**Levy Clip 3 Transcript**

DAN LEVY: All right, so here's what I'm going to invite you to do. Go to appendix two and see the two regressions that are there. And compare the r squares of those two regressions.

So the first study is an RCT of class size and test scores. The second study is an observational study. First question to you, in the absence of any other information, which of the two studies do you think is more likely to suffer from omitted variable bias? Number one or number two? Two. Marla, why?

STUDENT: Because two is an observational study. And in number one we have the benefit of randomization. So we can just assume that on all observed and unobserved characteristics the treatment and control were the same in the first study.

DAN LEVY: Good. So, if it's well-designed and well-conducted it has that going for it. Which of the two studies has the highest r square? Marla, I'll ask you that.

STUDENT: The second one has a higher r squared.

DAN LEVY: The second one. So the second one is observational, suffers from OVB, and yet has a higher r square. So while you might debate with the exact wording of this question I want to suggest, and this is where the emotional part comes in, that the r square is not very helpful to tell us anything about the causal effect of a variable on another variable. The r square is very good to do what Kat said, which is to tell us how much is it that we understand, predicts our y variable. But it's not very good at giving us any indication of how good our model is in terms of assessing the causal effect of a variable in another one.

If you remember, the regression that we just saw right here is a regression that has an r square of 86%. That's like unheard of in the social sciences. Do you think this

regression is a good regression to run to give us a causal effect of cigarettes on lung cancer? Literally has five observations. Five countries are in this regression.

So the r square is a very helpful statistic to tell you how much of the y variable are you explaining. But it's not very helpful in terms of giving you any indication of whether you have captured the causal effect of x on y. Are you OK emotionally with what just happened in this room? You're like, dude, I don't really care about any of this.

[LAUGHTER]

OK. We're going to move to a different topic here. It's related to r square. In a second we'll see how. But here's the motivation and here's why I think you should care about this topic.

We have been obsessing about what determines beta 1 hat. That is, what determines this number here. And whether that beta 1 hat is unbiased. And whether it captures the causal effect of x1 on y. And that all is very important.

But remember that if you want to assess statistic of significance, you want to know how large that effect is relative to this number here, which is the standard error. And so what we're going to do in the rest of this class is try to understand what are the factors that determine this standard error here.

And this is important for two reasons. If you're ever in a position to commission a study understanding how you're going to do to be able to get precise estimates is going to be important, and that relates to the standard error.

And the second reason it's important is if you're consuming a study and you see that the findings are or are not statistically significant. It's helpful to understand what it is that might be driving that statistical significance. So that's our goal. And this is what we're going to do next.

But let me first ask you a question. All else equal, we would like an estimator of beta hat that has a low standard error rather than a high one. Can someone explain why? Ignacio. Ignacio, what did you think about what Kat said about physics being superior to economics?

STUDENT: I totally agree with that.

DAN LEVY: You totally agree.

[LAUGHTER]

For those of you who don't know, he has a PhD in physics. OK.

[LAUGHTER]

Now you want to be his friend, right? OK. All right, go ahead.

STUDENT: I assume because the lower the standard error the higher the t statistic, all else being equal.

DAN LEVY: OK. And why do we care about that?

STUDENT: So the more likely you're going to have a statistically significant result.

DAN LEVY: OK, good. And why do we care about that?

STUDENT: To answer what question?

DAN LEVY: Right, so you're saying the more likely we'll have a statistically significant coefficient. Why is that important to us? Help Ignacio here. He doesn't have to answer all the questions. Raphael.

STUDENT: I think if your beta is not statistically significant then you actually don't have a model because you're not explaining the variation in the dependent variable with any explanatory variations in variable. So you sort of need that to have a functional model.

DAN LEVY: OK. But what if the variable has no effect?

STUDENT: Then you have omitted variable bias. Because then you don't have the proper explanatory variables for the dependent variable if the betas are not significant.

DAN LEVY: No, but what if the beta to your variable is not significant? Suppose you're trying to assess the effect of some program on some outcome of interest and you see that program, that coefficient and the treatment variable is not statistically significant. Is that necessarily bad?

STUDENT: It's information. So it tells you that you should drop the program probably and consider another program.

DAN LEVY: OK. OK. So I want you to remember what is it that we're trying to do when we do estimation? We're trying to estimate the beta j. Beta j stands generically for any of the betas. So contrast these two situations where we have the sample and distribution of our estimator. Let's have this situation here versus this situation here.

These are two scenarios. Scenario one and scenario two. And they both depict the sampling distribution of our estimator beta j hat. Remember, what is a sample and distribution? It's telling us it's a histogram of all the possible estimates. In which world would you rather be, scenario one or scenario two? Why? You're a policy maker, why? How many samples will you draw to do your study, typically? One. So why do you want to be in scenario one? Sarur?

STUDENT: I'm saying it's because I can say with greater confidence that it's beta j.

DAN LEVY: You can say with greater confidence that it's beta j. Another way of thinking about this is in scenario one most of the estimates that we will get-- we're only going to get one, but the likelihood of getting an estimate that's pretty close to beta j is pretty high. Scenario two, the likelihood of getting an estimate that's close to beta j is much lower. And in the end, we're doing all of this because we want to know what beta j is. So we would much rather be in this scenario where most of the estimates are very, very close to the beta j than in the other one. In words that are a little bit simpler. In scenario one we're less likely to get a weird estimate, an estimate very much away from the true value of the population parameters. We're good? All right.