Levy Clip 2 Transcript

DAN LEVY: All right, can you please vote again? Can you please vote again? OK, show you the results. OK, the results are different than last time. And I wonder if someone who changed their mind as a result of this discussion can you tell us why.

STUDENT: I initially chose D because I misunderstood the idea of r squared. I thought that r squared was to measure the goodness fate of any particular variable.

DAN LEVY: Uh-huh.

STUDENT: But then I was explained by our group mates that it actually is to explain the goodness of fate of the entire model.

DAN LEVY: Right.

STUDENT: So I changed from D to B.

DAN LEVY: Right here, right? So our model that explain.

STUDENT: Exactly.

DAN LEVY: Anyone else? So he changed his mind. Yes? Can you take the--?

STUDENT: Sure. So I first chose E.

DAN LEVY: Where's your name card?

STUDENT: It's back at home.

DAN LEVY: Back at home.

STUDENT: Yes.

DAN LEVY: Like, where home?

STUDENT: In Allston.

DAN LEVY: OK.

STUDENT: But I'll bring it next time, of course.

DAN LEVY: Go ahead.

STUDENT: So I chose E because I thought if a variable like the new one will be totally disentangled, will be something totally different than it might actually blur the score and lower the r too. But then as a result of the discussion I realized how, if this new variable will be completely inadequate and will not explain any of the slope, then the model will just ignore it and will just explain all of the slope by the first variable that we start with.

DAN LEVY: Very good. If the variable is totally relevant, say we're estimating the GDP in the US year by year and we introduce rainfall in Malaysia, you would think that that variable is not going to be very good at predicting GDP. If it's totally irrelevant, the worst thing that can happen is the model will set the beta hat for that variable at 0. Remember, OLS is trying to minimize this. So if you add a variable, the worst that can happen is you are at that minimum and set beta 2 hat to 0. Or beta x hat to 0.

So the idea here is when you add something to the regression the r square should increase or stay the same. Now in practice, when you add something to regression it's very hard for that something to be totally irrelevant. Because even by spurious correlation we might find a small, tiny relationship. So in practice, when you add something to the regression, when you add an explanatory variable, the r square will tend to go up, even by a little bit. But it would tend to go up. Faran.

STUDENT: Marla had a sort of counterexample that made me think harder about it. Say that you have a binomial relationship but you're doing an OLS. Is that an example of something that would actually marginally improve the OLS even though there's a binomial distribution between the variable that you're adding, for example?

DAN LEVY: So you're saying that is a counter-example as in the r square won't go up, or--?

STUDENT: Yeah, the r square won't go up, or that it would actually get worse. You're adding a variable that actually has, say, a binomial relationship with smoking. Would the model just ignore that, or would it actually--

DAN LEVY: If the variable is like totally useless, or even if it has that kind of relationship, the worst thing that can happen is the model sets the coefficient to 0 and then you still have the same sum of squared residuals. If the model finds a way to reduce the sum of squared residuals, then it will be able to reduce it. And hence, the r square will go up. So it cannot go down is the short answer.

Marla, he's putting words in your mouth. So I want to make sure that he represented them well.

STUDENT: That was exactly my question.

DAN LEVY: OK. And has your question been answered, or no? Clearly, no.

STUDENT: No, that is an answer to the question. No, now I'm just trying to think about if I can agree with that. Or like, I'd want to see an example of that.

DAN LEVY: Yes, OK. So I have a challenge for you. For the next week--

STUDENT: Sorry.

DAN LEVY: For the next week I'll allow you to run all the regressions you want, thousands, hundreds of thousands. And I want you to come to class and show us a regression when you added an explanatory variable the r square went down. I will save you some time, you won't find it. But you're welcome to try.

STUDENT: OK.

DAN LEVY: OK? All right. OK, so the problem with the r square, of course, is that if you just add garbage to your model, the r square will always go up. And so some people think that one better measure of the goodness of fit is to use what's called the adjusted r square. And some statistical software reports adjusted r square. And what this does is it lets you add explanatory variable power when you add any variable, but it penalizes you for every variable that you add.

In fact, the way that the adjusted r square works is if the t statistic for the explanatory variable you add is greater than 1, it will go up. And if it's less than 1, it will go down. That's kind of the arithmetic rule by which this operates. The bottom line is if you don't like the r square because it doesn't penalize you, the adjusted r square helps you with that.

So let me ask you one more question before we move away from the r square. And it links with a concept that we saw last class. So I'm going to ask you to answer it with your smartphone devices.

OK. There are enough votes here that I can already know what we're going to do. Please find someone close to you that has a different answer. It should be fairly easy to do that.