[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: So continuing on issues of functional forms, we briefly discussed what happens transforming the dependent variable to turn a nonlinear model into a linear model that we can estimate with OLS. You can also sometimes transform the-- leave the dependent variable what it was and change the independent variable. So when might you want to do that?

Well, suppose that you start with your data set and then you have some variables, say, demand for gasoline. How many people-- how many gasoline people are purchased and the price of gasoline, and you notice that you-- as good trained data scientists, you start by looking at some pattern in the data. And the first thing you decide to do is to run a kernel regression of the demand for gasoline, on the price of gasoline. And then you realize that it's really not-- doesn't look linear at all, that people have-- that in some places, very steep. In some places, much flatter. So it doesn't look that it's going to be characterized nicely by a linear model.

On the other hand, you wouldn't want to continue working necessarily with a kernel regression because it's a bit heavy going. It's harder to control for other things. And we know how we-- we know how to nicely represent kernel regression as bivariate things.

But then when it becomes multivariate, it's kind of harder to look at. So you would really move to a linear model. And the question is, can you do that? Is it the case that if the relationship between y and x does not appear to be linear, can you not do OLS?

And then in fact, you can do OLS. What you can do is to try to transform the variable x from just its basic incarnation as x to something a little bit more richer in order to be able to represent the nonlinearities and yet estimate a linear model. So this is why the linear model is beautiful, because it actually doesn't restrict us to estimate linear things. We can just modify X to go back to a linear model.

So there are only two things you could do to do that, two ways to go, and I'm going to briefly talk about both ways to go. The first thing you could do is to say, as you know, any functions can be approximated by polynomials. So the first thing you could do is to regress y, not on the x only, but on polynomials of x, or in fact polynomials of transformation of x's.

And once you do that, you're back to an OLS. So the simplest way you could do that is to say-- suppose I'm back to my gasoline example, which is actually a famous example, because it's when where Jerry Hausman and Whitney Newey who are both econometricians in the department, developed a lot of the theory underlying not what I'm going to say in a few slides, which is series expansion.

Take your nonlinear regression, it looks kind of funny shape, not at all linear. So then you decide to run a y on polynomials of x. Here, I'm assuming that x is just a vector, not a matrix. So it's, for example, a purchase of gasoline on the price of gasoline.

So we have the price of gas-- a constant, the price of gasoline, the price of gasoline squared. Cubic, and any number you want to put in. So you could put straight polynomials in there, or you could do orthogonal polynomials, or whatever transformation of x that you want.

If you assume that the model is known, for example, you assume that it's really a square or it's a cubic, then you're back to straight OLS, nothing special. You just have OLS, but your y variable is regressed on three regressors, which happens to be x, x squared, and x cubed. So all the theory that we have gone through goes through without any change. It's just that you assume that y and x are related by this particular shape.

So you're good. Once you've done it, once you've estimated your polynomials, you can plot it by taking for a number of ranges, say, of the price of gasoline, the predicted value for demand based on this coefficient. You have estimated beta 0 hat, beta 1 hat, beta 2 hat, and beta 3 hat, let's say, if you have done a cubic.

So you can plot it. You can take the derivative because we know how to take the derivative of this thing. So it's very nice.

So that's easy to assume. If you assume that the model is not known, in particular, you don't know how many-- what is the real functional form? You just don't know.

You want to approximate it with these polynomials, then we-- then you can still do a linear model. You don't have to do a kernel regression. You can do a linear model and just add powers x2, x3, x4, x5 et cetera till you have the shape you like.

And then the question you might want to ask is, well, but how many should I add if I don't know the-- do you have-- does anybody have an idea of how many you should add and how we should think-- we should even think about it?

When do I stop? So if I know the shape, I know it's a cubic. I just put a cubic. If I don't know the shape, then when do I stop adding polynomials? Yeah, go ahead.

AUDIENCE: We want to minimize the variance. So if you are adding a lot of polynomials, the variance--

ESTHER DUFLO: Almost. Go ahead.

AUDIENCE: My thought was this the size of the coefficient. If it's larger, then it's probably relevant. If it's not--

ESTHER DUFLO: You want to look at the coefficients, I think one might do that. He was a little bit closer to the way that is conventionally done. Yeah.

AUDIENCE: Wouldn't going beyond the cubed just basically mean more y-stretched versions of cubes and squares?

ESTHER DUFLO: No. Yeah?

AUDIENCE: Maybe you have the possibility [INAUDIBLE].

ESTHER DUFLO: Possibly. Lisa?

AUDIENCE: What machine learning people do is optimally split the data set between a test set and training set.

ESTHER DUFLO: Yes. So we could possibly go a machine learning route, just train the data set. We're not going to do that for this because we actually have a theory, which helps us think about that directly without splitting the data set. What we can do is exactly what we did for kernel. It's exactly the same problem.

He had about half of the answer. The other half is that, well, we have the variance. The more polynomial they add, the squigglier our function is going to be.

On the other hand, we also have bias because we don't know the actual shape of the function. So if we don't do enough polynomial, we have bias. So we need to minimize some combination of bias and variance. What is our standard way to do that is mean squared error.

So we're going to try and minimize mean squared error. How do we minimize mean squared error? Exactly like we did with kernel regression, which is we're looking at a cross-validation criteria. We calculate the sum of the estimated mean squared, and at some point we stopped when we have enough polynomials. So there are-- so that's exactly how we're going to pick the polynomials.

Other than that, that we need to pick the polynomials, the behavior of this linear expansion is very, very, very similar to OLS. So the theory is a little more involved because we have bias that remains, but the bias will go away as you add more observations and it goes away. And the variance, of course, will also go down as you add more observations.

So basically, it's the same idea as kernel. With the kernel regression, what you promise to do is to reduce the size of the kernel-- of the bandwidth as the number of observations increases. With polynomial regression, with series expansion, what you promised to do is to add more polynomial as your number of observations increases. So it's exactly the same principle.

And the nice thing of that is that's very easy to manipulate in theory and in practice, because it's just OLS. So it's even simpler than doing a kernel regression. Although, at the end of the day, it's Al who does it. So Al does a kernel or a series.

But in term of manipulating, it's easier to think about it. So it's very easy to take the derivative of this. Therefore, you can-- if you're interested, once you've estimated, you say, oh, suppose I'm interested in the elasticity of demand with respect to the price. You can get it straight from the function by taking the derivative.

So what else could you do? You could take the log of x. You could interact the x's in the same way that we see that we could interact dummies. You can interact dummies with a linear and a dummy.

We can interact a dummy with an x, any non-- any continuous x, and that's going to shift the slope. For when the group turns to one, it's going to shift the slope. In the same way, and you could interact two continuous regressors such that the slope of one depends on the level of another.

So if you have several x's and you want to-- and you have-- and it could be interacted in many ways, and they could be flexible in many ways, then you start to have, potentially, a lot of choices. As we know from what Sarah talked to us about, as you add more stuff in the regression, the r squared will mechanically go up, but the variance of all the coefficients will go down because if we add irrelevant things, it's going-- it's eating up degrees of freedom.

So we don't want to do an infinite number of those stuff. The issue has become how to choose, and that's where Lisa's point about machine learning was just prescient. That's when machine learning tools, which we are going to learn more about on Wednesday, can become useful, can become handy. So there are various tools. Some of them require training on the part of the data set and then using the other one to estimate, some of them actually doesn't even need-- doesn't even require that.

So that's one way you could go, is first take a series expansion if you have a single x, then start interacting them, interacting series expansion if you want to, et cetera. And all of that gets estimated by a linear model and therefore are easy to work with. The second thing you could do is using dummies for approximation.

So for example, suppose that you have some function like this and you want to work. For whatever reason, you don't work with-- you don't want to work with a polynomial, but you could say, well, let me estimate-- let me approximate-- let me just partition the range of x's into a bunch of intervals.

So this is what I have done here. I've partitioned the x's in a bunch of intervals. So there is a dummy equal to 1 if x belongs-- is between x0 and x1.

That's a dummy equal to 1 if x is between x0 and x1, and 0 otherwise. Then I can have a second dummy, which is equal to 1 if x is between x1 and x2, and 0 otherwise. And then a third dummy that will be 1 if x is between x2 and x3 and 0 otherwise, et cetera.

So first thing we could do is to just regress our y on a bunch of dummies-- on a bunch of those dummies. And here, I have not put the intercept. So why didn't I put the intercept in this case? Why did I knowingly omit the intercept? Yes.

AUDIENCE: Because it's really ambiguous between each side, which one [INAUDIBLE].

ESTHER DUFLO: Say it again.

AUDIENCE: It may be a little ambiguous. You introduced a lot of boundary conditions.

ESTHER DUFLO: Go ahead.

AUDIENCE: [INAUDIBLE] context, right, except for the first one?

ESTHER DUFLO: Yeah. And practically, if I put the intercept, what it's going to do?

AUDIENCE: It's going to be 0.

ESTHER DUFLO: Even more than that, if I put the intercept on all the dummies.

AUDIENCE: Oh. It's going to be multicollinear.

ESTHER DUFLO: It's not going to run. yes, it's going to be multicollinear. So what I could do is to put the intercept and omit one of the dummies. Then everything will be relative to that.

For example, if I omit the intercept-- if I put the intercept and omit the first dummy, that's going to be exactly the same as running exactly that. So when I run this regression, I'm basically estimating points at various-- at each level. So it's going to be like this, this, or it's like this.

Something like that. So that's one thing we could do. Another thing we could do-- so that's not-- that's not bad, but it's not perfect, because you can see that here, there is a step. The thing is a step function, and some of the steps are not there in the real data. They're not there in the true data. I just made them up because I used the dummies.

So instead, what could I do instead of doing this just estimating with the dummies? With just the dummies? To espouse the shape of the form a little bit better, not have these artificial steps. Yep.

AUDIENCE: Parametrically fit within each interval.

ESTHER DUFLO: Yeah, exactly. So what would be the simplest parametric fit? I could try to estimate the line in between each thing.

So instead of estimating this, I could say I'm going to define a piecewise linear variable as a dummy interacted with x This guy should be a subscript. A dummy between those two. So then when I do that, what I estimate is instead of estimating these steps, I estimate line, something like that.

Well, it is not very good looking. Something like that. So here, we need to center the variables. If we center the variables as I did here, which is we remove the threshold every time, this is going to ensure that the two fit.

Otherwise, every time it's going to go back to the dummies. So we're centering the variable by removing the threshold. This is this x2 here that is really not x squared. It's the 0x2. So we take the variable x minus the 0 and then we have piecewise linear pieces that are attached to each other.

So here, now we have-- just by putting dummies and estimating a linear regression where the regressors happen to be this little line, we have something that is entirely linear, and yet espouses the kind of complicated shape of that function quite nicely. And now you can have-- again, you can ask yourself the same question, which is that's all very nice. But how many intervals should we choose?

And the answer is the same, it should be very familiar by now, is that you know-- assume you know, so by assumption, you know what the interval should be, what are the relevant intervals. For example, because the shape of this function is given to you by-- I have to invent what this function should be. What do you think this function is? Does anyone has an idea? I don't know. Yes?

AUDIENCE: Maybe age.

ESTHER DUFLO: Something as a function of age very good. So for example, is your speed of--

AUDIENCE: Running a mile.

ESTHER DUFLO: Well, but then it goes back up again. Yeah, and then it flattens. So over the-- I'm trying to think of something. Or let's say it's the speed of-- oh yes, I know what it is. It's the ratio of the speed at which you can move over your IQ.

So when you're a baby, you really cannot move very much, and then it goes very up again as you start walking. But you are not very developed yet, so this is true. And then it-- actually, it should go down.

And then the point is that it increases again later because you start getting a car. So this is the moment where you need to take care of your children very carefully when they're two and just start walking, and when they are 17, and start driving.

So suppose this is the economic theory. Or in this case, I don't know, psychology gives us that idea. Then we just set up the interval and we are going to go linear function in between the points.

This is when I should have put a photo of my children. I was looking for an excuse to put a photo of my children, since she put a photo of her dog. That would have been-- that would have been the time.

So suppose we know that. Then we're just going to impose it and run the regression like that. And then we just, again, are purely in the OLS world, except we happen to have this function that has three, four pieces to it. So that's one.

So then we cut it the way that theory tells us to cut it and we're back to purely OLS world. Or we don't know and we are trying to guess the shape of this function. And then we are back to what we just explained with series, or what we did with OLS, which is the more-- we are now trying to guess the shape of the function. The smaller the interval, the more closely, we are going to be to the shape of the function. So the less bias, but the more variance for a given sample size, because it's going to start trying to fit little squiggles, which might not be really there in the real function.

So we, again, have this same usual trade-off between bias and variance. And as the number of observations goes up, the interval size will go down. So in fact, what we are doing here when we do that is run something called a third nonparametric method.

We've already seen kernel. You've seen series a minute ago. And this is local linear regression, which is exactly the same as kernel.

So we divide the sample into bands. We use-- and then the bandwidth will go down as the number of submissions goes up based on the cost validation criteria, exactly as a kernel, except that instead of running-- instead of taking the average of the y, which is the kernel method, where we take a weighted average of the y within the bandwidth that we are looking at, we run a little linear regression of the y on the x in that interval, still weighting it such that the observation nearer to the x have more weight. And we take the predicted value of this function as our prediction for y for this point.

So it's very, very similar to kernel, except that instead of being a mean, we're taking a linear regression. It turns out that this is actually a much better way to do things than kernel regression at the boundaries, because the problem at the boundaries is that kernel regression is taking a weighted mean, but there's nothing to the right, so it only takes the left. And therefore, it has a lot of bias at the boundary.

So the regression doesn't have that problem because it's extrapolating. So it is not biased at the boundaries. So it is an advantage of local linear regression.

You might say, well, most of the time, I don't really care about the boundaries. It is what it is anyway, and it's not so many observations anyway. So there is lots of variance. There is nothing I can learn from the boundaries, so kernel that is going to do just fine.

But one specific example where it's going to matter a lot is one we are going to see in a minute, which is where, for some reason, there is something in the middle of the data where there is a jump. For example, your probability to get admitted to a school-- to a particular school, say Boston Latin, strictly depends on the score at an exam. So your probability to be admitted to the school goes from 0 to something positive at the score.

That's the probability to be admitted. So if I look at SAT score later, I'm going to find if Boston Latin is, in fact, effective, I would find something where it goes like this and then it jumps at the boundary. It jumps at the threshold.

So here, and I'm going to describe that in much more detail in a minute. But if that's the world, I'm particularly interested in what's happening here. So I wouldn't want to use anything like a kernel over here, because I'm going to get biased on both ends. So I'm not learning anything useful.

So local linear regression has become a very popular tool among economists rather than kernel. Also, if you can run a kernel, you can run a local linear regression, because it's really just the same. Another advantage of it is that we get-- the derivative of the function is simply the coefficient on the local linear piece. So sometimes-- often, we're interested in derivatives and we get it free of charge instead of having to recalculate it.

So for this reason, like this. So this is-- why on Earth would you like to do that? For these two reasons.

So all of that to say-- was to try to give you both a sense of how flexible the OLS model is and how rich it is, and how everything is connected in a way between nonparametric OLS model and even machine learning.

Now I want to put all of this functional form discussion we've had together to discuss one tool that has become very popular in applied research in recent years, which is called regression discontinuity design. So it's really a nice example of applying all of these methods of playing with functional form, and dummies, et cetera, to understand something. So regression discontinuity design is appropriate in any circumstance such as the Boston Latin example where there is some treatment, for example, being admitted in Boston Latin, which shifts discontinuously with some running variable at a threshold.

So call the variable a. That's because I'm using someone else's notation. Usually, we use x notation. But call this variable a, for some reason, and call your treatment Da equal a1, if a is greater than some threshold. For example, here I put the threshold as 21, let's call the threshold alpha 0, and 0 otherwise.

So we're going to use, as an example, a very nicely done example in Josh Angrist and Steve Pischke's book *Master of Metrics.* They kind of dissect a nice article by Carpenter and Dobkin on [INAUDIBLE] which I'm going to ask you in a minute.

So it happens to be here that the threshold is 21 and Da is equal to 1 if you are above the threshold and 0 otherwise. What do you think the treatment is in that example? Something of interest that shifts discontinuously at 21. Yeah.

AUDIENCE: The drinking age.

ESTHER DUFLO: Drinking age, exactly. So this is something that is relevant to many of you. So you pay close attention, even if you're not interested in econometrics. Da is 1 above 21 and 0 otherwise.

So the idea is that we're going to look at an outcome. What do you think-- what would be an interesting outcome to look at that is related to drinking age?

AUDIENCE: Hospitalizations.

ESTHER DUFLO: Yeah, or even one step further.

AUDIENCE: Car accidents.

ESTHER DUFLO: Yeah. So they look at fatalities, actually. They look at motor vehicle fatalities. And the idea is that as you become-- that's related to my example with-- my example with mobility over IQ. So the-- well, that's not-- I guess because it's IQ that goes down once you've drunk.

So the idea is that as people become older, they probably become more reasonable and better drivers. So motor vehicle accidents should, in general, probably go down with age. Except that at 21, something discrete happens, which is you get the right to drink. So you might be starting to drink-- drive and drink at 21, which was not possible before.

So the idea in regression discontinuity design is that you assume that your outcome of interest does vary with something that determines-- that is going to influence whether or not you're treated or not. So the potential outcome is related to whether you're treated or not, because the same thing that affects the potential outcome also affects the treatment. But there is a discontinuity jump and the assumption that we are willing to make is that there is no other reason for the discontinuous jump, other than the fact that the treatment has changed. So there is no other reason why you would become a worse driver at 21, other than you might have drunk before. You might have been able to buy alcohol before.

So it's important to know that this is always an assumption that there is no shift at 21 in the outcome variable and that the assumption is not usually directly testable. You have to appeal to your knowledge of the institution to know that it is likely to be a valid assumption. Sometimes, you can provide some little auxiliary regressions to give you confidence in the assumption, but that's always an assumption.

So what can you do in this setting? Well, the simplest thing you can do is to use a dummy variable for treatment to shift the intercept at a0. So what you assume is that-- we're running the regression that we saw with dummy variable, we assume that the shift-- the slope of the relationship between your outcome on the left, and on the right, of the discontinuity is the same.

But then there is a treatment. It suddenly shifts up or down, depending on whether you can drink. So given what you were discussing before, before I show up the graph, how do you think that motor vehicle fatalities will look like as a function of age? If I'm taking the predicted value from this regression, what will the-- putting 21 here. How is it-- how should it look like?

Yes, exactly. So it's decreasing till here. Then maybe it's going to shift up because of the ability to drink, and it's going to be decreasing again.

So this is a regression that you rarely see in practice because quickly enough, people go and do more richer stuff. The very nice thing about Josh's book is that he's going step by step, and this is the graph that he ran for US. So the dots are for each age what is the average death rate from all-cause, actually. And the lines is the predicted value from this regression, and with the jump, with the intercept shift over here.

So that's the simplest thing you can do. So here, these lines are parallel because you've assumed that they were parallel. The estimated lines are parallel. In fact, it's not a bad-- it doesn't seem to be a very bad approximation of the graph, but it's assumed in this case. Yes.

AUDIENCE: Is this similar to the China and pollution example--

ESTHER DUFLO: Very similar to the China example, exactly. Exactly. This is exactly the same idea. What I was just describing about Boston Latin, very similar. So it's very, very popular. People use this at boundaries.

So that's the simplest thing you can do. And in this case, it's sort of nice. It's nice. It looks reasonably convincing that there is a shift upwards in mortality at 21.

However, the simplest analysis may get it wrong. So these are fake examples. These are fake examples of data where nonlinearities may disguise themselves as discontinuities.

So if you're looking at the top graph, if I force the data to have a jump, it might think that there is-- it might think that there is a discontinuity. But in fact, it's just that it might just be a nonlinearity.

But because I forced linear on both sides and I forced it to be with the same slope, I'm thinking that there is a jump, where, in fact, really-- when I estimate my regression, I'm going to really see a nice significant jump, when in fact, in reality, it's kind of-- it would be a-- it could just be well captured by a smooth quadratic.

It's very clear as well here that if you force here, it's not estimated linearly. But if you are in this one, see, if you force a linear trend, you're going to see a jump. But in fact, if you did something more flexible with just a cubic, or a cubic and a quartic, you would see that it's just a shape that is nonlinear.

So that's kind of a problem we might have if we run regression discontinuity analysis by imposing a linear trend on both sides. So what do we do? Well, we try to avoid imposing a linear trend on both sides. So the first thing you could do is to say, well, I'm going to add polynomials exactly as we said.

You could either estimate it on either side-- on the same one on both sides, so estimate the same polynomial on both sides of the discontinuity. I think this is what is being done here in the second graph, and imposing a jump. Or even better, you could say, well, let me estimate-- so under age 0, the hypothesis that there is no discontinuity, that's clearly the right thing to do.

Or even better, you could be even more flexible and say, let me estimate a polynomial that is going to be different on either side. So you center the variable, as we discussed previously, such that the pieces of the polynomial meet each other, and you run a regression where you have ai minus a0, and then ai minus a0 above the threshold.

ai minus a0 squared, ai minus a0 above the threshold squared, et cetera. And here you can do just-- you can do this version without the dots, where you have a linear piece on both ends, but you're not assuming it to be the same. Or you can be more flexible. You can have quadratic on either side, for example.

Here, again, this is an example of that. This is the same data we saw earlier, but now we estimate a linear model on both sides. But it's not the same slope.

So we allowing the slope to be different on either side by interacting with-- by centering the variables and then interacting with the treatment dummies. And then the graph that is a little bit smoother is a quadratic on either side, a different quadratic on either side. You can see here that it makes no difference because it seems to be a nice linear functional-- nice linear line either way that's almost parallel, except for the jump.

So that's what one could do. Sometimes by doing that, make your discontinuity disappear. So it's a little bit sad, but that is life.

Once you run that, what is of interest? So you always want-- if you're interested in doing regression discontinuity for your paper, all the power to you. That's great.

You don't have to find something. It could be a 0, no problem at all. That's actually a general point. Don't feel that you have to deliver a result that is not 0. We like zeros. I love zeros. I don't see-- I think that there is not enough around. So just show us 0, no problem.

But if you're showing a regression discontinuity, I want a graph. You cannot do a regression discontinuity analysis without a graph. If you don't see it in the graph, it's not there, even if it's in the table with significant numbers. So it's very important to show the graph.

But then typically with a graph, you're also going to show some numbers. So let's go over this table. So each row in this table is a different outcome. So this is not your typical way that regressions are represented, where in the row, we have the different axes.

Here, each row is a different regression. So each row is a different regression. Each column, column one and two are, looking at age 19 to 22, so around the discontinuity. Column three and four are restricting to age 20 and 21. And it's age in month over these two things.

Column one, you have just the age. So this is a linear on both sides without imposing the same slope, the first model we looked at. Column two is age and age squared interacted with over 21 dummies. So that's the second model we saw with more flexible control.

You can see that the first row is all death. And so what we have in the coefficient is the jump measured after those controls. And so we have-- for example, for the first row, first column, the effect of-- the effect of the 21-- of becoming 21 and therefore being allowed to drink is 7.66, and I think it's per 1,000 or per 100,000, or something. It's per some large number.

And it's very, very significant. You have the standard error below. So if you divide one by the other, it's something that is going to be quite large. So you can definitely reject the hypothesis that it has no-- the drinking age has no effect. So keep that in mind and be careful.

And you have-- so once you've done that, the good thing with this particular variable, again, you cannot test the hypothesis that there is no discontinuity-- that there is no other reason for the discontinuity except the drinking age. But if that were the case, then stuff that is not related to alcohol shouldn't jump.

So what you can look at is other outcomes. So first of all, because you are thinking that most of these things come from motor vehicle accidents, you can see that the coefficient for motor vehicle accident is, in fact, very large and significant at second line. Suicide also, actually, affected by alcohol.

You could say, well, that's a fail, but probably alcohol is involved in some suicide cause. Homicide is not at all significantly related. So it's not violence in general. Other all internal causes definitely should not be affected, is completely unaffected.

And of course, alcohol-related causes-- alcohol-related that are not motor vehicle accidents or death. So for example, people passing out in a coma and dying is a much smaller number because it doesn't happen very frequently, but it's also quite significant. So this gives you-- so looking at the other variable is not a proof that the specification is correct, but it gives you some confidence.

So that's the first way to solve this problem, is to impose a polynomial. Second way to solve this problem is by narrowing the estimate to a band around the discontinuity. That's going to be our bandwidth. And as usual, you're going to make the bandwidth smaller and smaller, so look at the lower and lower bandwidth around the discontinuity as you have more observation.

So that becomes nonparametric regression discontinuity design. So parametric regression discontinuity design is when you use a bunch of your data and impose a polynomial form on both sides, and nonparametric is then we're reducing your bandwidth. And then what you do, typically, is estimating a local linear regression, as we just discussed, on either side. And you're reducing the bandwidth as much when you have more, and more, and more observations. So again, very, very similar, but it's just you're reducing the size.

So in my series of advertising graduate student work that could give you an idea of things to do, here is a regression discontinuity design that Melanie over here ran, which is looking at something that's a popular design for regression discontinuity. A popular place where you find discontinuities is elections.

Because in elections, the first person wins and the second person loses, irrespective of how close they are from the first person. So that creates a discontinuity, assuming that there is a bit of randomness in whether some supporters are going to vote that day or not, et cetera. That creates potentially very discontinuous outcome at 50% if it's two candidates, or at whatever the margin of victory is.

So people have looked at this. In political science, regression discontinuity is really a tool of choice in the last few years, as people look at the effect of Democrats versus Republicans, having a union or not, incumbency effect, so being the winner rather than the loser of an election, et cetera. So what Melanie looked at is the effect of winning an election on the decision to run again, not on winning conditional on having a condition on running, but on the decision to run again. So which way do you think it goes if you narrowly win an election versus narrowly lose an election? Which makes you more likely to run again the next election?

AUDIENCE: Winning.

ESTHER DUFLO: Winning, yeah. So the winners are generally more likely to run. So what she is interested is to see whether that's different for men and for women, whether the winning margin-- whether the effect of winning an election is different for men and for women. So she uses data that's all publicly available, so local election returns from 1995 to 2012, and now there is even a few more. This is where the data is available.

This is a bunch of elections, like school board, city councilors, this type of stuff. So very, very local election. Usually for people, the first entry point into politics, so local elections. Sometimes, for a lot of people, actually, the last entry point into politics.

A nice interesting way-- the reason why these are interesting elections is that they are usually not party line-based. It's a lot of individual decision whether to run or not. So the interpretation of the result will be a little bit cleaner, in term of people's own decision, as opposed to party deciding to field people or not field them.

So here is what we find in graph. What has she done here? Well, she's estimating a parametric RD for men and for women separately. The women are on the-- are the green dots, the green diamonds, and men are the transparent round circles. And then they both have the polynomial fit on both sides. I think it's a cubic. We'll see the regression in a moment.

So it's a different cubic on both-- it's a different polynomial on both sides and it's different for men and for women. You could imagine running in a two separate regression, one for men one for women. In fact, I'm going to show you that she does everything at once using interactions. So that's going to remind us how to use interactions as well.

And so when you look at these graphs, what do you-- so let's focus on this one. What's your sense of what we are going to find in-- if you want, you can say for men, for women, and then the difference between the two in panel A here. Yep.

Men are more willing to run again, even if they lost.

Exactly. So men are more willing to run again, even if they lost. So for both genders-- how would I describe this? I would say for both genders, there is a pretty steep decline in the probability to run again if they have lost an election before, but the effect is larger for women.

And interestingly, this is one-- this is the number of observations, but it's a 1 if you run and you win, and a 0 if you don't run, and a 0 if you run and you lose. And what do we have-- what would be the conclusion, do you think? We don't have standard error on the graph, but what do you think the conclusion would be when we are comparing these two?

AUDIENCE: Maybe the women that run and win again are different than the women in the first graph.

ESTHER DUFLO: Yeah. So first, let's interpret the basic fact. The basic fact from this graph is that actually, if we use one running and winning, there is no difference left between men and women at the discontinuity.

So men are much more likely to run again if they have lost the election, but they are not likely-- they are not more likely to run and win, which, as we were told, has to be mechanically due to the fact that the marginal loser among men is more likely to lose again if they run, to lose next time they run than the woman. So basically, women do not run unless they are pretty sure that they have a good shot. So let's look at it in a-- yes?

AUDIENCE: I was just wondering, it just seems strange. Why does the margin of victory increase for a little [INAUDIBLE]?

ESTHER DUFLO: Yeah. It goes down again. So this part of it, I have no idea what this is. And it's hard to interpret because it might be that-- it might be that they change elections.

Also, they are not in the local election database. Maybe the very, very popular women are not snapped up by parties to run and other things. So I've just no idea.

It could also be very different people at this point. This is hard to interpret causally, because there is a lot of differences between people according to their margin of victories. The jump, we can interpret causally, because we assume it is random.

AUDIENCE: You said this-- based off the same question. Is this all California data?

ESTHER DUFLO: All California, yeah.

AUDIENCE: This is just for local elections?

ESTHER DUFLO: Just for local election. So there's not the--

AUDIENCE: So it's very possible that [INAUDIBLE] for more successful women, they just get--

ESTHER DUFLO: They moved up. Yeah, exactly.

AUDIENCE: --international elections.

ESTHER DUFLO: I think it is a plausible interpretation. I could dream up many other ones, but if you want to interpret this as causal, it's a plausible interpretation. So we could add them up, especially in California. Yeah.

So what do we see just at the time of writing it as a regression? So it's kind of a bit of a mouthful, the regression, if you want to write everything together because we are running it together for men and women. So everything is interacted with a female dummy.

So basically, if you remember how we interpret interaction with dummies, because here I'm putting a female dummy, all of the terms that are not interacted with the female dummy will give me the effect for men. And then the ones that are interacted with the female dummy will give me the differential effect for a woman.

So you have all the definition here. So the beta is going to be the effect of having lost for men after controlling for the polynomial on the left side of the discontinuity. This is the eta case here. The polynomial on the right side of the discontinuities are the data case side. So the data, etas and thetas give me the effect for men.

I mean, the beta is the effect for men and the other ones are the control for men. And then the gamma is the differential effect for females, so it's going to be negative in the effect of running again. Delta is just the fact that women may be less likely to run in general.

And then the control-- I'm also interacting the control for female to make sure that I have the right control on both genders. So if I run exactly this regression, actually, I'm going to get strictly the same coefficient, and in affront to regression separately.

So here we have everything. The reason why we are not going to get-- we're running everything today. We also include the year of election fixed effect.

So it's a bunch of dummies for each year the election was run in case something is different. And then county fixed effect. Forget about clustering. We didn't discuss that very much.

And this is the way we would typically represent this type of regression. You don't plot the polynomials. You don't describe the polynomial in your table. It's not very interesting.

But you're saying, well, this is my polynomial. Here, I just put a linear term. This is a quadratic, this is a cubic.

So I'm exploring the sensitivity to different types of polynomial. And she reports the effect of having lost. So remember how we interpret the lost dummy here. What's the interpretation of that?

It's the effect of a lost election for? For who? The lost dummy's the effect of having lost the election for?

AUDIENCE: For males.

ESTHER DUFLO: For males. The female dummy is what? The difference between the probability that a female runs and a male runs, among those who have won before.

Because it's not interacting with lost. And you can see that basically, if they have won, they are just as likely to run as men. The female times lost is the extra effect of having lost for female. So if I wanted the overall effect of having lost for a female, what would I do? What would I need to do?

AUDIENCE: By this female-- loss from female loss.

ESTHER DUFLO: Exactly. I'm adding them up. I need to add up the effect of lost and the effect of female time lost to have the overall effect for females, because the interaction tells me that females are 11% chance less likely to run than men if they have lost.

So on average, they are-- so in total, they are 18% or 28% less likely to run if they have lost than if they have won. And then you can see that adding more polynomials doesn't seem to do too much to the interaction coefficient, which is nice. That's what we are interested in.

And then you can see the coefficient of running and winning. There is still a negative coefficient and significant in the first column, but it disappears afterwards. So it seems that there is no additional female effect, or it becomes much smaller and insignificant for running and winning.

So then you can think about how to interpret this, why it might be the case. There are any number of interpretation one could give. Yeah.

AUDIENCE: So if [INAUDIBLE]

ESTHER DUFLO: So in a way, yes. What we are doing here is combining difference-in-difference with a regression discontinuity framework. Because we are comparing the losers and winners for male versus for female. And we are saying, is it the case that the female losers are more likely to not run than the male losers?

But in this case, we are not doing the difference-in-difference for identification purpose. We are just interested in the difference per se. The males are not a control here. They are something that-- they have their own effect, and then the effect for female is even larger.

So regression discontinuity is nice because there are actually a lot. Once you start looking for them, there are a lot of those discontinuities happening in the world that we can exploit. One thing that you need to remember-- two things that you need to remember.

One is that the effects you are estimating are very, very local. They are right at the discontinuity. Sometimes it's a problem, sometimes it's not.

Think about the Boston Latin example. If you use the discontinuity to estimate the effect of going to Boston Latin, you're comparing the last student to be admitted in Boston Latin to the first student not to be admitted. And this effect might be quite different than the effect for, for example, the best children, or even the average children admitted in Boston Latin.

So if, as a parent, you're interested in whether you should go to-- you should attend Boston Latin as opposed to any other school, any other Boston Public School, maybe the regression discontinuity might not be relevant if your child is doing very well.

In some cases, it doesn't matter. Or in some cases, it's exactly what we're interested in. But it's worth remembering that. The other thing that is worth remembering is, again, this is an untested assumption, and very fundamentally untestable. So you need to argue your best that it's a plausible assumption in your context.

I knew that I had some more material to cover from last time, but I didn't know that it was that many. So let's talk about omitted variable bias. So I'll give you another example that comes from Angrist's and Peschke's book, again, because it's done in a very nice pedagogical way, of the omitted variable problem.

So imagine that you are interested in estimating the effect of attending a private university as opposed to a state school on future earnings. And imagine that the true model-- the true model-- in reality, the true model is this one, is alpha, a constant, plus beta, going to private school. That's the effect of going to private school as opposed to a state school. I'll discuss that in a moment.

But you need to account for the fact that people who go to private school are different. So in particular, people who go to private schools may have higher SAT scores, SAT scores. And they may have different demographics. Their parents might be richer. They might be coming from different towns, et cetera, et cetera. So these are demographics, but think of these as parental income.

So I'll define the group-- what the group-- what this group thing is right now. What is this group thing? It's a bunch of dummies, 150. And group ij is a dummy equal to 1 if students, i, belong to group j.

So it's a set of group fixed effect. I had condensed the interpretation by putting just alpha j, you remember? So this is the dummy spelled out. And what are those groups? These groups describe the set of schools that the students have applied to and where they were admitted.

So basically, you have-- this is a data set called college and beyond, where people were looked at in 1996. And these were people who applied to college in 1976. All of them applied to college in 1976. And what you know from every person is all the schools where they have been applied and all the schools where they were admitted.

So the first thing that they do in these 151 dummies is separating the schools by state and private and by highly selective, kind of selective, not selective at all. And those dummies are described exactly with the type of school that people have applied to. For example, all the students which have applied to three selective schools and one nonselective school, and all of them were private schools, and got admitted to all the places that they applied to are in one group. And then you have another group made of one state school and two private schools, all of them selective and got admitted everywhere is also another group. So you form groups like that.

So a lot of groups have just one person. So these groups, basically, will drop out-- the students will drop out from the regression because they have no one to be compared to. But some of the groups will have-- 150 groups have more than one person in them.

So we are interested in beta here. We are not interested in any of these. In the effect of SAT, we are not interested in the effect of parental income and in future earnings.

One might be, but in this case, let's say we are only interested in beta. So why are we putting all the other variables in the full model? Why do all other variables belong to the two model that describes our data? Why is it the true model as opposed to just a private school dummy? Yeah.

AUDIENCE: If you omit them, then the error term may be correlated with other terms.

ESTHER DUFLO: Yes. If we omit-- so we are going to have-- they belong there because typically-- we'll go back to your potential outcome framework. Typically, people who go to private schools, private universities, are different than students who don't go to private-- who go to state universities.

First of all, they qualified, so they have large SATs-- maybe they have larger SAT scores. Second of all, their parents can afford it. So maybe they are richer. And we think that maybe SAT score would have an independent effect on your earnings.

Why do these guys belong to? We don't think that the-- do we think that there is a causal effect of-- there might be a causal effect of your parental income on your earnings, because your parents can buy you a factory or something. Do we think there is a causal effect of the fact that you applied to exactly three schools and got admitted to two?

No. So what does the group do for us? Why do they really belong to this-- why are they likely to belong to this model? Why is it the case that these dummies might not be zero, that they really belong to the model? Yep.

AUDIENCE: Why is earnings-- we have some student who is highly desired in five colonies. That might be indicative--

ESTHER DUFLO: Exactly. If conditional on SAT, a student, for example got admitted everywhere they applied, then that suggests that there is something about that student that may have been in their essay. By the way, if you do read this chapter, there is a fantastic essay for a 19-year-old who subsequently went to NYU. So you see this essay and you really want, using this person, he's going to do well in life.

So of course, we don't see the essay in the regression, but the fact that someone got admitted everywhere they applied might indicate something about their potential earnings, irrespective of their education. So that's one. And then-- so that's the admission part. And then even in the application part. Do you think there is something in the application part?

AUDIENCE: Is it a proxy for motivation--

ESTHER DUFLO: Exactly. Motivation, ambition, et cetera. So people who-- it could be that people who apply to more schools just are go-getters. Or it could be that people who apply to just one school, and it happens to be MIT, know exactly their worth. Or whatever it is.

We don't need-- the beauty of it is we don't need to exactly understand what's the mechanics of it, because we aren't interested in the group dummies per se. But we we're thinking that there might be a lot in those group dummies that represents something about the students.

So assume that happens to be the true model, that God told you this is the true model. Now we argued it because God has not told me anything of the sort. But this is the true model.

But imagine that we don't have all these variables. In particular, it's going to be very rare to have the-- to have all these application dummies. So typically, in the typical data set, we might have earnings and whether they went to a private school or a state school.

Maybe if we're lucky, we also have the SAT score. Maybe if we are very lucky, we have parental income, and we are unlikely to have all these group dummies. In this case, we happen to be able to compare for pedagogical purposes. But in a lot of cases, we won't have the group dummies. We would like to put them in, but we just don't have them.

So what happens when we run this regression? So I want you to start, maybe, with the column six. Column six estimates the true model, which has all of the group dummies over here. They are not reported, because that would be a lot of coefficients. So this is how we report in a regression, selectivity group dummies. Yes.

Then we have private school dummies. This is what we are interested in. We have the student SAT score, we have the parental income, and a bunch of other demographics, race, top school, top percent of their high school, blah, blah, blah.

So that's the truth. How do you interpret the first coefficient here? Yeah.

AUDIENCE: I had a question about one of the variables. For high school, range missing, that doesn't mean the data is missing?

ESTHER DUFLO: It means the data is missing. So we don't want to drop the data-- we don't want to drop the guy, since we're not so interested in that variable, per se. If it happens to be missing, we can just fill it up with zeros and add a dummy for whether this is missing.

That way, we don't lose the observation, but we estimate the coefficient properly for the person. We didn't go over missing data, but that's a good question. So the high school missing is a-- school rank missing is a dummy that the high school rank is missing, and therefore that this one was filled up with a 0.

So let's focus on the coefficient beta here in model six, which is the true model. What's the effect of private school on future earnings? Yeah.

AUDIENCE: Does it tell us that your earnings are 1.3% higher if you went to a private school than a public school?

ESTHER DUFLO: Exactly. And what is the standard error of that number?

AUDIENCE: 2.5%.

ESTHER DUFLO: 2.5%. So is it-- can we reject the hypothesis that the private school has no impact? No. So I'm very sorry, but there is actually no effect of-- it seems that in the true model, it seems to be no effect of going to private school or a state school in the true model.

AUDIENCE: What's the number real quick we were looking for? We were comparing 2.5% to 0.05%, or--

ESTHER DUFLO: No. We are comparing the ratio of 1.3 to 2.5, which is going to be about a half. We're comparing that to, if you're interested in 5% level, we're comparing that to 1.96. This is the typical-- I'm doing-- you will become very conversant with doing this in your head.

AUDIENCE: We just do it all in our head.

ESTHER DUFLO: We just do it all the time.

AUDIENCE: Yeah, that's what I wanted to ask.

ESTHER DUFLO: Divide the coefficient by the standard error.

AUDIENCE: And then--

ESTHER DUFLO: You get the T statistic. The T statistic is 0.5. Is it larger than two or smaller than two?

AUDIENCE: Two.

ESTHER DUFLO: Two. It's 1.96, if they are normally distributed, which is probably true. It's the largest sample anyway. So you can think of 1.96 being the threshold.

Is it below 1.96 or above 1.96? It's much below 1.96. So even if this regression-- these tables had stars, which it doesn't because Jess shares my opinion about stars, you would not have a star here.

In your paper, you should put stars. Because that way, you will get used to put stars to do these calculations. So this has zero stars. It's 1.3% if you went to private school, it's not very large. And it's not significant.

Now go back to the first column where we are just putting it, the private school dummy. And there, what do we find? We find 13.5% and that is very significant.

So if here, we have estimated the true model, there is no effect. But when we omitted a bunch of the control that we should have put in, we are seeing-- we're seeing a very large effect. And what this suggests is that this effect, actually, is a biased number coming from the fact that we haven't controlled for stuff that we should control for.

And you can see it. Now we're going to go progressively from column one to column six. You can see that already, when we put the student SAT, score the effect goes from 13.5% to 9.5%. Adding the parental income and the other demographics do a little bit more.

And then adding the group dummy, which is going from this guy to equation six, that's where the coefficient really goes down all the way to zero. So that's just the group dummies that capture like the spunk of the student on the one hand and their desirability, as you pointed out, on the other hand, makes a big difference.

Another thing that is pretty interesting, and we'll go back to that in a minute, is that once you've introduced those group dummies, actually, adding the SAT and the parental income doesn't make a big difference. So the group dummies, the application behavior plus whether or not they were admitted, seem to be doing enough of a job. So what I want to do-- so this is the omitted variable problem in a nutshell. If you omit variables that really should be in the model, the coefficients are very best.

And then this one tells us a slightly more complicated story, because we know that they belong to the model they are significant in the SAT score, and parental income are significant in the decisions to-- in the earning regression, but yet they don't belong once we put here. So we'll need a little bit more math before we understand that number, which I'm going to do in the next 10 minutes.

So what is happening? So by the way, a note that is important is that when we do the group dummies, we are missing a lot of observations because anybody who is in a group of one disappears. So it turns out that we can pretty much capture all of what the group dummies do for us by putting a much sparser control that we can have for all observations.

So we are comparing, really, the same sample. So the average SAT score where you applied when you send two applications, send three applications, send four or more applications. So these are dummies that you send two applications, three applications, or four or more.

That's pretty much enough to capture all of the group dummies. So the story that we give is exactly the same that we go from a very large regression-- a very large impact. This is now the same sample where all of these are defined.

So the coefficient here is a little bit different to what we had before, which was in the entire sample. And we go from 21% premium to a true one that is 1.7%. So this is the regressions that we-- these are the differences that we are actually going to try to explain now.

So now for some formula, I'm going back to a familiar notation for writing down the formula. Suppose the correct model has an x1, which you happen to be interested in, an x2 that really belongs there, but for some reason, you're not going to put in, for example, typically because you just don't have them in your data set, and you estimate a model that has just the x1.

I want you to define the ancillary regression as a regression of x2 on x1. And then you're getting this interesting formula, which is the omitted variable bias, which is the effect-- the estimated alpha minus the two coefficient, which is your bias, is delta 1 times beta 2. So in other words, the omitted variable bias depends, in this very simple way, of the influence of the omitted variable on the outcome and the influence of the x1 on x2.

Even though we don't have a huge amount of time, I might not go over the math of that. But I'll show it, it's in the slide, and show it to you in a minute. But I want you to think about what's the intuition for this result.

AUDIENCE: The correlation of the independent variable with the coefficient.

ESTHER DUFLO: Yeah. So that's just restating the results. But how do we give-- what's the intuition? So the intuition is that on the one hand, if we omit a variable that really doesn't matter, then it shouldn't matter. OK, fine.

On the other hand, if the variable that we omit is totally uncorrelated with what we put in, it doesn't matter either. Where it is going to matter is when it's very correlated with what we are trying to put in. Basically, x1 is going to have a very large coefficient because it's pooling the effect of x2.

So it's basically pooling with it a bunch of stuff that doesn't really belong to it. So that's why we have this product, is when the two-- so for example, if you're looking at a data set of kids and you just have height and you don't have age, you might think-- and you're looking at performance on some additions, on height, and you might think, wow, height is really correlated with performance i n addition. Well, it's because height is very correlated with age, and age is very correlated with addition results.

So when you put in the height, height is, unbeknownst to you, or perhaps beknownst to you, really proxying for the age. And so that's why the coefficient becomes large, even though coefficient is 0. So that's the formula.

That's just really great because once you have this formula, typically you don't have x2. Otherwise, again, it would be in the regression. But the formula helps us think about whether we should be worried about x2, and in what way.

The math for it is really, really simple. So we don't need to go over it. It's going to be in the slide. I put in for you the bivariate derivation and the multivariate derivation. So it's in the slide and it's easy.

It's better to spend our last three minutes to look back at our college example and see how it works. So this is a special example, because in this case, we have the omitted variable. We have the variable that would be omitted usually.

So what we can do is to run the auxiliary regression. So I'm going to run-- what happens? Well, I'm going to run the SAT score on the private school dummy, and the parental income on the private school dummy.

So let's say-- suppose I omit the SAT score and I'm running-- in the short model where I didn't have the group dummies, I omit the SAT score. The effect-- delta one here is 1.165. So people who are going to private school tend to have higher test scores, naturally. It's a pretty large coefficient.

Remember, this is table 2.3 that we are trying to compare to. Remember we have a 0.21 without controlling for SAT and then 0.15 with controlling for SAT. So what we're hoping is that the difference between the two is equal to delta 1, which is 1.165 times the coefficient of the SAT in the regression. And then miraculously, it's actually correct. So you can see that-- you can see the anatomy of the omitted variable bias in this regression.

Another thing that you can see from this regression is why the application dummies help. Because we have seen that the SAT score does matter. So why is it that it doesn't matter once we have the application dummy?

It's because once we control for the application dummy in the regression, the SAT score-- in column three here, we're controlling for the application dummy in the regression. And you can see that controlling for all the application dummy, the SAT score is not significantly related to the private score dummy. So in the omitted variable regression, in the omitted variable bias formula, what is happening here is not that beta 2 is 0.

SAT score does belong in the ending regression, but delta 1 is 0 once we have all the-- so in order-- once we have all the school dummies, the application dummies. In other words, the application dummies are doing a very good job capturing what should be in there, which is why the omitted variable bias disappears with the application dummies.

So how do we use the omitted variable bias in general? Most of the time, we don't have these variables. So otherwise, we would include them in the first place. So why is the omitted variable useful?

So it's useful because it guides our economic thinking on whether the bias should be important or not. When we're thinking that we're omitting something, is it something that we should worry about or not? Are we omitting variables that are important determinants of the outcome, and are they likely to be correlated with the regressor of interest?

I'm going to skip the example, but that's very fun. If you read the Well Blog-- I love the Well Blog on*The New York Times.* I actually do like it.

But one of the things that's very funny is that about every other day, there is an article that gives you a nice example of omitted variable bias. So here is a high-fat diet may lead to daytime sleepiness. Might well be true, by the way this is based on regression where 1,800 men who had filled out food frequency questionnaires reported how sleepy they felt during the day. So it turns out people who have more fatty diet also feel sleepy. This is controlling for a bunch of stuff. You could well imagine a number of omitted variable bias here.

But there's one every other day, like coffee makes you live longer, whatever. Might well be true. But this is usually on this-- based on these regressions, and that's a big problem with epidemiology in general, that it's just comparing people who are being very different behavior, and maybe there is a lot of omitted variable bias here.

I want to go over these slides quickly because this is our introduction for [INAUDIBLE]. Some hints of more advanced techniques to deal with omitted variables bias. First of all, you could-- so you could know exactly which variables belong, but there could be very many, and you might be wondering how to deal with them because you don't want a very, very, very long regression, because it's going to be very imprecise for your coefficient of interest to add a lot of regressors.

So in this case, one thing you can do is called matching, which is first, run a regression, which you can make as nonparametric as you want, of the probability to be treated, or your variable of interest, say, the probability to go to private school on a bunch of characteristics. Take the predicted value from this regression and then control flexibly, for example, with polynomial, with a series expansion, on this probability. So you do it in two steps.

Your first run an as complicated as you want messy regression of the probability to go to private schools on the regressors and then take the predicted value of that and just stick it in the regression. Which seems magic that it works, but it works. It's a result due to Rubin.

And then this is why I want this slide, is another thing you could do is try machine learning techniques, which is in this situation you would do when you have potentially lots and lots and lots of x's, but you do not know which one belongs, then you could use machine learning techniques to-- so of course, you're still interested in your private school, but you could do machine learning technique to select which one belongs in the model.

Depending on the technique you use, you might use-- you might do that in a training data set, as Lisa was describing to us, and then running the regression in the full data set. What I want to say about both of these techniques is that in both cases, they are just as good as the variable you do have at your disposal.

And the question-- you can add as much niceties as you want, but the question is always going to be what is it that is not in my data set which should really be there? And is it likely to be important or not important?

In some cases, it's not going to be important. You are pretty confident that you have a lot of stuff. In some cases, it might be important, in which case you might decide that you don't know until you run an experiment, or until you find another method. So that's where we are now. So we'll see sending on Wednesday.