CourseKata Video Transcript

Video Details

Video Title: Partitioning Sums of Squares
Video Link: https://player.vimeo.com/video/381975052

Video Transcript

Student (off screen)
So Ji, now you're making a model with a quantitative explanatory variable. How do you quantify the error around that model?

Dr. Ji
Yeah. And so we've now been able to make all kinds of different models. We started off with the empty model, but then we moved on to a categorical explanatory variable like sex, right? And now we're thinking about a quantitative explanatory variable like height. So I want to put that kind of in context. All the models we've gone through so far, and how we've quantified error in each one of them before we get to this one, right? So here I've put our tiny fingers data right into this little sample data window,

[ROSSMAN CHANCE APPLET IS CAST ON SCREEN. IN THE SAMPLE DATA BOX THERE IS A LIST OF 6 DATA POINTS FOR “Height2Group” AND “Thumb”. SPEAKER GESTURES TO THE LIST.]

and I've put in two variables, Height2Group, whether they're just short or tall, and thumb. And you could see here this person is in the short group, 0, and their thumb length is 56, right? And so those data points have been put on this scatter plot right here, and here we have Height2Group on the x-axis. So these are all the short people, right?

[SPEAKER GESTURES TO THE DATA POINTS IN THE SHORT GROUP ON THE SCATTER PLOT]

They're lined up with 0. Here are all the tall people lined up with 1.

[GESTURES TO THE DATA POINTS IN THE TALL GROUP ON THE SCATTER PLOT]

And here are their thumb lengths, right?

[GESTURES TO SHOW THE HORIZONTAL PATH FROM A FEW DATA POINTS TO THE Y-AXIS]

And you could kind of see, in general, the shorter people tend to have shorter thumbs than the tall people, but that's not totally the case. Like for instance, this person who's in the tall group has a shorter thumb than this person who's in the short group. But in general, it is the case that, in general, these people who are taller tend to have longer thumbs. These people who are shorter tend to have shorter thumbs, right? Okay. So let's first start off with an empty model, just our mean.

[A BLUE HORIZONTAL MEAN LINE APPEARS ON THE SCATTER PLOT]

It's a model we're very comfortable with, and here's our mean, 62. And in this case, we say I don't care how tall or short you are, just gonna say, your thumb length is 62. And we could quantify how much error this simple model has by looking at the residuals, and then looking at the squared residuals, right?

[BLUE SQUARED RESIDUALS APPEAR OFF THE LINE FOR THE EMPTY MODEL]

And so the sum of squares we saw was 82, and this is the best you could do when you just have one prediction that's the same for everyone, right? 82 is as low as it gets. Now, let's see if we could improve this by adding in a categorical explanatory variable. Height2Group, which is just short and tall, right?

[BLUE MEAN LINE AND SQUARED RESIDUALS ARE REMOVED FROM THE PLOT; A POSITIVELY SLOPED RED REGRESSION LINE APPEARS]

And so here I'm gonna show this regression line, and let's think about why this line is drawn in this way. So it's really saying, look, if you're a short person, I'm gonna predict that your thumb length is 59.67, right? And that is right here.

[USES MARKER TO DRAW A POINT IN THE SHORT GROUP WHERE THE REGRESSION LINE PASSES THROUGH]

And then if you are a tall person, I'm gonna add on 4.67 millimeters to that thumb length. And so I'm gonna predict your thumb length is right here.

[DRAWS A POINT IN THE TALL GROUP WHERE THE REGRESSION LINE PASSES THROUGH]

And so here what we see is that we're now not just making one prediction that's the same for everyone, but we're making two different predictions depending on which group you're in. For short people, we're gonna predict this number; For tall people, we're gonna predict this number. And then we're gonna look at, what's my error around these new predictions, right?

[THE RED SQUARED RESIDUALS FROM THE DATA POINTS TO THE REGRESSION LINE APPEAR ON THE PLOT]

And so I could look at the squared residuals, and you could think about it as if we compare it to the blue squares, these are much smaller squares, right? And so we could quickly look at the blue squares.

[THE BLUE SQUARED RESIDUALS FROM THE EMPTY MODEL APPEAR ON THE PLOT]

You could see that the blue squares are quite a bit larger than the red squares, right? So we've done a good job because we have reduced error. We have gone from 82 which we thought it was as low as it could go with one single prediction for everyone, but now we could give people a prediction if you're short. We could give people a prediction if you're tall. And so with this slightly more complicated model, we could reduce our leftover error to 49.33.

[POINTS TO “SSE = 49.33” ON THE APPLET]

So we've done a better job.

[SPEAKER PROCEEDS TO ERASE POINT MARKS MADE BY MARKER]

Student (off screen)
Ji, why does it go down when you just make a more complicated model? Why does the error go down?

Dr. Ji
When you make the model more complicated, the error goes down because you're able to make a slightly better prediction with a little bit of information you know about which height group they're in. So now let's think if you get a little better with a little more information, what if we had better information about this person? Like in the case of quantitative explanatory variables. Now we don't just know if they're short or tall, but we know exactly how short or tall they are, right? And so I'm gonna copy and paste in our tiny fingers data set now with their actual height, not just that they're short and tall.

[ALL DATA IS CLEARED FROM THE PLOT AND REPLACED WITH A LIST OF 6 DATA POINTS FOR "Height" AND "Thumb"]

And so now we could see, this person who has a height of 62 inches, they're short, right? But this person who had a height of 66 inches, they were also considered short. But they're different kinds of shortness, right? And so we could actually see that better now. So let's start off with creating our new model. And remember our new model is gonna make a different prediction for each person depending on their height. Before, they made a different prediction for each person, depending on which group they were in based on height. Now it's precisely based on their height.

[RED REGRESSION LINE IS PLOTTED THROUGH THE DATA ON THE SCATTER PLOT OF "Thumb" BY "Height"]

So I'm gonna create my regression line. And so now you could see, it's not about means, right? It's really I take this person who's very short, 62, right? 62 inches. And I'm gonna say, we're gonna take this negative 3.16.

[POINTS TO "*Thumb*^ = -3.16 + 0.9848 x *Height*" ON THE APPLET]

That's what we're gonna guess your thumb length is if you add a height of 0, right? And so that doesn't exist. But then we're gonna add on about 1 [POINTS TO "0.9848"] for every 1 unit of height. And height is in inches, right? And so we're gonna add on a millimeter for your thumb for every 1 inch of height. And so for this person who has a 62 inch height, we're gonna multiply 62 by this number, 0.9848, right? And so we're gonna predict for this person, the prediction is now going to be this number right here, right?

[DRAWS A POINT ON THE REGRESSION LINE ABOVE ONE OF THE DATA POINTS]

Student (off screen)
So what do you mean "prediction"? Because you already know their thumb length.

Dr. Ji
I know. Isn't it so funny?

We already know their thumb length, but we're making predictions about it. What we're actually doing is we're checking our model, right? And so we're generating predictions from our model to say, how good is our model at matching the thing that we already know, right? So it's a way of checking our model. And we do that for the purpose of calculating error of our model. Because then we could quantify how good, how bad is our model. So

now let's think about how off is our model predictions from the actual data points that we have. So here I'm gonna show you the residuals.

[RESIDUALS APPEAR ON THE PLOT AS RED VERTICAL LINES FROM EACH DATA POINT TO THE REGRESSION LINE]

And so notice the idea of the residual is very much like what we saw when we saw the empty model and the categorical explanatory models. Basically, it's just whatever your prediction is, how off are you from that. We talked about how the regression line is very much like the mean because it balances the errors from both sides. The negative residuals balance out with the positive residuals, right? Now, let's take a look at the squared residuals. All we're doing is we're taking this little residual right here, and we're making a little square out of it.

[SPEAKER USES VERTICAL HANDSHAPE TO DEMONSTRATE ONE RESIDUAL THEN LAYS THE HANDSHAPE HORIZONTALLY TO SHOW THE SQUARING OF THAT RESIDUAL]

And so visually you could even kind of see these squares look a lot smaller than the squares we saw before. And so you could see that now our sum of squares has been reduced to 28. So let's see how far we've come from our empty model. I'm gonna show you the empty model just overlaid right on top of this regression model, right?

[BLUE HORIZONTAL LINE APPEARS ON THE PLOT THROUGH THE MEAN OF “Thumb”]

So here's our height model here, but here's our null or empty model right here, right? And now you could see the squared residuals off of that. These squares are much larger than the squared residuals off of the height model, right? And you could also see, because the squares are larger, the sum of squares is 82. Compare that to 28 [POINTS TO “SSE = 28.82” ON THE APPLET], we've come a long way by knowing a little more information.