Slide 1 – Title Slide

Hello and welcome to Week 6, Part 6 of EGM101: Interpolation and Extrapolation. In this lesson, we'll discuss how we can use regression to make predictions with our data in a bit more detail.

Slide 2 – Recall: Mathematical models

Earlier this week, I showed this slide to introduce the idea of mathematical models, and to explain how we use them to help understand or explain our observations, as well as make credible predictions based on those observations.

In this lesson, we're going to focus more on the predictions part of this, and see how we can use tools like regression to make predictions.

Slide 3 – Definitions

First, we have two definitions to cover. The first is interpolation, where we estimate the unknown values of a response variable inside of the range of the observed values of the explanatory variable.

On this figure, the black dots represent our observations, and the red line is the least-squares regression line – that is, the predicted value of the response variable, based on the value of the explanatory variable.

All of the black dots indicating our observations are contained within the blue shading – this is the range of the observed values of the explanatory variable. Inside of this range, the red line represents the interpolated values of the response variable.

The opposite of interpolation is extrapolation – where we estimate unknown values of the response variable outside of the range of the observed values of the explanatory variable. In this figure, these are represented by the black shaded areas – here, the red line represents the extrapolated values of the response variable.

Slide 4 – Interpolation Example: Salmometer

Let's say that we're out fishing, and we manage to catch a large salmon. We want to know how much the fish ways, but we have a slight problem: first, we don't have a scale, and second of all, we're not allowed to take the fish home to weigh it – it's catch-and-release only fishing where we are. How can we figure out how much the fish weighs?

As we have discussed previously, it turns out that in the wild, the weight of a fish tends to depend on its length. That is, there's often a clear relationship between the length and weight of a fish.

So, we could measure the length of our fish now. Then, the next time we come out fishing, we could bring a scale and be sure to measure and weigh every fish that we catch, and work out the relationship between the length and weight of fish in this particular area. Then, using the length of the fish that we first caught, we could estimate its weight using the relationship that we found.

Obviously, this is a bit of a silly example, but it's not so far off of something that is actually done. Recognizing that anglers might want to know how much a fish weighs, but that they might not always have reliable access to a scale, regulatory agencies will sometimes provide tables so that anglers can estimate the weight of their fish using just the length.

One such example is provided by Fisheries and Oceans Canada, known as the Salmometer. This is an example of interpolation in the real-world, where we can predict or estimate unmeasured values based on a best-fit regression.

Slide 5 – The dangers of extrapolation

Looking at this figure, we see that we have a number of observations in blue, and a regression line in red. The R-squared value of the blue points is very high, meaning that the points are all fairly close to the regression line, and the model does a reasonably good job explaining the variability in the data.

When we interpolate values in between the two groups of blue points, we could, of course, find new observations that give us large errors compared to the predicted values. But, given that we have observations on both sides of the gap, and the values are generally close together in range, it's usually not a big risk.

What about over here on the other side of the data, though? The problem with extrapolation is that, unlike with interpolation, we can only ever be sure about the "shape" of the relationship within the range of our observations.

With apologies to the late Donald Rumsfeld, outside of this range, we don't know what we don't know. The relationship could be non-linear, as illustrated by this example, where additional observations show that this isn't actually a linear relationship at all.

As we will see on the next slide, extrapolating might also lead us to make some ridiculous conclusions.

Slide 6 – A classic example

In a letter to Nature, published in 2004, Tatem and all looked at the winning Olympic times in the 100 meter dash between 1900 and 2004, and noted that the decrease in the women's winning time was happening at a faster pace than the men's winning time – meaning that, at some point in the future, the women's winning time was likely to be faster than the men's winning time.

Using linear extrapolation, they concluded that by around the year 2156, the women's winning time will likely be faster than the men's winning time, based on the steeper slope of the regression line for the women's winning times.

From the spread of the data, they determined that this could happen as early as 2064, and as late as the 2788 Olympic games – assuming, of course, that humans are still holding the Olympics that far into the future.

The problem with this from a statistical standpoint is that the authors are attempting to extrapolate values far beyond the range that they have data for – 2156 is farther away than the full length of the

available data that they have – we have no reason to think that either the men's or women's winning time will continue to improve at the same rate for that long into the future. And, perhaps more importantly for this example, in a response letter published that same year, Rice notes that in 2636, the model also predicts that the winning time will be less than zero seconds – a truly remarkable achievement for not just sport, but our understanding of the laws of physics.

The point of this example is to illustrate that while it might be tempting to use our data and tools like linear regression to make predictions far into the future, in practice it's a bad idea – we don't have enough information here to make any kind of credible prediction that far outside of our observations.

Slide 7 – So when is it okay to extrapolate?

Extrapolation is not always bad, though – it depends entirely on what you are trying to model.

A reasonable example of an extrapolation might be predicting that the sun will rise tomorrow. Based on our experience, where the sun rises and sets every day, it is perfectly reasonable to assume that this pattern will continue tomorrow. We might not see it because of the clouds, but the likelihood of the sun not being there is fairly small.

On the other hand, this graph illustrates an unreasonable example of extrapolating. Just like with the example of the Olympic winning times, we have no reason to believe that the pattern shown is going to continue very far into the future, and it would be foolish to assume that it will.

In general, we should avoid extrapolating very far outside of our observations unless we have some kind of theoretical basis for our model. Our understanding of how marriage works tells us that we're not very likely to marry a new person every day, so we shouldn't assume that the observed pattern will continue in the same way.

Our theoretical understanding of astrophysics, on the other hand, tells us that the sun is not likely to go dark for another 5 billion years or so, meaning that it will most likely continue to rise tomorrow, and the day after that.

Slide 8 – Summary

In this lesson, we've discussed how part of using regression is to predict values. If we make predictions within the range of known values, we call it interpolation; outside of the range of known values, we call it extrapolation.

In general, as long as we have good data and a good model, interpolation is generally "safe" – that is, our predicted values aren't likely to be very far from the correct values.

On the other hand, we discussed how we don't know what the relationship between two variables looks like outside of the range of our observations – for this reason, extrapolation should be used sparingly, and generally only in cases where we have a solid theoretical reason for the model that we've chosen.

Slide 9 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Caswell, Chapter 9.4, and Weiss, Chapter 4.2.

There's also a link to an article about using regression analysis to make predictions, as well as links to the two articles that looked at extrapolating Olympic winning times to both predict when the women's winning time would be faster than the men's winning time, as well as when the winning times would become negative. And finally, you can find the Salmometer, provided by Fisheries and Oceans Canada, at the link shown on the slide.

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!