[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: What we are going to do today is first, finish off a little bit of testing in the linear model. So that's in particular go over the t-test, which you will always see in any regression output. Second, go over a bunch of practical considerations that you will encounter when running regressions.

So in particular, how do we deal with dummy variables, issues of functional forms, et cetera? Then putting them all together into one very popular technique that's based on the linear regression, which is the regression discontinuity design. And then if I have time at the end of today, I'll give you one more of this project preview from a student, which is a regression discontinuity design based on data that's readily available and accessible. So that's the plan for today.

On Monday, we'll talk about oncogenicity and instrumental variables, et cetera. And then soon, we'll have the two guest lecture by Sendhil Mullainathan on machine learning. Then we'll basically be almost done. The rest will be fun stuff that one can do with data like some stuff on data visualization, a little bit on conducting surveys online, that kind of stuff. So that's the plan for the rest of the semester now.

So let's talk about the t-test. It's something that Sara barely mentioned when she was talking about the various things we can do in regression. And yet you always see t-test in regression. So where do they come in? It must have some use.

So the t-test is a test where we don't use as a basis for creating critical value, the normal distribution, but instead, the t distribution. So when do we need a t distribution? You remember, maybe, that in small sample, if we knew that the errors have a normal distribution, then we also know that the betas have a normal distribution.

But so if we knew the variance, then any test based on the beta and the known variance would also have a normal distribution, but we don't know it. We need to estimate it. So the error variance is typically estimated. And so when we substitute when constructing a test, the 2 sigma squared by sigma hat square, the resulting distribution is not normal anymore. It's t.

As the sample become large enough, it really doesn't matter. The t distribution would become very, very similar to the normal distribution. More generally, we're not always willing to assume that the errors are normally distribution, not normally distributed, we don't need it to justify the OLS model. We just need that they are IID. So they don't need to be necessarily normally distribution.

So in that case, we can still do the t-test. Essentially as a way of being a bit more conservative and not assuming-- the test is not assuming that-- doesn't assume normal distribution of the error. So what's happening with the t-test is that the tails of the distribution are a little bit fatter. So the critical value will be a little bit more conservative than with the normal-- than with the normal distribution.

So how does a t-test look like? Well, by now, it should have a reasonably familiar form. If we want to test the hypothesis, for example, that a particular coefficient is c, where c might be 0 or any other constant that we know, the t-test is simply the estimated beta, beta hat minus c divided by the standard error of beta hat, where the standard error is given by sigma squared x prime x minus 1 to the square root of that.

So that's going to pick the diagonal element of the variance covariance matrix that are the relevant x squares. This is the ii here. In practice, we don't know sigma squared. So we are going to replace it by sigma squared hat. If it's a more complicated matrix, more complicated hypothesis, which has potentially several restrictions as we wrote for the more general form of hypothesis testing, then its R beta equals c. So it's-- the test then is R beta hat minus c over the SE of R beta hat. And SE of beta hat is, again, picking the diagonal, the right element, not necessarily the diagonal element, the right element of the matrix R, x prime x minus 1R prime, which is going to pick up the relevant element.

So for example, if R is-- if our hypothesis is that the 3 beta 1 plus 5 beta 2 equal to c, the R is going to be 3 for the beta 1, 5 beta c. And that's going to be-- we are going to have R here and R prime here. So it's going to pick the relevant linear combination of the x prime x minus 1 inverse. So why do we care?

Well, it's a simple hypothesis, simple linear hypothesis, just testing a particular coefficient, for example, beta j equal c, 0 being a leading square of that, the f-test and the t-test are equivalent. The t-test statistics are the square root of the f-test statistics. So it's the same. So you can use either. You can construct the f-test as we saw-- as we have done at the end of last lecture.

But it's easier to use the t-test for a single estimated coefficient because you have it right there in the coefficient for you. Sometimes it shows up in little stars. And so you don't have to worry about that. One thing where you do need a t-test, because that's not-- this is not something that you can do in the f-test formula is when you're trying to do a one-sided test.

So you're not testing 0 versus anything but 0, you're trying to test for reject, for example, h0 is the hypothesis that g is greater than 0. So it's greater than 0 versus smaller than 0. In that case, you cannot this doesn't fit nicely in the framework of an f-test. So you're going to construct a t-test for that.

So in any regression framework, you will find an f-test that's for the-- that you find in the-- in Stata it's at the very top. In R, It's at the very bottom. What is that f-test? What's that f-test is testing, the one you get for free like, how does it know which one, which one you care about? How does R know, like f statistic could be for any kind of hypothesis you might test. Why does it decide-- by default, which one does it test?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Of what?

STUDENT: The h [INAUDIBLE].

ESTHER DUFLO: Yeah, but any particular one? All of them. So it's the hypothesis that all of the coefficients are equal to zero. And then what you're getting here is-- so you're getting that for free, the f-test. And another thing you're getting for free is the t value for coefficient after coefficient.

So when you have just one coefficient, then-- if you just have one coefficient in the regression, then as I explained before, the t-test and f-test should really be the same thing. So hopefully they give you the same answer. So you should be able to check that if you take the square of that, in fact, it's going to be 15.3. So with only one single coefficient, the t-test that is plot-- that is plotted here on the graph is the square root of the estate for the entire regression.

When you have lots of regressors, the f-test is going to be that all of them are zero. And the t-test is going to be efficient by coefficient, testing that this one is zero. So this is something that is typically of interest. Both of these things are typically of interest. Now, if you're interested in something else, like two coefficients are equal, or two coefficient are equal or opposite sign, you'll have to construct those tests, either as test statistics, or t-test, or as f-test statistics. In practice, you're going to ask your software to do it for you. And they are going to do it equivalently with an f-test or a t-test.

STUDENT: So before we move on in this case, because the t-value is the 1.96, we're comfortable regression was not due to chance? That the effect was not due to chance?

ESTHER DUFLO: Well, what we know is we can reject the-- so this coefficient here is 0.2. We can reject that it's zero at the 95% level because that coefficient is, in fact-- is greater than 1.96 or whatever the critical value is for the t statistic, which is a little bigger than 1.96. Another way-- another thing we can do from these statistics is we can say, well, let me construct a confidence interval for our coefficient.

From here, if you want to construct a 95% confidence interval, how would you-- what do you-- what you're doing from-- how do you construct the confidence interval from having the coefficient? From what we have here, how do I construct the confidence interval of gss data here? Of the consequent interval of the coefficient?

STUDENT: It's a function of the effect. And then also the z-score--

ESTHER DUFLO: You have everything you need here to construct, with-- I guess, I should give you the critical value of, let's say with your critical value of 1.96, you can construct the confidence interval, you have everything you need now to construct a 95% confidence interval for that coefficient. So what would it be? Yep?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Exactly. So that's the confidence interval. What else do we have that is of interest? So we immediately know that we can reject the-- we can reject the hypothesis that this is zero. And then we can, in fact, here we have the probability that it's greater than t. We can know at what level we are rejecting the hypothesis that it's equal to zero. We know that it's quite. This is a large t-stat. So this is a very small p-value.

And that's reflected by having lots of little stars here. If you cannot read this number, which of course, involves a lot of numbers, you have the little stars. What else do we want to see in that regression that would be interesting?

STUDENT: So another quick question. Your p-values are calculated by not by Fisher, but by the other test? [INAUDIBLE].

ESTHER DUFLO: Oh no, no, there is no Neyman here. This is all based on-- I go back to the relationship between Neyman and what we get in OLS regression. This is based on the same t statistics that we have here, which involve the square root of-- the square root of n. So this is the same. There is no Neyman here. It's all the same. It's based on the same t value, which is the coefficient minus 0 in this case divided by sigma squared hat and then the square root of n, the number of observations.

I'll give you the relationship between what we did with randomized control trial, which is Neyman, how is it different from the standard t-test of significance. It's very, very related, obviously. OK. So some practical considerations that you-- now you basically have all the machinery to do all the regression that you want.

So some practical issue that you might run into, one is dummy variables. I'm going to talk today about dummy variables on the right hand side. I'm not going to talk much at all about when the dependent variable is a dummy variable, which has-- which brings a bunch of other difficulties of interpretation that have to do with the fact that the errors are certainly not normally distributed because it's either 1 or 0.

So by definition, they can not be normally distributed. So I'm not going to talk about when the y is a dummy variable. But I'm going to look at situation when the x is a dummy variable. I'm going to look at other functional form issues that we might encounter on the x or on the y's. And then one example of putting all these things together, dummy variable, functional forms, et cetera, which is a very popular design called the regression discontinuity design.

So what's a dummy variable? A dummy variable, which we also call sometimes an indicator variable, is a variable that is 1 if the observation belongs to group A and 0 if it's in group B. Examples we can think about, very popular examples, especially with me, is the RCT example, where a dummy variable could be 1 if the observation is in the treatment group and 0 if in the control group.

But you could also run regressions that have nothing to do with randomized trials, where you have dummy variables. For example, you might be interested in the gender, so the dummy could be 1 if male, 0 if female. You could run regressions that involved times. So you could be wondering about financial spread before and after the Great Depression. So you would have time series data. And 1 would be after the Great Depression, 0-- 1 before the Great Depression, 0 after or the opposite.

You could be-- if you're interested in manufacturing of drugs and pricing of drugs, so one would be before the Generic Substitution Act was passed and maybe the price of medicine is higher. And 0 is after. You could describe attributes of goods. For example, 1, if the house has a deck in the backyard, 0 otherwise.

So without any control variable, if you run this regression, it's very easy to verify, once you write down, you go back to Sarah's early notation when we have the bivariate model, it's very easy to verify that beta hat, when you run this regression, beta hat is numerically, strictly numerically, the difference between the mean of the group A and the mean of group B. If d is a dummy that is 1 when the observation belongs to group A and 0 otherwise.

So in other words, writing down your observation, group A, and then you have the average of y. And then group B, just writing a table here, and this is yB. The difference yA minus yB is beta in this regression. So that means that when we discussed about randomized control trials, we were discussing of taking the means between-- taking the mean of the treatment and the mean of the control, sample average of the treatment and the control group, and taking the difference.

Well, you can do it by hand like that. Or you can run a regression with 1 for the treatment group and you're going to have the same answer literally. And that's what, actually, a lot of people do. Just for reasons that-- I actually don't know exactly why a lot of people do that. But that's why a lot of-- what a lot of-- that's what a lot of people do in general.

One little wrinkle here is the standard error you're going to get for the difference yA minus yB, if you're doing the Neyman regression, do you remember the standard, the estimate, the best estimate of the standard error with the Neyman? Not a regression, but do you remember what it was?

STUDENT: Variance of [INAUDIBLE] divided by the number of control plus the variance of the treatment [INAUDIBLE].

ESTHER DUFLO: Yes, minus 1 in both cases. Exactly. So its estimated variance divided by n of the size of the control group, the same thing with the [INAUDIBLE] plus the same thing minus the size of the-- divided by the size of the treatment group. Now, if you do an OLS regression, what is going to be the estimated variance of beta hat?

STUDENT: Can you repeat the question?

ESTHER DUFLO: If you do an OLS regression, if you run this as an OLS regression, your estimate of beta hat is going to be numerically the difference in the average between the two groups. But what's going to be the standard error?

STUDENT: The error [INAUDIBLE].

ESTHER DUFLO: Yes. And then by what are we going to divide?

STUDENT: By the whole number of treatment plus control.

ESTHER DUFLO: Exactly. So the difference is that the regression, when calculating the standard error, when calculating-- doing any t statistic-- any standard error of the estimate calculation, we'll have not the sum of the standard error for two groups, but the standard error of the difference, which is going to have as adjustment of degrees of freedom the n minus 1, where n is the sum of the two.

So if the samples are very large, it's not going to make any difference. But if the sample are small, and in particular, if the samples are very large or if they have exactly the same size in the treatment and the control group, it's not going to make any difference. Because you get it both sides anyway.

If instead, the samples are different sizes, for example, you have one control group that has five observation and one control group-- and one treatment group that has 100 observations, the OLS estimates of the standard is not going to penalize that-- you for that at all. But the Neyman would, because you would have a very large estimated variance in your control group and a smaller in the treatment group.

So the-- so in small sample, they are going to be a different, the estimate of the standard error. And the Neyman are correct. The OLS would be slightly misleading in this case. But it's a small sample problem. As long as the sample is large enough, and in the special case, which is quite frequent, where treatment and control group are in any case divided equally between the group, it's not going to make any difference.

So that's the relationship between Neyman and OLS. Someone was asking about that. Now, do you see a relationship between Fisher and just OLS? The Fisher exact test and OLS, and what we would do in OLS?

STUDENT: I think Fisher would punish you.

ESTHER DUFLO: Well, it's very different. Those are very different framework. So there really no way that they really look like-- resemble each other. But what you could do is you could run-- you could do a Fisher test by running regressions instead of taking the difference in mean. So you could use-- you could run your regression. You could first run your regression and estimate the true difference between the treatment and control group then draw a different assignment, all the possible assignments if you want, 100, or 1,000, or however many you are patient to do.

And for each assignment, run this regression instead of taking the difference in mean. And you will have a distribution of the betas that you will obtain from this simulation of running every possible assignment that you are looking at. In this case, it will be identical to doing it with the difference in mean because there is-- at no point do you calculate an estimate of the standard error. You are getting it directly from the sample. So in practice, if you wanted to do a Fisher test, you could do a Fisher test with-- by running regressions here.

Now, what happens if you don't have treatment and control group? But for example, you have three-- you have one treatment and two treatment and one control group or you have 51 states. So your variable that you're interested in putting potentially on the right hand side here is not a 1, 0. But it has a value ranging from 1 to 51. Or it could be are you dissatisfied with the product, moderately satisfied, quite satisfied, very satisfied, in which case, you have value ranging from 1 to 4.

So usually, it doesn't make sense in this case to include that variable directly as a regressor in the right hand side. If it's states, it doesn't make any sort of sense. Because there is no way in which the states-- the states are ranked from 1 to 51. They might be by alphabetical order. Or that number, that numerical number doesn't make any sense. It's just a category.

In a randomized trial where you might have several treatments, you might order them. You might decide that if you have one control group, one treatment, which has a low dose, you could call it number one, treatment which has a higher dose, you could call it number two. Now you might be tempted to just put that on the right.

But that would be odd because it's-- the coding between 1 and 2 is a little bit arbitrary. For example, it's not clear that the high dose is even twice the low dose. So the coding is arbitrary. So the data wouldn't make too much sense. So if you have one of these categorical variables, the states, or different one, two, three, four, et cetera. What would you do if you want to deal with these variables? What would make sense to do?

STUDENT: Is this collinearity?

ESTHER DUFLO: So before getting to collinearity, what would you do with your variable that goes-- I've given you a data set that has my outcome for e equal 1 to e equal 100. So my outcome is 12, 12, 15, 9. And then I have my treatment category, which could be 0, 2, 1, or 3. One control, three groups.

What do I do? How do-- what should I do before running a regression? If that's the data set, I've given you, very small data set, it only has two guys. What should I-- yeah?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: I appreciate your politeness. Don't fight. Yes, what would be the columns I would add?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Exactly, t equals 0, t equal 1, t equal 2, and t equal 3. So t equals 0, this is a 1000. t equal 1. this is a 0010. t equal 2, this is a 0100. And t equal 3, this is 0001. OK, so now I've transformed my variable, which is categorical, which is discrete, but where I have different categories into four categorical variables, which are now all dummy variables. And now, I want to introduce this guy in the regression.

But what's going to happen if I introduce all four of them together? Then we have our friend who was ahead of us. Then I introduce my collinearity because they sum up to the constant. So if I introduce the constant and the four variable, that's not going to work. So in practice, what the software would do is to just drop one.

But which one I'm going to drop? They're just going to decide which one to drop. Different software have different convention. But the point is you're not deciding which one to drop. So that's a bit-- that's not what you want. So what you're going to do instead is decide which one you want to drop. For example, you could drop here t1 equals 0 and only include three dummies for t equal 1, t equal 2, t equal 3.

In that case, if I run this regression, what would be the interpretation of beta 1, beta 2, and beta 3, if beta 1-- if I'm running a regression of y I equal alpha plus beta 1 t equal 1 plus beta 2 t equal 2 plus beta 3 t equal 3. What's the interpretation of each of these numbers? Yeah?

STUDENT: Each of them relative to the one that you left out. So rather than [INAUDIBLE].

ESTHER DUFLO: Exactly, so it's a difference between-- it's again numerically, the difference in between the average of the outcome in the group that I have that I'm looking at versus the one that is omitted, which is why you want to omit one yourself and not let your statistical package decide which to omit. Otherwise, you're going to make strange-- you're going to compare to someone you haven't chosen, which might not be the one that is relevant.

So for example, in the case of one control and several treatment, we are typically interested in looking at the treatment relative to the control. Then of course, once we run a regression of this form, nothing prevents us to then compare, for example, t1 versus t2. We can just compare the coefficient. And we can test whether t1 and t2 have the same effect by testing h0 that t1 equals t2 or t1 minus t2 equals 0. Or we can likewise do t1 equal twice t2 or whatever it is that we think is relevant.

Now, if we have other variables in the regression, we could run something like that where we have an outcome. We're regressing on a constant plus beta times a dummy variable plus Xi gamma. So I-- sometimes I write the-- I put the coefficient on the other end. So that the matrix formulation makes sense. So Xi here is a matrix. And gamma is the coefficient.

So in that case, what's the interpretation of beta? In that case, the interpretation of beta is that it shifts. So suppose X had only one member. So this is X is just one variable. Then what beta does is that it shifts the slope between-- it shift the intercept. So this is group A. This is x. And this is y. And this is group B.

Here, beta is actually negative if group A is the dummy. So beta is the difference between-- group B is shifted down. So that's what, if we introduce the dummy variable and we don't interact with anything, it's an intercept shift. So again, going back to RCT, often people want to-- so often, the way you see RCT presented is that the first table of the paper shows you some axes that we don't see a sign assignment. We know it doesn't affect assignment because assignment is randomized.

But might turn out to have been different for a reason or another, the samples are not immense. So sometimes there might be some difference leftover. So you see differences between some variables of interest, which we tend to call covariates. And then if we see that there are some stuff that are different, typically the way you're going to analyze RCT is show one column where you estimate this equation without the X.

So that's the standard difference between treatment and control, which is justified by the completely randomized design. Typically, people cheat on the standard error and report the OLS standard error instead of the Neyman standard error. But that's fine. And then you're going to have a second column, where you have some control variable introduced. And then the interpretation is that the treatment shifts the intercept.

So it increases the-- increases or decreases the value of the outcome in a common way across every value of x. So here, this is estimated under the assumption that the treatment doesn't change the effect of x. The treatment just changes-- puts you one value up. So for example, if the treatment is aspirin and it reduces-- baby aspirin, and baby aspirin might reduce the chance to have a heart attack. One might say, well, men and people above a certain age, et cetera are more likely to have a heart attack. But the effect is going to be lower for everyone. So that would justify this-- that would justify this model.

So that is, again, something that you're going to see very frequently. And of course, it's something that it doesn't-- non-RCTs as well you're going to often see, for example, you're interested in whether men or women have different wages, conditioning on doing exactly the same on-- controlling for the fact that women have different levels of education maybe.

Women may join different fields maybe. Women have different potential experience maybe because some women may have taken some time out of the labor force to take care of kids. Conditioning on all of these things, is it the case that women tend to have lower wages than men?

You would run exactly this regression, which is-- does the fact that being a woman gives me a lower intercept to start with, and then I have all these other variables that are going to have their effect that might also turn out to be different between men and women? So that's one way to handle other variables. It's just not you-- just add them separately. In which case, the dummy variable just shifts. Yep?

STUDENT: This is like a practical question. When you have a variable like age in practice, is it more common [INAUDIBLE] pertaining to a certain age brackets and variables, [INAUDIBLE] and ages?

ESTHER DUFLO: So it's a great question. I'll go back to that because there-- it's a more general question of functional form, where in the case of age, it's not a discrete variable, except if you are in a very small-- if you're looking at a very small age range, like 20 to 25, it's too large to be discretized such that 19, 20, 21, 22 are separate.

But you might or might not think that a linear relationship between y and age is appropriate. So let me go back to that. That's kind of other function form issues. And I'm definitely going to talk about that in a minute. That's a very important practical questions. And really in this class, I'm really trying to go there. There's no really-- nothing really theoretical.

We are all going, kind of trying to understand how we use these tools in practice and how researchers. So just interrupt me and ask me any odd questions. And then I'll distribute them during the class or answer them right away. So just don't be don't be shy, please.

Now, another way that you might think of introducing dummy variable is not just as an intercept shifter, but interacted-- what we call interacted with the x's. So a simple way to look at it to start is assume you have-- both variables are dummy variables. You have, for example, a model where you have a treatment that might be randomly assigned.

And then you have a group, for example, you have male versus female. And you might be interested in knowing whether the treatment affects men different than women, differently than women. In which case, what you're looking at is not just an intercept shifter, but to know whether the effect, the treatment has a different effect in one group than another. In which case, you're going to run a model with what we call interaction, which is you're going to multiply 1 variable by the other.

So here, we're running, for example, the outcome. Let's say it's our D is a training program for people who are unemployed. And Y is probability to be-- is whether or not the person is re-employed after six months. And M is a dummy for whether the person is a male. So I'm running Y, the person is re-employed after six months on a constant plus beta, they were treated, plus gamma, they're male, plus delta, they are treated.

It's a one-- this one is a product between the two. It really literally is the row product. So it's going to be a 1 if the person is a male and if they are treated. And a 0 otherwise. So that's what we call an interaction. And here, it's like by categorical variable. So how do we interpret the coefficient?

What is alpha, beta, gamma, and delta in this model? What are they numerically equal to in term of differences between groups if there are no other variables in the regression, and these are categorical variables.

STUDENT: Is the difference in sample means again. treatment and control.

ESTHER DUFLO: Between treatment and control and--

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Not quite. So it's the difference between treatment and control among female. Gamma is what?

STUDENT: Differences in treatment for male [INAUDIBLE].

ESTHER DUFLO: It's the difference between male and female in the control group when treatment is equal to zero. What is alpha? What is going to-- what's the-- sorry?

STUDENT: Difference between male and female for the treatment?

ESTHER DUFLO: No, not quite. So alpha is the constant here in the regression.

STUDENT: For the entire treatment control group.

ESTHER DUFLO: It's the average for the control group for female. And what is delta?

STUDENT: Averages of the male treatment group?

ESTHER DUFLO: No.

STUDENT: Difference between that and [INAUDIBLE].

ESTHER DUFLO: Yes, exactly. So let me put it in-- let me put it in a 2 by 2 box. So this is a very popular design called difference in difference or an interaction variable. So you're going to see it again, and again, and again. So it's good to remember the-- you clan prove what I'm saying. But it will also, once you look at it, it'll look reasonably intuitive to you.

So suppose this is a treatment. And this is 0. And this is 1. And this is a male, female, and this is male. And I'm going to put the treatment as the first in this and the gender as the second in this. So this is Y00, the average for treatment group, for a control group among female. This is Y01. This is Y10. And this is Y11.

In this regression, this is going to be-- this is equal to alpha. This is going to be the difference between male and female. So the difference between the two between male and female is going to be-- so this I'm putting the difference, which is this one minus this one, so one is 0 minus Y00.

That guy is going to be-- what did I write in the regression? It's going to be gamma, the difference between male and female. The difference between treatment and control in the omitted group is Y01 minus Y00 is beta. And then the difference in difference is gamma, which is Y1-- let me put it Y11 minus Y01. So gamma is equal to Y10 minus Y00 Y11 minus Y01 minus Y10 minus Y00. That's a difference in difference.

So the reason why this works is that you can calculate it, but the reason why this works is that imagine that we are putting both what we call the main effects, which are the dummy for each of the groups and the interaction. So what this tells you is that the beta is the-- if we didn't have anything else and we run it only in the female group, in the female group, we would get beta.

To get the effect for male, how would we-- how do we get the effect for male? So beta gives us the effect of the treatment for female. How do I get the effect of treatment for male? Yeah?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: So how do I get-- from this one, how do I immediately get the effect for male? How do I compute it? You can also see it in this table. The effect for male is Y10 minus Y1-- that's the effect for male. So you can see it from here, this is gamma. And this is beta. So if I want the effect for male, I just have to add up exactly gamma plus beta.

Gamma plus beta is the effect for male. So what this delta is-- so beta is the effect for the omitted group. M is the difference between male and female for the control group. And delta is the additional effect for male compared to female. So that the total effect for male is simply beta plus delta.

If it's not, it could be-- I'll give you another example where it's not a treatment and control for any situation where you have two groups that you, for example, you could have a law, state that pass a law, eventually before and after they pass the law, and a state that never pass the law, before and after they pass the law.

So then you could say that D is going to be dummy for pre versus post. M is a state that do it versus don't do it. Beta is the difference between pre and post for the states that don't-- never pass the law. M is the difference between states that pass the law and states that don't before the law is being passed. And delta is the extra difference after the law is being passed, which is going to be, in many cases, a reasonably good estimate of the effect of the law. Yep?

STUDENT: You've got two gammas on the board. One [INAUDIBLE] has a delta, right?

ESTHER DUFLO: Yeah, one should be a delta. This one should be a delta.

STUDENT: So the gamma is the difference?

ESTHER DUFLO: Yeah, the gamma is the difference. No, opposite, in my-- it's my-- my notation is not-- the delta is the difference. The coefficient of the interaction is a difference. I'll give you an example soon.

So we went over all of this. And we discussed the-- so yeah, we discussed how to get the treatment effect for male, which is beta plus delta. If I wanted to know the average for male from this coefficient, how would I get it? If I wanted an estimate, just from using this coefficient, I wanted the average of the male.

STUDENT: The average what?

ESTHER DUFLO: The average of the outcome for the males? I could just calculate it from the sample. But there is also-- it can also be found from this regression.

STUDENT: Delta plus beta plus alpha.

ESTHER DUFLO: Correct. No, not delta. The average for the male, so the average for the male in the treatment group is going to be alpha plus beta plus gamma plus delta. The average for the male in the control group is going to be alpha plus gamma-- and then the forgetting the actual average for the male, I would have two, take the average between the two. Yeah?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: So suppose that I'm interested in the effect of the male, I need to put a 1 for the relevant places. So if I'm interested, for example, in the average effect for the male in the control group, I'm setting D equal to zero. Alpha is the constant. So it applies to the male. And then gamma is 1. So the average effect for the male in the control group is going to be alpha plus gamma.

Now, if I want to know the average effect for the male in the treatment group, they have a 1 for being a male. And they are 1 for being treated. So the average effect for the male in the treatment group is going to be alpha plus beta plus gamma plus delta. They have 1's everywhere. Now alpha is the average effect for women in the control group. If I want to know the effect for women in the treatment group, what is it going to be?

STUDENT: Beta plus alpha.

ESTHER DUFLO: Beta plus alpha, exactly. So I think you're getting the hang of it. So that's the basic difference in difference model, which is very popular in applied work. Because often, you have a law that is passed, for example, in some of the states. You're not willing to assume that the state that passed the law are different.

Take the current controversy about transgender bathrooms, for example. And you want to know the effect on the number of gay men and women who decided to take a vacation in those states. So you don't-- you're not willing to assume that in the absence of the law, the same number of gay men and women would visit the states because maybe it's not such a nice place to visit anyway because there is a reason why those laws got passed. You might not willing to assume that over time, before and after the passage of the law, the trend in visiting to this would have been the same, say because gay men and women are more willing to travel in general because life has become easier in general.

But you might be willing to make the assumption that in the absence of the law, the state might have been following the same trend. So in the absence of the law, the difference that we observe in-- which passed in, North Carolina? The difference in a state that passes the law and a state that doesn't pass the law in term of the number of gay men and women who visit for vacation would have been similar.

In which case, you run a regression that is like a difference in difference. But D is post. M is whether you pass or not the law. And M times D is post times passed the law. And in that case, delta is going to be interpreted immediately as the effect of the law under the assumption that in the absence of the law, the difference would have remained the same. Because the only reason why the difference has changed is because the law was passed.

So I want to give you one example of using that method from some of my work. And that's going to, again, give us some practice in difference in difference. So I looked at-- Yes?

STUDENT: I was going to ask you a quick question, just before we move on. Couldn't you-- so in this case, you have two groups. Or you have treatment and no treatment. And you have male and female. But could also do diff in diff for just a regular-- just any old regular regression where you have two groups and a treatment or--

ESTHER DUFLO: Yes, you could do anything.

STUDENT: And everything-- we only cover diff in diff today. But we've looked at lots of regression before today. Would a diff in diff be more convincing in an article or a journal article, just looking at the means of treatment or the effect?

ESTHER DUFLO: Diff in diff is just another regression where the regressors are the two main effects for the groups and the interaction between the two. So it is justified, in some cases, for example, it is justified in all sorts of cases. It is justified in some cases where you're interested in the differential effect. It doesn't even have to be a treatment. It could be that you're looking at, for example, having gone to college or not in male versus female.

None of that is randomly assigned. You can still look at whether the wages of someone, as a function of whether or not they went to college or not, they are male versus female, and the interaction of the two. In that case, the interaction is going to tell you whether men have a bigger premium when they've gone to college than women, or maybe the opposite is true.

So just the college premium is an interesting question. Potential difference between men and women wages is interesting. And then the interaction of the two is also interesting. These are different questions that you might ask yourself. So that would be one setting where you would look at interactions, where you're just looking at differential effects.

Another setting where you set up differences is the one I just described, where you set it up because you're interested in specifically-- you're willing to make the assumption that in the absence of your reform, you have two groups, one of them that eventually passes the reform and one that doesn't, and you're willing to make the assumption that in the absence of the reform, they would have been the same.

In that case, the difference in difference gets very convincing. An example that you and I discussed in detail, and that's sort of the famous examples of using difference in difference, I'll go through them in words now, are the Mariel Boatlift experiment and the New Jersey-Pennsylvania minimum wage experiment.

So what is the Mariel Boatlift experiment? At some point, in a very short period of time during very few days, Fidel Castro authorized whoever wanted to leave Cuba from the Port of Mariel to just go. So hundreds of thousands of Cubans suddenly reached Miami in very few days. And they settled, for the most part, in Miami because that's where they had landed.

So suddenly, you have many more people who have reached Miami. And the question that card is asking is whether that affects the labor market outcomes of people who live in Miami anyway. So what it does is that it takes a bunch of cities that are reasonably similar to Miami's in terms of their size, their education makeup, their racial makeup, et cetera. And you run the difference in difference.

I wanted to give you a table, but it didn't actually run it this way. So I thought it would introduce extra complications. But it runs a difference-- it runs a difference in difference specific of the type. The wages of someone who is a native from Miami or one of the other cities that he uses as a control. On a dummy for whether the city is Miami or not, a dummy for whether it's pre Mariel Boatlift or post Mariel Boatlift, and then the interaction of the two.

So if you're willing to assume that in the absence-- of course, there is business cycles or things change over time. But if you're willing to assume that in the absence of the Mariel Boatlift, Miami and the other comparison city would have moved roughly in sync, then the difference in difference coefficient delta here tells you what is the extra wages that people in Miami get post Mariel Boatlift.

Or the hypothesis, I suppose, is that delta is 0, h0 is the delta h0 against the alternative that delta is actually negative because all these Cubans are fighting for jobs with the native. So the interesting result in this paper is that delta is zero. So that doesn't seem to be-- although lots and lots and lots of new people arrive potentially competing for wages, there is no effect on the wages of native Miami people.

So this has been enormously influential study because it was one of the first use of this difference in difference idea in a very convincing-- potentially very convincing set up. As a result, it has also led to considerable debate on whether it was correct or not. And this debate is interesting because you set up-- you see all of the issues with this type of technique. For example, well, is it really true that these other cities would have been on the same trend? Maybe they would have been on another trend? We have no way to know. So we have to-- this is an assumption.

Another very famous experiment of difference in difference is also by-- involved Card, David Card, who is a professor of labor economics at Berkeley and Alan Krueger. And they look at minimum wage. So they compared fast food restaurant workers in New Jersey and Pennsylvania. And they are looking at employment after an increase in the minimum wage in one of the states. So they are, again, comparing New Jersey versus Pennsylvania before and after the minimum wage was passed and see whether people lose their jobs post the minimum wage.

STUDENT: I was wondering, would it be-- if you're not sure whether two cities would be on the same track, would be more convincing to choose one city and then compare that to an average of the two cities?

ESTHER DUFLO: That's an excellent question. And the answer is yes. And in fact, you can do this. There is subsequent work on difference in difference that does this very systematically to find an average of-- the idea would be to find an average of other cities, which is in fact, a weighted average. You just-- you don't take just a simple average. You take a weighted average, such that your weighted average looks very, very, very much like Miami.

So then when you run your difference in difference, you will get almost-- you get the beta is 0 because you've constructed a synthetic group of cities, which is quite similar to the original city. And and then there is an effect of post. And then there is the interaction. So this is some work that was done later by Alberto Abadie, who is actually coming here to teach here next year, to join the faculty. It's called synthetic-- I'll write his name down. And I'm sure it's going to teach an excellent applied econometrics class. So you can-- if you want more of his work, Alberto Abadie. And it's called synthetic control-- synthetic control group method.

Where synthetic comes from the idea that we create a control group out of dispersed. Another thing you can do, and I'm going to show that to you in your example, is to-- a testable implication that the cities are on the same trend is that they were on the same trend before. So you can at least look at what happened before.

So here is an example of that. So this is an example of the use of difference in difference where I look at school construction in Indonesia. In the late '70s, there was an oil shock, the oil shock turns out to be a boon for Indonesia because Indonesia is an oil producing country. So they had suddenly lots of money.

And they decided to use it, in part, to construct schools. So in five years, they constructed 62,000 schools all over the country. And before that, they had no money. So they were not building any schools. So that's one-- we have a pre, post, which is pre '73, post '73. And then the way they allocated the school is they try to allocate the school in places where education levels were low to start with. So that creates a difference between regions that got less schools, fewer schools, and regions that got more schools.

I looked at-- I used data that was collected in much later. So we see people who have completed their education. And what we are going to be interested in today is the effect of education. So there are two factors that affect the intensity of the program in this case. The year of birth because if you were seven or younger when the schools were built, you get to go to them.

Otherwise, you already-- so if you are 12 or younger when they were built, you get to go to them. Otherwise, you have finished school already. And then your region of birth, which is you get more-- some places get more schools than others. So here is the equivalent of this where-- so you can focus for now on the top of this table.

And what you can see on the top of this table is the regions divided by high program regions and low program regions. And then we have kids who were young in 1974. So they were young enough to go to the school. And then a group of kids aged 12 to 17 in '74, so they were too old to go to the schools.

So first thing we can do is compute the average years of education that people get in each of these groups. And then we can compute the difference. And then we can compute the difference in difference. And here the standard error that is below the coefficients are from a regression from running the regressions.

So what you can see is that the places-- people who were born in programs that got a lot of schools are generally less educated than people who are born in regions that got few schools. Does that makes sense? So yes, it does make sense because they were building the schools in places where education levels were low.

And then we also see that over time the younger people are generally more educated than the older people. That also makes sense because the country was growing. But what we also see is that this growth is faster in places that got more schools. So the difference in education level is 0.47 for the younger kids versus the older kids in high program region and 0.36 for younger kids versus older kids in low program regions. And the difference in difference is 0.12.

So we could say, well, if you assume my hypothesis, my assumption, I'm going to maintain is that in the absence of the program, the difference would have been the same over time. Or another way to put it is the difference between the regions would have remained constant over time.

Then if we assume-- willing to assume that, then this 0.12 is the effect of having more schools rather than fewer schools. One way to test that is to say, well, let's look at what was happening before. We can compare the-- we can do a sort of placebo experiment by comparing the older kids to kids who are even older.

None of them benefit from the school. So hopefully there is no difference in difference in that group. And in fact, that's what we're finding. We're finding now a much, much smaller difference in difference, 0.13. So this is reasonably-- this helps in-- it's not really a test of the assumption because things could have changed later. But at least it is a little bit reassuring.

So that's one use of difference in difference. More generally, if our variables x are not constant, but are things that are discrete, are not dummy variables, but things that are discrete, the interaction between a dummy variable and some variable X tells us the extent to which the dummy variable shift the regression function of that regressor. So what I wrote on the board doesn't make all that much sense. I did it for you.

Suppose that I have one regressor, one continuous regressor X and my dummy variable. If I run a regression Yi equal alpha plus beta Di plus gamma Xi plus gamma times Di times Xi, then what I'm getting-- what this gamma tells us here is by how much the slope of the function is changed by the dummy. So we have a-- should be delta, yes. Something that's not gamma.

So we have group-- this is, say, group B, which is the D equals 0. And then maybe this is the regression function. So this is the relationship between Yi and Xi. This is group A. So the intercept, the shift difference is given to us by D. And the intercept-- the shift-- the tilting of the slope is given to us by delta.

So my intercept is-- let me go to the intercept. This is D. So this is really the intercept. So I have to evaluate that at zero. This is the change in the intercept. And then the difference in the slope is the difference is this coefficient delta.

So here is one example, just continuing the example we had before where we run that, we can run the education of an individual on a bunch of-- we could just-- we could do it linearly. So we could do-- because before that, I had separated the regions somewhat arbitrarily between high program region and low program region.

But in fact, for every region, there is about 300 regions in Indonesia, and for every region, I know how many schools they got. So a thing that-- first thing that I can do is to run exactly this regression by saying I'm going to put a dummy for being young. And then instead of a dummy for high program region, I'm going to put the number of regions that-- the number of schools that were built, which I'm going to call-- I call here P .

I'm going to call that number-- I'm going to put the number of schools that were built in my region of birth and then the interaction between the two, young times P. And then this delta is going to tell me whether the extra-- the difference in the slope between the effect of the school before the program and after.

Of course, before the program is just telling me that there is something different about these regions. But after, the difference between before and after is really the effect of the schools. So that's what we could do. The way-- what I've written here on the slide is almost that. Except I've replaced the young dummy by a bunch of year of birth dummies.

Of course, the young is just a linear combination of the year of birth dummy. So if I put each year of birth dummy, it's going to be a little bit more precise. As per the discussion we had before, I'm going to omit one. And then I'm going-- instead of putting the number of-- instead of controlling linearly for the number of schools, I'm going to put a bunch of region of birth dummies, again, omitting one.

So this is what this alpha j and beta 1k are. They are a set of year of birth dummy, a set of region of birth dummy, and then the interaction between the number of schools built in my region of birth and whether I'm young enough to go to these schools. So this is going to give me something like that. This is simply a graph and then a linear regression between-- for each region, I computed the education of the high-- of the young guys minus the education of the old guys. And I'm regressing on the-- I'm regressing on the number of schools built in the region of birth. So that coefficient is going to be exactly that. Yeah?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Anyone, can omit anyone. Because in this case, you're not very interested in the first dummy. So you're not going to try to interpret them. Traditionally, you omit the young-- the oldest guys, so everybody is compared to the oldest guys. And for the region's dummy, really anyone, doesn't really matter at all. Yeah?

STUDENT: Do you get the a number of interaction terms equal to the number of different buckets that you have?

ESTHER DUFLO: So in this case, I have lots of potential buckets because I have 300 regions and then 25 regions of birth. So if I did all the interaction, that would be a lot of interactions. But I'm only interested in one, really. So what I've put here on the right, so it's not really a difference in difference anymore. It's an interaction effect.

What I've put it on the right is what I'm interested in, which is the region the number of schools built in the region of birth interacted with whether you're young or old. So although I have lots of region of birth dummy and lots of year of birth dummy, I only have one interaction coefficient, which is the one that I happen to be interested in.

STUDENT: OK, so is it fair to say that when you run a regression with one interaction term between two variables that have two categories each, that is a difference in difference.

ESTHER DUFLO: Exactly. That is, strictly speaking, a difference in difference. Where is it? That is strictly speaking, a difference in difference. And that is strictly speaking, a fixed effect regression. People call it a fixed effect regression because these guys, a bunch of dummies for the region of birth, they are called region of birth fixed effect.

It is systematic differences that don't vary over time between regions. And then the beta 1s, which are a bunch of year of birth dummies, are a bunch of fixed effect-- fixed effect for year of birth. So we tend to call this one fixed effect regression.

STUDENT: All ones or zeros?

ESTHER DUFLO: All ones or zeros, exactly. I was a bit lazy to write down some sign with all of the dummies. But this is-- so this is kind of a collapsed way to write fixed effects. I wrote it in the text. So it's a bunch-- it's a dummy for each year of birth minus 1. And so for observation i, that dummy, all of the dummies are 0, except for one, which is for the year in which that person was born, where it's 1.

So this is how the table could look like. So there are several regressions here. But you can focus on the top here. This coefficient, 0.12-- 0.124 with a standard error of 0.0250 is the interaction between the number of school built and young versus old. So this is, in English, this tells us that for every school built per 1,000 kids, because I normalized per 1,000 kids, for every school built per 1,000 kids, the year of education of the young cohort was 0.12 years-- the young cohort got 0.12 more education than the old cohort for every year. And that's my estimate of the effect of one school.

All right, now if we want, you could do one more step and say, well, I'm now going to also separate the cohort, not just between young and old, but between all the cohorts, the people who are 23, the people who are 22, the people who are 21, et cetera. And I'm going to run this regression always comparing cohort by cohort.

In that case, I should see no effect until people are young enough. So I'm going to get a bunch of zeros for the older cohort and then start positive numbers as the people are young enough to go into the schools. So this would be a more sophisticated version of the test we did before that there is no-- that the trend is different only thanks to the school. That make sense?

STUDENT: So where is that in the table?

ESTHER DUFLO: It's not there. This is just-- I did it-- I thought you, by that time, you would be fed up of seeing regression. So this is just-- this is I just did young versus old. But in the same way that you did young versus old, you could do 24 versus 23, 24 versus-- in 1974, 24 minus 22, 24 minus 21, et cetera, and estimate that. You can even estimate that as one big regression.

STUDENT: Did you do that in the paper?

ESTHER DUFLO: Of course.

STUDENT: Do you see zeros--

ESTHER DUFLO: Yes, it's very nice, like that. So that's kind of pretty reassuring. Yeah?

STUDENT: [INAUDIBLE] regressions.

ESTHER DUFLO: Yes, so this is extra control variable introduced. So in the first, one might be worried that there is something different that-- so it's another way to deal with the trend with potential trend differences. So going back on Liz's ideas, maybe we should take an average, a way that you could try and get closer to an average is try to control for things that were different to start with. And that might-- and then interact that with the year of birth.

So in that case, you are allowing the trend to be different for the regions that get treated. You know that they had lower enrollment rate in 1971. So you can account for, you adjust for the fact that they had lower enrollment in 1971 by controlling for year of birth time enrollment rate. And then this one is because they had another program that they started roughly at the same time, and that I was worried would also affect education potentially. And that also has it could have been in the same region. So this one controls for the water and sanitation program. So as it turns out, the coefficient becomes a little bit higher for each control you add. So if anything, without introducing this control, you slightly underestimate the effect of the reform.

STUDENT: Another quick question on that regression you show here. What are the p-values associated with-- [INAUDIBLE] could you plot them together, is that [INAUDIBLE]?

ESTHER DUFLO: Meaning? For each coefficient--

STUDENT: [INAUDIBLE] you're comparing.

ESTHER DUFLO: Yes, you can plot them. So each of them, there is a coefficient for each year of birth. So the younger people are here. And the older people are here. And then for each coefficient, there are associated standard error. So you can plot the associated standard error. So you plot something that-- you get something that's like that. It's not as nice. And you can do it separately. Or you can do it in one massive thing where you introduce all, as you were proposing, introducing all the interaction one by one.

So as other functional form issues, quickly. Sometimes, your economic theory tells you that the model is not at all linear. We started by saying the linear model, the OLS model, incredibly powerful. You can have lots of stuff you can do with it. It's blue. It's like-- we love it. But sometimes, the world is not linear. So what do we do?

So sometimes, there are cases where it's not linear, but by transforming the dependent variable, it could become linear. So for example, suppose you have a traditional what we call Cobb-Douglas function, which is, say, production function, where to produce widgets, you need two inputs. And they enter with these power functions.

So that's not nice and linear. But if you take the log of that, bing, it becomes linear again. So the log of Y in that case is linear with the coefficient between-- of each of the log X being interpretable as elasticities, which is-- when you vary X1 by X percent, Y change by beta 1%. X1 by 1%, sorry, Y change by beta 1%, so it would be a 1 here.

So that's one thing could do. So how do you know that it's like that? That's where you usually-- you're starting from a model. That tells you that that's a reasonable way that your outcomes are related. Sometimes you have-- so the log log is very popular because of that elasticity interpretation. It's unit free. So it's very convenient. And it tells us that if I vary my variable by 1%, the outcome will vary by beta percent.

Another formulation that is popular is this log linear relationship, where the log of the variable is regressed on the variable-- on a linear version of dependent-- of a regressor. This is very popular, for example, to estimate the impact of education simply because it happens to be quite-- to represent things very well.

So this is telling you that for each year of education, wages increase-- for each extra year of education, wages increase by about beta percent. So if you go from 0 to 1 year of education, your wages increase by 7%. If you go from 1 to 2 year, wage increases by a further 7%. And any one-to-one increase, one year increase of education is associated with beta, beta 1 increase in education. You might think it's crazy. But it happens to be reasonably close to being true.

So here are other example that we don't need to go in detail, but you will encounter if you continue to do economics. One is the Box Cox transformation. I'm going to skip all together. And one is the discrete choice model. So the discrete choice model is used when you have data that is the percentage of-- you have people can choose between Honda Civic, and Toyota, and whatever other small car that corresponds to a Honda Civic. And you are wondering what they are going to choose from.

And your data comes in the form of the fraction of customers that chose this particular model. Now, it turns out that a very convenient way to represent that is with this formulation, which is called extreme value formulation, where the choice depends on the attributes in this form. And then if it turns out that this is the form that it takes, then when you take the log of Py over 1 minus Py, then you get a linear form that you can regress.

So if you're given a data of the form x percent of people choose this type of car and 2% people choose this car, this type of car, and 5% choose this type of car, and 7% choose this type of car, you can then regress your-- transform the formulation, and think if it's 5%, it's 5 over 95%. And then this is actually a linear model. You might think, why would this be true? Well, it happens to be true [INAUDIBLE] quite naturally from choice models. And it also again happens to be-- often seem to correspond to reality quite nicely. So this is a way in which you start-- another example where you start from something very non-linear and end up to something quite linear.

Now the next question you can ask is-- actually, why don't I skip that. And we'll start from here because that's kind of a new topic. We'll start from here and do regression discontinuity design on Monday. And then we'll move on straight to omitted variable bias and IV if we have time. Or we'll skip IV for after the machine learning.