is a good fit, we don't want to have really any pattern. And I don't really see any strong patterns here. I see a point here, which had a very large residual. But the rest the points seem to be in kind of a random scatter.

So I would say that this is a fairly random residual plot, which gives us evidence to show that our best fit line was a good line. So the previous residual plot showed no pattern. So it seems that the best fit line appropriately models the data.

If the residual plot showed a pattern, like a heavy curve, or showed uneven variation, a linear model might not be the best choice. So what those might look like is maybe you would have a residual plot that kind of showed a curve. If you're seeing a curve on the residual plot, chances are, your data values were also kind of curved. So it might be more appropriate to fit a curve rather than a straight line.

You also want to be watchful for uneven variation. So if your residuals start pretty small, but the variation gets a lot wider as x increases, so I'll say that this is your residual plot, really widens out like that. That, again, might be an indication that a line is not the best way to model the data. So these are both types of residual plots to be aware of.

So we've looked at how to calculate residuals, also how to make a residual plot for assessing a line of best fit. So that has been your tutorial on residuals. Thanks for watching.