Slide 1 – Title Slide

Hello and welcome to Week 6, Part 2 of EGM101: Correlation. In this lesson, we'll discuss how we can measure the correlation between two variables, and also why it is not causation.

Slide 2 – Covariance

In the previous lesson, we looked at different ways that we could describe the relationship between two variables, including the direction, or sign, of the relationship and the strength or consistency of the relationship. It turns out that these two things, direction and consistency, are exactly what we can learn from the correlation between two variables.

Before we define the correlation between two variables, though, we'll start with the covariance of two variables. Remember that the variance measures the dispersion of a single variable – it tells us how far away, on average, the values of a variable are away from the mean value.

Covariance, on the other hand, is a way to measure the dispersion of two variables, and it's very closely related to the variance, as we will soon see.

Last week, we saw the formula for the sample variance, s^2, looked like this.

Instead of writing this with the term in brackets squared, though, we could instead write it like this, with the term in brackets multiplied by itself.

The formula for the covariance, written like this, shows you how these two ideas are related. For each pair of and y values, x_i and y_i, we calculate the distance, or difference, between x_i and the mean of all the values, and multiply this by the distance between y_i and the mean of all the y values.

We then take the sum of all of these products and divide by the number of pairs minus 1 – in other words, the covariance tells us how far away, on average, each pair of x,y values is away from the middle, just like the variance tells us how far away, on average, each value of is away from the middle.

The advantages of covariance are that it can tell us the direction of the relationship – the sign of the covariance (whether the covariance is positive or negative) tells us the direction of the relationship.

On the other hand, the covariance has squared units, like we saw with the variance. We also can't directly compare different variable pairs – if we have two variables with values in the thousands, and two variables with values in the single digits, the covariance for the first pair might be much larger than the second pair – but what does that actually mean? Because of this, it's harder to determine the strength of the relationship from the covariance.

Slide 3 – Pearson's Correlation Coefficient

Because of these issues, we often turn to something else. To help make the covariance of different variable pairs more comparable, we can divide the covariance by the standard deviations of and y to give us Pearson's correlation coefficient, denoted by a lowercase r.

Pearson’s correlation coefficient, or pearson’s r, has the benefits of being unitless, meaning that we can compare variables with different units, or with different magnitudes of values.

Effectively, Pearson’s r is the proportion of the dispersion of each of the two variables, which means that the value ranges from -1 to 1, which makes comparison between different pairs of variables easier.

Finally, it also tells us the direction and strength of the relationship directly. Values close to -1 or +1 indicate a strong relationship, while values close to zero indicate a weak relationship. Because the standard deviations are always positive, r has the same sign as the covariance, which means that the sign of r tells us the direction of the relationship.

Unfortunately, this does not give us everything. It only tells us about the linear relationship between variables, and like the variance and standard deviation, it can be very sensitive to outliers.

Slide 4 – Pearson’s Correlation Coefficient

As we covered on the previous slide, Pearson’s r tells us the direction of the relationship between two variables. If r is greater than 0, we have a positive association between the two variables, shown in the top row of this figure.

If r is less than 0, we have a negative association between the two variables, shown in the bottom row of this figure.

Pearson’s r also tells us the strength of the relationship. If the absolute value of r is close to 1, it means we have a strong linear relationship, as indicated in the first two columns of the top and bottom rows of this figure. And in fact, if the absolute value of r is equal to 1, it means that all the points lie on a straight line.

On the other hand, smaller values of r, with an absolute value close to 0, indicate that we have a weak linear relationship, indicated by both the middle row and the right column of this figure. In the middle row, even though the first column looks like a straight line, it’s nearly flat, which means it has a low linear correlation. You can also see that as we move toward the right, the points get more and more spread out, and look less and less like a line.

These different examples illustrate how as points become more and more spread out or scattered, the correlation decreases – but, as the horizontal line indicates, this isn’t the only cause of low correlation values.

Slide 5 – Correlation of Nonlinear Relationships

That brings us to an important point – Pearson’s r only tells us about the *linear* relationship between variables. In the examples in the top row here, we can see how the points all appear to lie very close to a straight line, and the correlation values are very close to 1 (or -1) for the two sets of points.

In the bottom row, though, we see two sets of points with low correlation values. On the left, this is because the points are spread out, seemingly at random. But on the right, the points all appear to lie on a line – just not on a straight line. This is because the relationship or association between the

explanatory and response variable in this plot is nonlinear – there is still a relationship, even though the value of r is low.

Slide 6 – Spearman's Rank Correlation Coefficient

The other commonly-encountered correlation coefficient is Spearman's rank correlation coefficient. Instead of using the values of the data, Spearman's rank uses the difference in rank of pairs of data. In other words, we rank the values of each variable from largest to smallest, like this. Here, we see that variable has values 8, 5, 2, and 7, with corresponding ranks 1, 3, 4, and 2. Variable *y* has values 3, 5, 4, and 9, with corresponding ranks 4, 2, 3, and 1.

The formula for Spearman's rank correlation coefficient is exactly the same as Pearson's correlation coefficient – it's the covariance divided by the product of the standard deviations. The difference is that we are calculating the covariance and the standard deviation of the rank values, rather than the values themselves.

If all of the rank values are unique integers – that is, none of the values are tied in rank – then we can use the approximation shown here instead of the full formula.

Just like Pearson's correlation tells us about the strength of the linear relationship between variables, Spearman's rank correlation tells us whether there is a monotonic relationship between the two variables. By "monotonic", I mean that the direction of the association doesn't change.

In the first example here, we see that as the value of the explanatory variable increases, so does the response variable – and this is always the case. It's the same for the second example, just in the opposite direction. In the third example, though, we see that as the value of the explanatory variable increases, so does the response variable, at least for a little while. But then, the response variable begins to decrease. Because the direction of the association changes here, it's a non-monotonic relationship. Just like nonlinear relationships and Pearson's correlation, Spearman's rank won't work as well for these kinds of relationships.

The advantages of using Spearman's rank are that we don't actually need to know the values of the variables – we only need to know their ranks. We can also use non-numeric data if we can rank them – in other words, if the data are ordinal. And, because outliers don't have much of an impact on the ranks of the values, it means that Spearman's rank is a good choice when we're working with data that have lots of extreme values.

The biggest drawback to using Spearman's rank correlation is when we have ties. Ties make things more difficult in part because we can't use the approximate formula shown here. But, more importantly, if we have tie values it can actually increase the correlation value, which we don't necessarily want, either. For only a few tied ranks, it's not as much of a problem, but if you have data with lots of tied ranks, there are better options out there.

Slide 7 – Pearson vs Spearman Correlation

In general, Spearman's r is a good approximation for Pearson's r, as long as we're dealing with linear relationships. In the first panel here, we can see that the values for r and r_s are basically the same.

The second panel shows an example where Spearman's actually works a lot better than Pearson's correlation – this is an example of an exponential relationship. We can see that the points all lie on some kind of line, but not a straight line. Accordingly, Pearson's r is 0.63, while Spearman's rank is 0.99 – indicating that the two variables are highly correlated, just not linearly. Logarithmic functions are another example of a type of relationship where Spearman's correlation is better suited than Pearson's.

Slide 8 – Correlation is not Causation!

I have said this several times now this week, and I will probably continue saying it. The only thing that the correlation value can tell us is whether or not there is a specific type of relationship between two variables.

It does not care which of the two variables is the explanatory or independent variable, and which is the response or dependent variable. In other words, correlation cannot tell us anything about causality – it can't say whether or not the changes in one variable are caused by the changes in the other variable.

This plot provides a great example of this – this plot shows the number of Math PhDs awarded each year between 1996 and 2008 in red, and the amount of Uranium stored at US nuclear power plants in black. You can see that the lines look very similar, which means that if we plotted them as a scatter plot, with number of math PhDs on one axis and millions of pounds of Uranium on the other, it would look like a pretty straight line. This is an example of a spurious correlation – a correlation between two things that have nothing to do with one another.

You will probably read or hear about studies that "prove" that one thing causes another on the basis of some kind of correlation. Be very skeptical when you hear claims like this, especially if there's no solid explanation for why this would be the case.

You might even be tempted to make these claims yourself – avoid the temptation.

Slide 9 – Summary

In this lesson, we've discussed how correlation helps describe the relationship between variables. Very often, we'll be using Pearson's correlation, which means that we'll be investigating the linear relationship between variables.

We've also discussed how we can calculate certain kinds of correlation using either numeric or ordinal data, and how correlation is not causation.

We have also discussed how correlation does not tell us anything about causality, because even if it looks like it might be, correlation is still not causation.

Slide 10 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 12.3; Caswell, Chapters 9.5 to 9.7, and Weiss, Chapter 14.4.

Chapter 8 of the Huff book from the recommended reading list, titled "Post Hoc Rides Again", and Chapter 4 of Bergstrom and West both provide a good further discussion of some of these topics, and why you should take claims of correlation "proving" something very skeptically.

If you want to see more fun examples of nonsense correlations, I've also linked to the website where the Math and Uranium example comes from.

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