Slide 1 – Title Slide

Hello and welcome to Week 6, Part 4 of EGM101: The Coefficient of Determination. In this lesson, we'll learn about one of the ways that we can evaluate how well our linear model fits our data.

Slide 2 – How good is my model?

So far this week, we've seen how we can assess the linear relationship between different variables using correlation, and we've seen how we can fit a linear model to observations, a process called linear regression. And we've also discussed how this might not always be appropriate. Here, we see some data that are very weakly linearly related, with a "best-fit" line that seems less obvious than drawing a constellation.

The question we want to answer now is, how can we evaluate this? How can we assess how well the model fits to the data?

Slide 3 – Variability

To answer this question, we'll start by looking at something called the variability of each individual y value, which is just the difference between that value and the mean value of y.

We can think about the variability of y as having two different components: the first part is the difference between the predicted value and the mean, which we can think of as the part of the variability explained by the model.

The second part is the difference between the observed value and the predicted value, which you should remember is the residual. We can think of this as the part of the variability that is unexplained by the model.

Last week, while learning about the dispersion of a variable, we noted that the sum of all of the differences between the values of the variable and the mean is equal to zero. Similarly, in the previous lesson, we saw that the sum of the residuals for our regression line is also equal to zero.

So, to calculate the total variability, we can't just sum up the differences. Instead, we use the sum of squares, like we saw for the variance and standard deviation.

What this equation shows us is that the total variability, or the sum of the squares of each individual variability value, is equal to the total variability explained by the model, plus the total variability that isn't explained by the model.

Slide 4 – Coefficient of Determination (general)

Rather than writing out each of these terms, though, we can simplify the notation a bit.

First, we write the total variability as SST, for “total sum of squares”; we write the model variability as SSR, for “regression sum of squares”, and finally, we write the residual variability as SSE, for “error sum of squares”.

The statistic that we most frequently use to evaluate how well our linear model “explains” the data is the coefficient of determination, also written as R-squared. The R-squared value is the proportion of total variability that is accounted for by the model – in other words, it’s the variability explained by the model, divided by the total variability. And, using this equation here, we can also write this as 1 minus the error sum of squares divided by the total sum of squares, which is another form that you might find in a textbook.

The R-squared value is often expressed as a percent – in the absolute best-case scenario, where the model perfectly fits the data, the R-squared value is equal to 1, or 100%.

In the worst-case, or “baseline” scenario, where the predicted value is equal to the mean value, the R-squared value is equal to zero – none of the variability is explained by the model, and all of it is equal to the residual or error variability.

As a final note, despite the name, it is actually possible to have an R-squared value less than zero, but those are special cases that we’re not going to worry about here.

Slide 5 – Coefficient of Determination (simple case)

In the case of simple linear regression, where we have only a single explanatory variable, we can calculate the R-squared value in a different way.

Here, the math simplifies things a bit, and the R-squared value is actually equal to the square of Pearson’s correlation coefficient. In other words, for simple linear regression, all we have to calculate is Pearson’s correlation coefficient and then square it to get the R-squared value.

This is probably the form of the coefficient of determination that you are most often going to come across, and most textbooks will pretend that it’s the only form. Fortunately, though, you know better.

Slide 6 – Interpreting the R^2 value

What does the R-squared value actually tell us, though? For starters, it tells us something about the scatter of the data points around the best-fit line. When the R-squared value is close to 1, which means that most of the points are fairly close to the best-fit line. The closer to zero that the R-squared value is, the larger the scatter of the observations around the best-fit line.

The R-squared value also tells us what part of the total variability in the data the regression model accounts for. If we have a low R-squared value, most of the variability is not explained by the model.

It’s important to remember, though, that this doesn’t actually say anything about how “good” the model is – we can have a good model that has a low R-squared value, and we can have a high R-squared value for a model that is completely unsuited for the data. On the next slide, we’ll see one of the ways that we can judge whether a linear model is appropriate for our data.

Slide 7 – High R^2 does not mean it's the right model!

One of the best ways to check if we are using a "good" model, or a model that is appropriate for the data, is to look at a plot of the residuals.

On the plot here, we see that the linear model explains 74% of the variability present in the data. If we then plot the residual value against the predicted or fitted value, we can see that the residuals are scattered around zero, indicated by the dotted line. Some of the values are positive, some are negative, but overall, there's not a clear pattern or bias apparent in the residuals. This is an example of random errors – errors were we don't see a pattern.

Now let's look at a different plot. Here, we see that our linear model explains 93% of the variability in the data – you would think that this means that this is a "better" model, right? But look at the distribution of the residuals – they have a clear pattern. This is an example of systematic bias, or non-random errors, or a pattern in the residuals. Systematic errors in the residual plot indicates variability not accounted for in the model – in other words, the model is a bad fit for the data. In fact, the data here are a parabola, which means that the linear model is not an appropriate choice.

When we see patterns in the residual plot, it's not necessarily a bad thing – what it means is that we need to find a different model to use. This could mean that we need to look for additional explanatory variables, or even to consider adding non-linear terms to the model. In the second example here, we would want to fit a parabolic model to the data.

Slide 8 – So what's a good R^2, anyway?

So that leaves us with an important, often-asked question: what is a good R-squared value, anyway? In the previous slide, we saw how a high R-squared value isn't necessarily a good thing, but is there some minimum value to determine a "good" R-squared value?

Ultimately, it depends on the context and our goal. First, we'll consider the goal of understanding the relationship between variables.

Different fields of study will have different research questions which have different levels of explainable or un-explainable variability. What defines a "good" R-squared value entirely depends on the research question, and how much of the relationship between different variables can actually be explained. Remember that the point of R-squared is to say something about the variability that's actually explained by the model. A lower R-squared value doesn't necessarily mean that the model is bad – instead, it might mean that there are additional factors or explanatory variables that you should consider.

Second, we'll consider the goal of predicting unknown values. Here, we might place more weight on higher R-squared values, because higher R-squared values tend to mean there is less variability due to error – another way of saying this is that the model predictions are more precise. If we are making predictions using a model with low R-squared values, the model predictions might not be precise enough to be meaningful.

Rather than just asking what a good R-squared value is, we might come up with some more relevant questions, such as how much of the variability in the relationship can actually be explained?

Another important consideration is whether the modelled relationship between the variables is “statistically significant” – something that we will cover more in the weeks to come. If we do find a statistically significant relationship, then the R-squared value doesn’t really matter as much.

We might also ask how precise the predictions of the model are – something that we’ll try to cover more in the next two weeks. In this case, the R-squared value can tell us something, but it’s not necessarily the best way to answer this question.

Slide 9 – Summary

In this lesson, we’ve learned about the coefficient of determination, or R-squared value. We have seen how the coefficient of determination measures the scatter of the data around the regression line, and tells us how much of the variability in the response variable is explained by the explanatory variables.

Perhaps most importantly, we’ve discussed how a high R-squared value does not actually tell us whether the model we’re using is the right model for the data – in fact, we can have a high R-squared value for a model that is completely incorrect for the data that we’re using.

And finally, we’ve discussed how interpreting or evaluating the R-squared value depends on the context and our goal – there’s not really a “one size fits all” value that we can use to say “this is a good model” and “this is a bad model”.

Slide 10 – Additional resources

You can read more about the topics we’ve discussed here in the textbooks – Illowsky and Dean, Chapter 12.3, and Weiss, Chapter 4.3.

I’ve also included a link to a Khan Academy video that also discusses the coefficient of determination, as well as three great short articles that talk about the R-squared value: how to interpret it, what a high R-squared value means, and why having a high R-squared value isn’t always the best thing.

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!