Slide 1 – Title Slide

Hello and welcome to Week 6, Part 3 of EGM101: Linear Regression. In this lesson, we'll talk about how we can estimate the relationship between two variables using a process called linear regression.

Slide 2 – Refresher: Lines

Before we jump into linear regression, though, we should talk about what we mean by "linear" or "line." On the graph here, we see that we have values of an explanatory variable, x, graphed along the horizontal or x axis, and the values of a response variable, y, graphed along the vertical or y axis. On the actual graph, we have two points: x0, y0 and x1, y1, and we have a line connecting these two points.

The general equation for a line can be written like this, read as "y equals beta plus alpha times x." You may also see this written as y equals mx plus b, or some other variable names – the actual names don't matter, though – what matters is the form that the equation has.

We have a slope, which we've called alpha here, which tells us how quickly the line goes up or down as we move along the x axis. If we know two points on the graph, we can calculate the slope using this equation – we take the difference between the y values and divide that by the difference between the x values. The difference between the y values tells us how far up or down the line goes between the two points, and the difference between the x values tells us how far over the line goes between two points.

In other words, the slope describes the relationship between the change in x and the change in y.

Once we know the slope, we can calculate the intercept using the values of any x and y on the line using this equation, read as "beta equals y minus alpha times x."

Slide 3 – Mathematical models

In science (and yes, I'm including social science in this definition), our goal is usually to understand or explain some set of observations, and to make credible predictions based on that understanding.

One of the tools that we have at our disposal in order to meet these goals is mathematical models. In a mathematical model, we take some inputs – usually the observations that we have; then, we do some math with those observations, and we get some outputs that we can analyze in turn.

This is slightly abstract, so let's turn to a more concrete example. Let's say that we are running a pizza business, and we need to calculate the surface area of our pizzas in order to figure out how many toppings we can fit on the pizza. This is simple, right? Pizzas are usually circular, so all we have to do is measure the radius of the pizza, and use the formula for the surface area of a circle. Job done.

Is it, though? This model assumes that we're going to be putting toppings on the entire pizza, but we normally don't – we normally leave a crust where there aren't any toppings. Okay, that's fine, we just have to measure the width of the crust, and subtract it from the total radius of the pizza, and we can still calculate the surface area. Job finally done.

But is it, though? Because try as we might, we can't always make the crusts an even thickness, we don't always make pizzas perfectly round, we probably aren't going to completely cover the pizza with toppings, and we can keep coming up with ways that this really simple model won't perfectly explain everything.

But – does the model have to perfectly explain everything? Or is it enough for us to assume that the crust is an approximately even thickness and that the pizza is mostly round?

A famous aphorism, usually attributed to the statistician George Box, says that "all models are wrong, but some are useful." Our goal with modelling is not to explain everything 100% perfectly all the time. Instead, our goal with modelling is to be able to explain our observations while still understanding how and where the model is "wrong" – that is, the limitations of the model.

Slide 4 – Linear Relationships

So, now that we've discussed lines and models, let's take a step back to remind ourselves what a "linear relationship" is. And, let's say that we have some observations of an explanatory variable, and corresponding observations of a response variable, and that when we plot them in a scatter plot they look kind of like this. And if you squint hard enough, you can hopefully see that these points make a kind of line – in other words, they appear to have some kind of linear relationship.

Another way of saying this is that we can find some line with a slope and intercept that "fits" the data – it passes through, or near to, the points. With the equation for this line, we can explain the relationship between the two variables, but we can also predict unknown values, like the values of the explanatory variable in between where we actually have observations.

In other words, this linear relationship is a model for the relationship between the two variables!

Now, the biggest problem we have is that there are actually many such lines that can do this. For example, this line would seem to fit the points reasonably well, but then again, so does this one, and this one, and this one, and on and on it goes.

So, the question that comes out of this is: how do we find the "best" line that fits the data that we have?

Slide 5 – Finding a Linear Model

To answer that, we first have to figure out what we mean when we say "best". To do this, we first need to remember that the model is a prediction, and that we have plenty of observations to compare against this prediction.

The difference between the predicted value, represented by the red line here, and the observed value, represented by the black dots, is known as the residual, or error. Remember that all models are going to have some mismatch between the values predicted by the model and reality. Most of the time, the "best" model minimizes the total prediction error, or the total residual – that is, the *average* error is small, even if the residuals for a few observations might be quite large.

Slide 6 – Simple Linear Regression

So, regression is just the process of finding the “best” fitting model to the data. And, as you might guess, “linear regression” is just regression using a linear model.

It turns out that we can actually do linear regression with any number of explanatory variables – for this module, though, we will focus on simple linear regression – the case where we only have a single independent or explanatory variable.

Similar to our equation for a line, the general linear model for two variables (one explanatory variable, one response variable) looks like this equation, which should be recognizable as our equation for a line, with two small differences.

The first difference is that in this equation, we have a y with a hat on it, called “y hat”. This is the predicted value of y, which is our response or dependent variable. The other difference is that we also take into account the residuals or error between the predicted and observed values of y. The residuals are represented by the greek letter epsilon, which looks like a small upper-case E.

So, our goal is now clear: the best-fitting model is the one where the calculated values of alpha and beta, the slope and the intercept of the line, give us the smallest residual value. Okay, how do we find those values?

Slide 7 – (Ordinary) Least-squares Regression

One very common solution is to minimize the sum of the squares of the differences between the observed and predicted values of y, known as “least-squares” regression. There are many different ways of doing this, but probably the most common variety is known as ordinary least-squares regression.

I won’t go through the math of how we come to this solution, but it turns out that the value of alpha that minimizes the sum of the squares of the residuals is given by this equation: alpha is the covariance of x and y, divided by the variance of x. We can actually simplify this a bit using the formula for pearson’s correlation coefficient, in which case alpha is equal to pearson’s r times the ratio of the standard deviations of y and x.

Writing this out in summation notation, it looks like this – again, you don’t need to memorize this formula, but you should at least have some idea of what it looks like, and you should remember that the covariance involves the difference between x and the mean value of x multiplied by the difference between y and the mean value of y, and that the formula for variance looks a lot like the formula for covariance.

Anyway, to find the intercept value that minimizes the sum of squares of the residuals, we use the mean values of x and y, and the value of alpha calculated using the formula above, and solve for beta in the same way that we saw previously.

As with all things mathematical, we do have some caveats to keep in mind before we start doing least-squares regressions all over the place. The first is that we are assuming that there is, in fact, some kind of linear relationship between the two variables. If there isn’t, then these formulas will give us a result,

but it won't be the "best" solution. Along a similar line, least-squares regression works best when the residuals are normally-distributed.

The formulas shown here use the covariance and the variance, which we know are sensitive to extreme values – this means that least-squares regression is also sensitive to extreme values – we will see some examples where a single outlier point far away from the rest of the points in the distribution will determine the slope and intercept of the whole set of observations.

And, finally, the explanatory variables can't all have the same value – we can't fit a vertical line using this method, because in order to be able to divide by the variance, it has to be non-zero.

Slide 8 – Properties of Least-squares Regression

The big things to take away from this discussion of least-squares regression are that the slope describes the relationship between the changes in the explanatory and response variables. It tells us how much the response variable will change if we change the value of the explanatory variable.

As we saw, it depends on the covariance of x and y – if the covariance is negative, then the slope is negative (because the standard deviation is always positive). Similarly, if the covariance is positive, so is the slope.

If we have an intercept that is not equal to zero, then the line will always pass through the middle of the data (defined using the mean values of x and y). Not only that, but the sum (and therefore the mean) of the residuals is zero, and the residuals have no correlation with the explanatory variable. Remember, though, that this all requires that the assumptions described on the previous slide are valid – if they aren't, then these things won't necessarily be true.

Slide 9 – Summary

In this lesson, we've discussed how we use mathematical models to help explain and predict the "real world", with varying degrees of success.

We also discussed how our goal with linear regression is to find a model that explains the variation in y based on the variation in x, that can predict values of y based on a value of x, and that minimizes the difference between predicted and observed values – in other words, the best model is the one that has the smallest possible error.

And finally, we discussed how least-squares is probably the most commonly-used way of finding the "best" model, but it is not the only way. But that's a story for another time.

Slide 10 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapters 12.1 to 12.3; Caswell, Chapters 9.1 to 9.4; and Weiss, Chapters 14.1 to 14.3.

I've also included a link to an article about the seven classical assumptions of ordinary least squares linear regression, which goes into more detail about the things that need to be true about our data in

order to be able to use ordinary least squares regression correctly. Finally, there are links to two videos from Khan Academy about residuals and least squares regression, and a worked example of calculating a regression line.

That’s all for this lesson – I hope you found it interesting, and if you have any questions, please don’t hesitate to e-mail me or post in the discussion forum on blackboard. Bye!