Cyberseminar Transcript

Date: April 3, 2019

Series: HERC Econometrics with Observational Data

Session: Fixed Effects and Random Effects

Presenter: Liam Rose, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C:\Users\VHASLCMyersK\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.IE5\73RXHSNF\l)

Dr. Liam Rose: As Todd mentioned my name is, Liam. I am a Health Economist with HERC. I am presenting today on Fixed and Random Effects, and I want to mention that, Joe Jacobs, who is also a Health Economist here at HERC, did a lot of the work on these slides. And then Todd Wagner is our host today, and he is a fellow Health Economist of HERC and Director of HERC, and he’ll be handling the questions. And I encourage everyone to send in any questions you have, I'll try and get to as many as I can at the end and leave some time. And then Todd will direct any prepping questions that we might have during the talk. Okay, so everyone thank you so much for joining today, I really appreciate it. This is a lecture about Fixed and Random Effects. And I want to start by saying that these terms can be a little bit loaded and they can be somewhat confusing because they’re used a little bit differently across disciplines. As you might expect, today we're taking this from an economics and econometrics standpoint. So this is kind of how economists are going to be using this term.

And so I'm going to start off with a poll and that is going to kind of guide us a little bit. I'd like to know how much you know about it going in, and that will kind of push in which direction we want to go with this topic.

Rob: And that poll is launched, Liam. The question being: How familiar are you with the concepts fixed and Random Effects? Answer choices being; very familiar, somewhat familiar, and not familiar at all. And we have about 70% of your audience voted. It usually levels off not too higher than that, so we’ll give a few more moments to make their choices. Yeah it looks like it's leveled off at just under 80%. It’s at 80% now, so I'm going to close the poll and share the results out.

And Liam, 12% of your attendees say that they are very familiar, 73% say they are somewhat familiar, and 15% say they are not familiar at all. And now we’re back on your slides.

Dr. Liam Rose: Perfect. Okay, thank you everyone for answering that. And with that, tells me that we're going to take a little bit of a ground up approach and we’ll try and go through this as carefully as we can in the time we have. So I'm going to start with talking about panel data and the regression that comes out of that. And then how we might have heterogeneity and we might have confounding variables in that. And then we'll step right into the Fixed and Random Effects models and how to choose in some of the terminology in them.

So, to start off this is what panel data is. It's basically just repeated measures of the same observation. So this could be a lot of different units. Here I have a person. So there's two different people, I'm observing them every year and so I write down what year I observed them, what age, sex, income, and education. It doesn’t have to be that, it could be a repeated measure of individuals, households, countries, firms, any type of unit being repeated, we measure it over time.

And so what we do with this, this is a very basic panel linear regression model. It's like any other regression model, but we write it with different subscripts. So, YIT, here this is going to be outcome variable. Going back to our data from last slide, outcome variable would be income for individual I at time T. And here I'm going to take X, which could be any explanatory variable, I'm going to be talking about a relationship between education, income, longstanding question economics. So we have XIT here, where it's going to be education for individual I at time T. And then our error term is going to be basically everything else. This is, you know, it could be their age, it could be their sex, things we have measured. It could be things we don't measure, such as their motivation, or innate ability. And what we’re measuring here, what we’re interest in, is Beta, Beta hat, which is just going to be the estimate of the change in income associated with the unit change in education. So if you're comfortable with regressions this should be very familiar. It's all the same, except we add the time at aspect, because we’re, like I said, we have this repeated measure, we have the same, we have repeated measures of the same unit. And finally for Beta one to be unbiased estimate of the casual effect of education on income, X, this education, it has to be exogenous.

What does that mean? In math it means the expectations of EpsilonIT is 0. So the conditional mean is 0. It just means that the error term and our explanatory variable, in this case education, can’t be correlated. And when does that happen? Usually it happens. So we have XIT and EpsilonIT, so education and everything else. There is correlated when those omitted variable bias. So, you guys can't see me, but I would think of all, you could think of all the different variables that might be able to go in this regression that better explains income. Other types of things that we can have are sample selection issues. They could be correlated in certain samples than others. And simultaneous causality. Although that one is a little bit less frequent. So the opposite of exogeneity, or exogenous, is that if they’re correlated, and then X is endogenous.

So in our example here, we could say that, our variable education and the all other factors are endogenous, because they’re correlated in terms of things like demographics, family history, innate abilities or even things that are a little bit found on an individual level. We could have the state level of things, geography, demographic. Or if we're talking about education as we are here we could say maybe the school district or maybe parents education. Lots of different things can be stable over time, but not in our regression.

Okay, so what does that imply for our progression? So at the top here, the top equation, is our normal, is the regression we started with two slides ago or three slides ago. And then below it we break up this, all other terms thing. So this is our EpsilonIT, this is you know, our state level factors, our individual level factors. It’s all those things that we just mentioned in the last slide, and we break it into two things. There's the time bearing error component, and there's the stable component. So one of these things changes over time and we're going to call that U. And then we have this second one that's unique to an individual that doesn't vary over time and we’ll call that Alpha. So it’s very simple up here, if you’re not a super mathy person it’s just we're going to take the all of the factors and break it into two pieces. The parts that are different over time and there’s parts that are the same over time, and we call Alpha and U.

Okay, so then it's the covariant, meaning that if they relate to each other of X and Alpha, meaning in our example if they are equivalent to stating that the motivation or any other exogenous factor not related to education is equal to 0. So what this means is that they're not related to each other at all. So then if, but if they are related if the covariant is not 0, then putting this term in the error, i.e. not breaking it up like we just did that's going to be problematic and it's going to lead to biased estimates.

And here is a graph that’s showing what, this is from Clark and Linzer, this is a graph showing exactly what I was talking about. So if we don't account for this unobserved heterogeneity if we keep them both in the error term, as our previous epsilons, then we could have three separate estimates not knowing what this actual correlation is. So this is a simulation figure. If you don't take account of this and the correlation is negative then you’re underestimating if the correlation is 0 then you're getting 1. And if the correlation is positive then you are overestimating. So there's a couple of ways to deal with this but it's difficult because we're not always sure what the correlation is, especially if our error term includes something like motivation, or innate ability. It's extremely difficult to measure if not impossible. So there's some ways of doing this maybe ID is instrumental variables. But here panel data offers us a method of dealing with this. And the next couple slides are showing how.

So with that if we don't have an instrument or additional data if you don't have data on literally everything including their state, their Democrat graphics, their motivation and all those things, having repeated measure of the same unit allows you to control for these factors that don't change over time. And there's two ways of doing this. And that's what fixed and Random Effects are. This is our term output that I talked about. This is not the things that change over time these are the things that are stable over time. And we're going to talk about fixed affects first and then Random Effects.

Okay, so back to our model here. So this is a Fixed Effects model. Right now we have it broken down like I said before. So we took our all other factors term Epsilon and turned into Alpha and U. Which is the time varying U and the stable which is the Alpha. So instead of just having a single unobserved error which is the Alpha, we put in a set of fixed parameters. And I'm going to call those Mu in this scenario. And you just add all of those to the model. And what this does is control for the meta effect of all unobservable factors that are different across units but are constant over time. I think the easiest way to think about this is something like innate ability. It's definitely going to be different across people, or individuals. But we tend to think it's not, it's something if we're talking about innate ability, it's going to be the same over time for the same person. And that's why sometimes you hear this design being called the Within Design because we're looking at within a unit. In this case an individual. So each unit in this case is going to have their own intercept that’s estimated separately.

So how does this work? How do we actually figure out how to do this in practice? It's actually pretty simple. You just make a dummy variable for each individual observation or group. So whatever your unit is, in this case it's an individual, you make a dummy variable and if it's that unit then you make it equal to 1 otherwise it's equal to 0. So in our original data example I had two individuals, and so then they would give you two dummy variables and it would just be equal to 1 for each observation in a time period that it is for that individual and not the other one. So then our model just becomes the same as before except with as many dummy variables as you have units in your dataset.

So the only drawback to this is that, oh sorry back up, what this does is it’s going to absorb all that variation, that time invariant information, those things that change over time. It's going to control for all those things. The things like ability, they're going to be soaked up by this dummy variable that's the same for the individual and this regression. Practically though, this can be a little bit hard if you don't have a huge amount of degrees of freedom because as you might imagine if you have a really large dataset you could have this really, really large regression with a dummy variable for each unit.

So what I'm going to show you right now is that the Fixed Effects estimator is the same. This is what's usually done by a lot of packages which I’ll show you in Stata or R fast and it will be mathematically equivalent, but I think the dummy variable intuition that I just showed you is also useful. So what you do is you take, you get the time mean for each component. And what that means is at a time point what is the mean for each individual part of your regression. So we can call these, just call it the same variable as before but put a bar over it. So we have Y bar and it's just going to be the time mean of Y at that over T years. So this would be, in the example we showed earlier this is like taking income for person 1 over three years. So we had like 2010, 2011 and 2012. And we just take the average income over those three years. So then we have average income for that individual, we have average X, which is education in this case the individual. And then we have the error and the time invariant error.

Okay, but the nice thing about this though is that with the way we defined this, Alpha, is time invariant. So it's the same in every time period. So it's just the same as the Alpha as we said before.

So what you do is you take this Within Effects transformation, Within transformations, and you just subtract by the mean per individual and then you run that regression. So this is what's called a Within estimator, or a Within transformation. So we subtract the time means, like we said from the last slide, and then run a regression on these demeaned values. Demeaned not meaning that we were rude to them, meaning that there was, we subtracted the mean off of the value. So we're running LOS regression using this transform data in the form of deviations from that mean. To rephrase, we're taking the difference between that person's income in one time versus the average income for that person over the entire time period for which we observe them.

Okay, so to further show this, the regression doesn't include the constant terms for this because the timing of these factors will be the same as the actual values. So when you take out the demeaned part for over here on the right which are highlighted in red, the Alphas, which are the time invariant parts, if you take out the time invariant mean it just equals 0. This is kind of a really complicated way of saying the average is equal to the, an observation if the average is the same as every observation. So you're just taking one thing minus the same thing and it's 0. And then it goes away. And that's really nice because those are the factors that we don't observe that are stable over time, such as motivation or innate ability, but they’re no longer in our regression. Okay. So in Stata you’re a Stata user, this is actually quite easy which is really nice. You just have to set up, you set up your data to be panel data. I think that's XTset. And then you do XTreg and you put in the reg command both for whichever variable and commaFE which stands for Fixed Effects. And this will give you the regression that we’ve just outlined in the last few slides. So what it's done is it's in the background it's demeaning all the variables by individual, subtracting them off, doing the regression on the transform data and getting rid of all that unobserved, endogeneity that you don’t want. And this will give you this Within Unit estimator.

And equivalently if you wanted to do this in a little bit less black box way you can write this regular term if you're a Stata user you know of this one, reg for regress, Y your outcome, X your explanatory variable, and then I dot variable where variable would be your unit. So in my case with the example we've been going off of I dot variable is the I dot individual. And what that does is I dot in in Stata creates a dummy variable like I showed three sides ago for each individual. So this would be this very long regression with a dummy variable for each person or a Fixed Effect as we call it. And that will give you the same exact results as XTreg, FE.

Okay, so let's look at some of the pros and cons of the Fixed Effects approach. The biggest thing that I and a lot of economists would say is the best thing about it, is it gives you this unbiased estimate of a coefficient when the covariant between your explanatory variable and the time invariant variable is not equal to 0. This is just saying it gives you an unbiased estimate when you think that your explanatory variable is related into things that don't change over time but are not observed. So in our example again, this gives us an unbiased estimate when we think that our explanatory variable education is not related to things that don't change over time, like motivation and ability. So there's a couple cons. Estimating so as it says here estimating time invariant explanatory variables, the variables that change very little over time, is not possible. So there's, like if you look at the demeaning process you're taking those, that variation out. As you saw two slides ago, it's one thing minus the same thing and it gives you 0. You can't really get the effect of the state or something like that. And then the next con is that their estimates that can be subject to high sample-to-sample variability, meaning in different samples you might not get the same the exact same estimate. Especially when you don't have very many observations per unit. In our example dataset we only had, we observed those people three times. So that's not a ton, ideally you want this really long panel where you observe the same unit over time and you're seeing them again and again and again and again that's really nice. It gives you a much more stable estimate. But oftentimes it's not what we get and so when you do, you know you set up your panel data nice and tidy and whatever data software you’re using you might only have a few observations per unit. Sometimes it's used too. So that can be, it doesn’t give as good sample-to-sample estimates that are stable. And the last one is that out-of-sample predictions are not possible. The math behind this is pretty simple, but we won't go through it. But the idea here, is just that, that you don't know exactly what's happening if you were to observe that same unit again. So if you’re trying to do out-of-sample predictions, Fixed Effects are not the way to go. And that's not necessarily the case in Random Effects, as we will show in a minute.

So let’s go through a real world example, and see what happens. So this is from fairly recent ledge here, Oberg et al. where they’re looking, a study looking between, the association between labor induction and autism. So they have really great Swedish register data, if you know anything about this a lot of Scandinavian countries keep this really great panel data on people. So they follow these people from obviously their birth and they follow them over time with a lot of repeated observations. So they have one point three million births, twenty-two diagnosis and they're doing this within siblings comparisons.

So some things they know in this, they can control for some observable factors like they know which year it was, they know a lot of things about what's going on in that person's life, they know about the mom’s life down to their BMI and early pregnancy and other health factors. But, of course, there's still unobservables. Some environmental factors, and all genetic factors, are within families but there is going to be some variation there. And so what they do here is they look at within the maternal siblings pairs for any differences within the siblings.

And here is what happens. So this is a Hazard Ratio. First they run with all of these really rich set of covariates, we see this positive and significant association between labor induction and autism, as you can see in those first three boxes. It’s really quite striking, and this would be no, I think, policy relevant to a lot of people. So basically what’s happened here, they started with a model where they, where it was relatively small. They didn't have a lot of covariates then they added pretty aggressively more covariates, you can see the list right below. So model 3 is this kind of kitchen sink model where they have everything on it and it’s still significant.

But then what happens when they include the maternal siblings Fixed Effect, labor induction no longer associated with offspring autism. This shows that there was some unobserved, time invariant factors, shared by maternal siblings. Maybe genetic factors, maybe family level characteristics, we couldn't quite measure, even Swedish people with their great data couldn't quite measure. And that was showing up in models 1 through 3 when we were doing normal OLS but then this is biasing those estimates upwards. And that wasn't happening anymore once we include those Fixed Effects.

So a couple other notes about Fixed Effects before we move on, I will go through this very briefly. You don't only have to have one Fixed Effect, you could have a Fixed Effect for more than one dimension. And here I've included Fixed Effects for a person and a year.

And Fixed Effects is also a way of expressing difference in differences. This is often done in some clinical trials where you have repeated measures, and so you have that you can do a Fixed Effect being after the treatment and the group and then an interaction between them. And generalized you can do this for you know having a group, an individual and a group, and a time; this often works in situations where it's perhaps a person living in a city or a person in a particular arm of the trial in a certain time period. So I just wanted to mention that those things are intuitive.

So Random Effects, there's the other approach. So we've seen that Fixed Effects will tend to produce these unbiased estimates where the time invariant omitted factors are correlated with their aggressors. But what if that's not the case? What if you can assume that the time invariant stuff and the explanatory variable you care about is 0? Well in that case you can't simply run a pooled OLS, pooled OLS being just a regular OLS but with the panel data. This is where Random Effects are going to help us. If you can assume that there is no correlation between what you want to estimate and you unobserved heterogeneity, you don't need to use this Fixed Effect estimator. I'll just mention again, you can't use just pooled OLS, this is serial correlation issue. That just means that you're getting, that just means that the error term is correlated over time, because you're, and intuitively you might think it’s, “Oh, because we’re doing the same unit more than one time it’s obviously going to be correlated over time.”

So what happens there, unfortunately, is that it's inefficient. If you can think way back to your introductory stats course, that was one of the benefits of OLS is that it's efficient but if there is serious correlation in the error that's no longer true. It hurts our standard errors, and there's implications for hypothesis testing that are not good. So we don't want to just choose pooled OLS.

So instead we do this Random Effects model. It’s like Fixed Effects, it doesn’t eat up so many degrees of freedom. But it still accounts for the unobserved heterogeneity. Again with that assumption that there is no correlation between the time invariant factors and our actual variable of interest. It's kind of a compromise between those two models. The Fixed Effect that we got rid of this boxed Alpha term by doing this Within transformation that was the demeaning process. The Random Effect assumes that this Alpha term is a random variable in of itself. Usually assumed to be normal and it is going to have some mean 0 invariants of some arbitrary Sigma squared.

So just a little bit of math we’ll go through this too quickly, I just want to show, a little bit quickly, I just want to show that the two components that come into play. So Random Effects essentially transforms the Fixed Effects system with an inverse variant waiting Lambda, that's what that term is the squiggly L kind of thing, that's a Lambda. Then one Lambda is going to be 1 minus the square root of the variance of our error over the variance of our error plus T times the variance of our unit effect, essentially. So we're waiting by what we think the errors are, again we impose this structure, this normal structure, on our Alpha, excuse me on our time invariant characteristics, time invariant variable. So this is kind of a quasi-time demeaning of the system because again we're going to do this thing where we're going to subtract off, as you can see on the bottom of this screen here, by this Lambda term that we've calculated.

This Lambda is going to be between 0 and 1 because it is a waiting. So getting rid of these equations it's just going to show that, how it is a compromise between pooled OLS and Fixed Effects. When that term that we impose on our Random Effects is 0 and Lambda is 0 and Random Effects is equivalent to pooled OLS. So basically this is saying that, “Hey, let's completely ignore these time invariant factors.” Going the exact opposite way, if the variant is really high and our Lambda is 1 then Random Effects is exactly equal to Fixed Effects. Most of the time it is going to be somewhere in the middle. That's why I say it's a compromise between the two.

So essentially what’s going to happen is that the groups with, depending on what’s happening for a specific group, some will be closer to pooled OLS and some of them will be closer to Fixed Effects. And the effects here are going to be greatest when you don't as have many observations for a particular unit, so you know you only have two or three observations for a particular unit and when it estimates the variance for our time invariant thing are closer to 0.

Okay so how do you do this? So this is just kind of like what's going on in the background of our statistical package and you can see on the bottom here same thing, XTreg except FE you do RE. So what's going to happen is that first it's going to obtain an estimate of its waiting Lambda and then we will put that in and transform our OLS subtracted off and it will do this transformed regression. Again you don't actually have to do this, you could, but Stata will do it for you fast R they’ll do it for you. In Stata it’s XTreg,RE.

Okay, so some pros and cons here. First, as we noted, the estimates of Beta will not have as much variance because it's this compromise, right? Again, if we're going towards the Fixed Effect model that's when we're taking the full variance of our Beta, of our sorry excuse me, the full variance of our time invariant factor. But in Random Effects we're kind of shrinking that down and it will be closer to the pooled OLS estimate and so that gives us less variance on the data which could be very nice for some situation. The second pro is that we can actually include those time invariant covariates in the model. So you know if you're interested in the individual state that doesn't change over time, you can include that in the model and get an estimate for that. The third is that it doesn't use as many degrees of freedom so it doesn't take into account where there is this unreliability associated with estimates from small sample within units. What that means is, hey if you have only two observations from the same unit that Lambda that we estimate is going to take that into account. It will be weighted less because you only have two observations. Okay so what are the cons? The big con here is, and this is the main reason that a lot of economists and others aren't really interested in Random Effects, is that there is likely going to be some biased in your estimates of Theta. The problem here is you can never truly know the correlation between your variable of interest X and the time invariant factors Alpha, such as, motivation. And so if those are really highly correlated there's going to be more biased. And there's really not a great way to test that because those factors are unobserved. And then finally, again, a con may be that we don't may be that we don't actually estimate these unobserved factors they're treated as random variables.

Okay so I'm going to go to a poll here. The poll will be a little bit abbreviated on your screen, but from an econometric standpoint when is it appropriate to use Random Effects in place of Fixed Effects? And the first is, when the unobserved unit specific factors, Alpha as we call them, not correlated with the covariates that we care about, or when they are correlated or can they be used interchangeably?

Rob: And I just launched that pole, Liam. Audience members you can see that I had to concatenate some of the terms, but I think you get the point. First choice, when unobserved factors are not correlated. Second choice, when they are correlated. And third choice, they can be used interchangeably. And we only have about 35% of the attendees voted so far, Liam, so I guess we're going to have to give people a little bit more time this time then they needed last time around.

Dr. Liam Rose: Yeah. This is a tougher pole. Go ahead and vote if you're unsure, because it shows what the efficiency of the talk is. Go ahead and vote if you're unsure.

Rob: Things have leveled off at just about 70%. So I'm going to go ahead and close the poll and share out the results.

And what we have is that 51% of your attendees choose the first option that it’s when the factors are not correlated, 38% choose the second option when they are correlated, and only 11% say that they can be used interchangeably. And now we're back on your slides.

Dr. Liam Rose: Okay, great. So the correct answer, at least from an econometric standpoint in theory, is that when they are not correlated with the covariates in the model. And I will show a situation in just a moment where that might be true. Thank you everyone for answering and I will try and answer any questions from anyone who thought that 2 or 3 or were unsure. So this really is the key assumption that is between the two models regardless of any, you know like, math or anything like that. This is really what differentiates them too. And not necessarily a hard and fast rule but this is the big thing that’s going to separate Fixed Effects and Random Effects. Where the first one is going to be some Random Effects and the second one is going to be for Fixed Effects.

So let's talk about a little bit about how we might choose if you are on the fence in any way. The Hausman test is often used to choose between them sometimes you have to put it in your manuscript at least in your plan. So what this gives us is a measure of the difference between the Fixed Effects estimate and the Random Effects estimate. So our null hypothesis is that they’re the same. So whatever you're estimating, say if we want to go back to you our education on income example, say we get an answer of 5, whatever that means, it should be the same in both cases, that's our null. And if you reject the null and then the models are different and you reject the Random Effects model in favor of Fixed Effects.

So it’s a fairly simple model. The idea basically is that, hey I'm just going to run both of them and see if I get the same answer and then test to see if that answer is statistically different. It's nice intuitively because it is simple. And however, it does have a few drawbacks. There's a rejection of the null hypothesis, meaning that you get different answers from each model. It might be because there's not enough statistical power to detect, to partition the null. So say you're doing this on a fairly small dataset and your effect size for whatever you’re estimating is not particularly large, you may not just, the null hypothesis may not mean that they are the same. So that’s, you know, something that's inherent in all statistical tests but it does show, it does rear its head, in this one in particular. And then the other problem is that with Fixed Effects and Random Effects there's a tradeoff between bias reduction and variance reduction. Hausman's does not help in evaluating this. To give an idea of what I'm talking about here, is that the Fixed Effects is going to really reduce the biased, like we said, because it's not making that strong assumption that the factors that don't vary over time are correlated but it does have a little bit more variance the Random Effects doesn't do that so it doesn't really help with this trade off that really might be what you're doing. Clark and Linzer have a reference to it at the end they have a variant paper about Fixed Effects and Random Effects, and they offer three considerations. These are a little bit more rule of thumb than formal but I think they're very useful. So the extent, the first one is the extent to which variation in the explanatory variable primarily within unit as opposed to across units. So if you’re interested in what happens to an individual over time, as opposed to what happens versus individual A versus individual B. If it's within unit you're much more likely to go Fixed Effect, as opposed to approximates, you might want to go Random Effects. In our data, if you have a lot of data almost always better to go with Fixed Effects. And then there's also the goals of the modeling exercise. That could be, as I mentioned before, out-of-sample predictions are actually available in Random Effects because you impose some distribution on the unobserved, oh sorry not the unobserved, the time invariant factors. So you can do the out of sample predictions. So if that's something you're interested in much more towards Random Effects.

So this is just a little bit, kind of, repeating. When variation is primarily within units, decide based upon the purpose of the research. If you have any bias in the Random Effects, could be compensated by increase in efficiency. Practically, what this means, like, if you're trying to use Fixed Effects and you’re not getting a very clear response then you’re getting a null result, with a very wide competence interval and it's not really telling you anything the Random Effects might be the way to go.

And then they have this very nice chart, that's basically, this is saying the same thing. This is kind of like a decision tree, kind of chart. I won't say too much about this but it does kind of give, these are not hard and fast rules but, they are nice rules of thumb if you are deciding against the base kind of work through.

It can also go with which field you were trying to publish in. Econ literature very much not likely to accept Random Effects as much. And then medical literature, a little bit more likely to accept it. Although some of this comes down to you doing randomized clinical trials. And the economists generally work with far more with observational data, where the Fixed Effects is going to be preferred. Whereas in medical journals there is a lot more randomized control trials where the Random Effects might be of use.

So on that note, the mixed models are often used in this scenario. It is possible to include both. And this is often done in a clinical trial where you have a repeated measure that might be correlated. So, an example I gave, is if you have the same memory test repeated over the course of the treatment. For example, if I asked you to take a memory test today, and tomorrow, in the next day, you might want to have a Fixed Effect and a Random Effect. And in this scenario you might have observations that are correlated within unit, but not across the group. So if we’re talking about the memory test, the larger circles here, in this bottom figure, could be the individual. But my memory test results might not have anything to do with your memory tests results. So then we might want to implement a mixed model.

In matrix, just really briefly, this is to know it exists for the most part. In matrix notation, we have the outcome Y, we have X containing the Fixed Effect, Z is containing the Random Effects. In practice use ‘xtmixed’ in Stata or ‘lme4’ in R. As an example, we can use the effect of a treatment on a test score with a Fixed Effects for the week, that's a test to take in; and the Random Effects for the site, if perhaps we think that the site is administering the tests a little bit differently. So the data code is on the bottom here ‘xtmixed’, ‘testscore’, ‘treat’, ‘week’, with a Random Effects for site\_id.

So my last thing I'll say, is that some sometimes it can come down to terminology. Gelman, in 2005, gives a nice overview of different ways they’re defined, in some cases the terms are applied to different concepts. So it can be pretty confusing if you’re someone that works in the interdisciplinary context to figure out what the researchers are talking about. So, one thing in particular, is noted here is that some context a Fixed Effect is like a population average effect. And Random Effects means that subject is a specific effect. Sometimes those could both mean a Fixed Effect in a typical sense. So, yeah, it can be a little difficult. The best advice I can give is, be specific and ask questions, to make sure everyone is on the same page about how the model should be formulated.

And then I will point out a few references. Of note on this here is the Clark and Linzer paper, ‘Should I use Fixed or Random Effects?’ So we used a lot of that in forming this. And then the Gelman paper that talks a lot about it, in terms that was kind of more of discussion state of the literature, kind of thing. So now I would like to open up to questions, I think I left a good amount of time so please, please, please, give as many questions then we can double back to anything and talk about anything you guys would like. Thank you for listening.

Dr. Todd Wagner: Thanks, Liam. Well right now we don't have any questions coming in, so I’ll ask one and then as people type it in. So if we go back to when you first gave your first example, which was income and education, could you walk us through just sort of what you were thinking in, most of the time we think of that as a Fixed Effect, you’d have a person Fixed Effect and that you sort of want to control, so you said, the motivation of the person and then you're essentially looking at a small deviations within the person that change over time, correct?

Dr. Liam Rose: Right. So, in that scenario you’re trying to say like, the basics of it are if someone went to college versus someone went to high school, there are many, many, many, reasons why one person is able to go to college and one person only completes high school. And some of them will be related to income and some will not. And some will change over time and some will not. And so one reason might be that this person, this particular individual has, if they went to college has maybe has more inherent motivation. So the Fixed Effect help us get rid of that, if we think that, that’s something that’s not changing over time.

Dr. Todd Wagner: Great, now in some of those cases we still have to make an assumption of exogeneity. So one, it might not obviate the need for instrumental variable, so there are ways to do Fixed Effects with instruments. I wasn't sure, I know you're not going to talk about that now but, do you want to say anything about the exogeneity assumptions built into Fixed Effects?

Dr. Liam Rose: Yeah, right. So I mean it comes with down to how you think those within person effects that change over time might be correlated with both the outcome and your variable of interest. One example you might give is, if you are somehow unable to observe the person's demographic status, and that’s something that may change over time. That could be correlated with both the person's income and the person’s educational achievements. And that would give us a confounder that’s not going to be fixed by a, including a Fixed Effect.

Dr. Todd Wagner: Yeah, we did a study years ago, because you know VA it's got a 104 medical centers and these centers adopt things at different times. So at one point we use a Fixed Effects model to say “Hey, what about these certain medical centers that adopted this news sort of treatment model for mental health compared to those who didn't?” And we used a Fixed Effect model but it assumes in some sense that those who did were as good as randomly assigned and in that case you just might not like that assumption.

Dr. Liam Rose: Yeah, that’s very true\_\_

Dr. Todd Wagner: You don't get around it\_\_

Dr. Liam Rose: Very true.

Dr. Todd Wagner: You don’t get around it with that. So in the ideal scenario you would have an instrumental variable that says here is something that nudged the certain facilities to adopt and will leverage that nudge, if you will.

Dr. Liam Rose: Right.

Dr. Todd Wagner: Another question for you is, thanks for the very informative presentation. If you have two levels of geographic areas, let's just say state and county, which is more appropriate to choose for a Fixed Effect?

Dr. Liam Rose: Wow, what a great question. With that I’m going to give a very bad answer, which is, it depends. So if you have, it really depends on what level the policy or whatever thing you might be looking at, hits. So if you think that, and if you want to look at results that are within county, or with an individual, within an individual and within county, for statewide policy that might inform you a little bit more about how it's going differently across counties. And then vice versa, if you want to completely ignore counties and you want to go across states, than you might want to do it that way. I further will say that, as long as you are making the assumption that the individuals are not moving over time of your study, it’s actually not a problem to do county and state Fixed Effects.

Dr. Todd Wagner: So mostly it's just a data issue, right? Do you have an update\_\_

Dr. Liam Rose: Data in the Fixed Effects\_\_

Dr. Todd Wagner: County and state. Yeah

Dr. Liam Rose: Yeah.

Dr. Todd Wagner: So and that might prohibit one or the other. Great answer I agree with you. Can you please give an example, of a Fixed Effect in the population? Are we talking about a stratified sample in this case?

Dr. Liam Rose: That's a good question. I'm assuming by population you mean like the large population. Is there such a thing as a Fixed Effect in the population? Geneticists might say yes, but the idea here is that, wow.

Dr. Todd Wagner: And you might qualify it, Liam, within relatively short panels of time, there are things that are fixed. So you could think of yourself and how much you've changed over your life span, but over the past few years there are things that haven't changed in you. And so it might be a fairer assumption over a very shorter period of time. Yea\_\_

Dr. Liam Rose: Right.

Dr. Todd Wagner: And I love your example of genetics. Cause I think that geneticists would often argue, except for mutations, that largely the sort of set when you’re created.

Dr. Liam Rose: Right. So the way I think I would answer that is that it depends on the individual. So there's very much a sense of a population level Fixed Effect. If, for example, your unit is not an individual but perhaps a firm or a country which is something that’s done a lot. So then that becomes a very different calculus about whether you think that there is time invariant factors across maybe a country. So if you are doing some analysis across countries and you include this country Fixed Effects, you’re doing within country over time. And in that sense there often is a very different, you know, if you look at 10 years the variation between countries just overwhelms the variation of countries over time. So in that scenario, yeah, countries Fixed Effects, or something like that, would very much be appropriate and that one I can kind of think of as a population level Fixed Effect.

Dr. Todd Wagner: Okay, people are still writing in questions, this is great. Can you elaborate on which type of effect is useful for calculating overall population estimates versus individual estimates?

Dr. Liam Rose: So, I will say, if the concern here is about out-of-sample predictions, then Random Effects is the way to go. If you are more concerned about something like a, treatment effects for the group that takes the treatment or the treatment on the treated kind of thing. Let’s see population level estimates. Sorry, but having a little trouble considering what population level means. So the fundamental part of this is, is that it has to be panel data, right? This has to be repeated observations of the same unit over time. And so fundamentally when you were trying to do that it's by unit. So you want to know the effects for an individual. If you mean by population you mean by for all individuals, that’s a little bit difficult. I would still say that Fixed Effects is the way to go there. So, because, we’re trying to make in oftentimes a Fixed Effect are used for countrywide, or nationwide, surveys or studies, a few things like that. Where you can wipe out the variation within individual, if that’s your unit.

Dr. Todd Wagner: Great.

Dr. Liam Rose: Does that make sense?

Dr. Todd Wagner: Yeah, and we’ll, I think if the person has additional follow ups, they can email us, if we aren’t answering it. But we have a couple more questions I want to get to too. So, we don't talk a lot about nesting, do you have anything you wanted to add here or would that be a separate topic. So you said, counties and states, there is an implicitly sometimes correlation between counties and states and we didn't really talk about nesting. So the person is asking about: How do you account for the nesting?

Dr. Liam Rose: So that's actually a great question. It's okay to do that. So a lot of times you can see some papers, where instead of Y sub IT it can be Y sub IGST, where a person I, in group G, in state S, at time T. The nesting is okay. You might remember I mentioned, we don’t really get an effect for the Fixed Effect itself. We’re not recovering an estimate for that. We’re really going for an estimate of our variables of interest. One example I can think of is if you are, for example there’s a program in Mexico where they were giving unconditional cash transfers and they were doing this by village. So they want to know the effect for person I, in family F, in village V, in state S, in time T. So then you can have a Fixed Effect for the family, for the village, for the state, and for the year. And those are all okay, those are taking out all the within family variation it takes out all the within state variation all those things that doesn't change over time. The things that do change over time, those are still there but the things that don’t change over time within any of those units can be put into the model and you can just wash out that variation in that Within transformation that I showed earlier. So it can be quite nested, and I've seen them maybe five different levels but again you’re not getting an estimate for the Fixed Effect itself you’re really focused on that variable of interest.

Dr. Todd Wagner: Yeah, so we have, we’re not going to have enough time to answer all of them. So thank you for your presentation. My initial understanding is of Fixed Effects estimators was that this is an estimator that adjusts for patient characteristics or you could say facility characteristics that don't vary over time. Is this the case that this model is in essence removing the effect of that unit? That is, yeah that’s right.

Dr. Liam Rose: Over time, right. So I agree very much with that characterization. So it doesn't remove the part of that individual that changes over time. It's just the part, remember when we broke up that error term in the time varying and the non-time varying so it's just getting rid of the non-time varying. So we're still getting that within person variation or within facility variation we're just not getting the part of the variation that is so different, maybe across units but doesn’t change within the institution over time or individual.

Dr. Todd Wagner: That’s right, and if you took, for example if you had a patient level Fixed Effect and you include that Fixed Effect in the model you wouldn't be able to have any estimate as you mentioned, Liam, on things like race that don’t vary over time because that Fixed Effect already accounts for that.

Dr. Liam Rose: Yeah, completely takes that out.

Dr. Todd Wagner: Yep. So if, last question, I hate to ask, put up the last question and then we’ll have to follow up on things. So a Fixed Effect has two parts, time invariant values and time varying, time variant values and time invariant effects. The first seems easy to assess but what do you make of the time invariant effects that you estimate?

Dr. Liam Rose: So backing up that Fixed Effect itself is that part that is time in-varying that's the fixed part of it. So the time varying stuff we don’t, it's still part of that error term. Remember we broke those up into two. In general what you do is you're going to say it's okay to have both, basically. It's not a penalty it's not gonna fix everything, but like I said before there's both aspects of that. So take for example if you had a medical center Fixed Effect. There's differences across medical centers that are quite large. And they don't change over time. However if we include the Fixed Effect for the medical center however, if we include that we still have the within center variation. So, you know a center is a very large unit for VA patients or for I guess any other hospital, there's going to be very different patient experiences within center and you still have that variation in your model that's going to be a part of the error term and part of the explanatory variable. So, regarding like what part is taken out and what part isn’t taken out it’s not really up for us to decide because some of it is going to be unobserved and some of it’s not. So there’s a portion of the error term that's going to be in that time in-varying part and some portion that's not. But if you include this medical center Fixed Effect you're taking out all the time in-varying part but leaving the time varying part and that's all you can do about it.

Dr. Todd Wagner: Great, I think we’re over the hour by two minutes so we'll have to leave it there but that was great thanks, Liam.

Dr. Liam Rose: Okay, I will mention then that my email is liam.rose@va.gov. So I welcome anyone who didn't get their questions asked or if you had follow-ups or anything like that please email me. I’ll be sitting here answering them or if you have it at a later date go ahead and email me again at liam.rose@va.gov.

[ END OF AUDIO ]