Cyber Seminar Transcript

Date: March 22, 2017

Series: HERC Econometrics with Observational Data

Session: Fixed Effects and Random Effects

Presenter: Josephine Jacobs, 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%3A%5Cl)

Molly: I would like to introduce our speaker. We have Dr. Joe Jacobs joining us, she’s a health economist for the Health Economics Resource Center, known as HERC, and she’s located in VA Palo Alto Health Care system.

Dr. Joe Jacobs: Thanks very µch Molly and thanks Wade as well for all of being backup here. And thanks everyone for joining today’s seminar on fixed and random effects. I’m going to start off by noting that these two terms sometimes generate confusion, because they can be used really differently across different disciplines. And today we’re approaching the topic from an econometrics standpoint. But at the end we’re going to note how the terminology can be applied pretty differently depending on where you look and what your background is.

So I’m going to start off with just a brief poll asking everyone how familiar they are with fixed effects and random effects models: very familiar, somewhat familiar or not familiar at all?

Molly: Thank you. So for our attendees, as you can see up on your screen, you do have a poll question. So please just go ahead and click the response right there next to your answer. And it looks like the responses are coming in quite quickly. We’ve already had an 80% response rate, so that’s great. We’ll wait just another second and close it out. Okay. It looks like I see a pretty strong trend.

I’m going to share those results: 15% feel they are very familiar, 66% somewhat familiar and 19% not familiar at all. So thank you to those respondents and I’ll turn it back to you.

Dr. Joe Jacobs: Okay. Great. Thanks Molly. Okay, so we have a bit of a range here. And hopefully for those who are newer to the topic this will be an introduction to some new tools that you can use. And for those more experienced hopefully it will provide maybe some new resources or it will approach the topic maybe from a different perspective than what you are used to.

So since we’re approaching the topic from an econometrics standpoint, we’re going to start today’s seminar off with a brief overview of panel data, which is one type of data that fixed effects and random effects models can be used with. We’ll also go over the panel linear regression model and use it to understand a type of omitted variable bias, unobserved heterogeneity, that can be addressed using fixed effects and random effects models.

We’ll go over the fixed effects and random effects models and talk about how we might choose between the two models. And finally, we’ll conclude with some notes about differences in the use of terminology across disciplines.

So let’s start with a brief overview of panel data, which is one type of data we use with fixed and random effects models. Panel data is the pooling of observations for the same cross-section of individuals or households, countries, whatever the unit of analysis is, over several time periods. In this case, we have data at the person level and we observed each person over three time periods: 2010, 2011 and 2012.

Some of the factors you’ll note, like age and income, are changing over time, while others, like sex and education in this case, are not. I’m also going to note here, and we’ll touch on this again later, that the models we’re going to explore in the remainder of this seminar can also be applied to clustered data. Where instead of following a unit of analysis over time, we have observations that are clustered into groups.

Now last lecture with Christine, we recapped the linear regression model. And you may recall the model above helps us predict a change in Y for a unit change in X. This time, because we’re dealing with a time dimension to our data, you’ll note that we have an additional subscript, *t*. So now we have Y*it* , which is the outcome for individual *i* at time *t ,* X*it* , which is the explanatory variable of interest for individual *i* at time *t*, and epsilon *it,* which is the error term for individual *i* at time *t*.

Now you may recall that ε*it*  contains all other factors, besides X, that determine the value of Y. And, as Christine noted previously, in order for β hat to be an unbiased estimate of the casual effect of X on Y, X has to be exogenous. And we’ll recap what that means briefly. This means essentially that X*it* and epsilon *it*cannot be correlated with one another. So the explanatory variables in your model can’t be correlated with the error term.

Christine noted last lecture that there were instances where the error term will be correlated with the covariates in the model, and one such case relates to omitted variable bias. So if you’re unable to include explanatory variables in your model and these omitted variables are correlated with an explanatory variable in your model, then you’ll have biased predicted values of β.

So say, for instance, you want to look at the effect of education on income, as in our sample panel from the previous slides, we can control for age, sex and education. But we don’t have data on innate ability or ambition. And since innate ability or ambition may also be correlated with education, omitting these factors will lead to bias in our model.

And today we’re going to talk about a special form of omitted variable bias called “unobserved heterogeneity” and how some panel data techniques can be used to address this kind of bias in a regression.

“Unobserved heterogeneity” refers to omitted factors that may vary across individuals, but that remain constant over time. For example, at the individual level this may include some demographics like race, ethnicity, family history or maybe innate abilities if they don’t change over time. At the state level this may geography, or over a shorter term, say demographic, educational or the religious composition of a state, though surely those will change over the longer term.

So working with the equation we previously outlined for a panel linear regression model, we might think of this unobserved heterogeneity by breaking apart our error term, epsilon *it*into two components. First, a component *uit*that varies over time and by individual. But also a second component that is unique to each individual but does not vary over time, α *i..* Alpha *i*is what we refer to as time-invariant unobserved heterogeneity.

Now if α is not related to other explanatory variables in the model, then αis like any other observed factor that’s not systematically related to Y and is soaked up in the error term. And we really don’t need to worry about it. If, however, α is related to other covariates in the model, then relegating αto the error term will be problematic. And it will result in biased estimates of our coefficient.

So referring to our income estimation example, if we believe innate ability is not related to education and other covariates, we can ignore it and just run a pooled OLS. But if we believe it is, this will be a problem.

Now Clark & Linzer, which is a reference I have at the end, demonstrate what we mean by “creating bias” when we don’t account for unobserved heterogeneity quite nicely with some siµlation models. In these siµlation models they impose different degrees of correlation between the individual time-invariant unit effects, so α,and independent variables, X.

In the center is case B. We see a case where there’s no correlation between αand the independent variable, X. And we see that an unbiased estimate of β would be equal to 1. In case A we have a situation where the independent variable X is negatively correlated with the unit effect, α*.* And the result is an estimated β that’s smaller than the true β value.

In case C we have a situation where the independent variable is positively correlated with unit effects, α*.* And the result is an estimated β that’s larger than the true β. So in both A and C we can see how the estimated coefficients would not reflect the true value of β. So we have biased coefficient estimates.

Now previously Christine talked about trying to deal with this type of omitted variable bias through the use of instrumental variables. But as Christine noted previously, it’s often very difficult to identify and appropriate instrumental set variable. So panel data offers methods to deal with this special type of omitted variable bias, time-invariant unobserved heterogeneity.

So now that we’ve seen how this form of omitted variable bias can impact estimates of our regression coefficient, we’ll look at how panel data can be used to try and deal with unobserved heterogeneity, or α, from our previous slides.

There’s two standard approaches to modeling variation α *i*. There’s the fixed effects model and the random effects model. So the first approach we’ll talk about is the fixed effects model. And with the fixed effects model, basically we replace the unobserved error component, α *i*, with a set of six parameters, let’s call them µ1 to µn, one for each of the n units in the sample. So µ1 would represent the net effect of unobservable factors on Y that are constant over time for one unit, or for unit number 1.

These n fixed parameters control for the net effects of all unobservable factors that differ across units but are constant over time. And this is the general idea behind fixed effects. Because fixed effects models control for all time-invariant differences between individuals or units, estimated coefficients won’t be biased because of these omitted time -invariant characteristics. Fixed effects models, then, are designed to study causes of change within a person or unit. You can think of this as using each unit as a control for itself.

So referring to our running example about estimating the effect of education on income from earlier, if innate ability was constant over the time period in question, then µ1 would control for that plus all other factors that don’t change over time for individual 1.

There are a couple ways of operationalizing fixed effects models. And perhaps the most intuitive is through the Least Squares Dummy Variable Estimator. In this case we simply create a dummy variable for each unit. For instance, if our unit of observation is an individual person, we create a dummy variable for each person in our data set. And we then simply regress our dependent variable Y on the dummy variables and all of the other explanatory variable in the model. And this, again, gives you an idea of the intuition behind fixed effects, where we have a control for each individual unit that soaks up the effect of all factors unique to that individual that don’t vary over time.

And we can interpret the coefficients of these individual or unit specific dummy variables as different intercepts for each individual or unit. And for this reason we’ll often leave out the β-not term, or the constant term. Typically though if we have many individuals and not a lot of time periods, adding a dummy variable for each individual eats up a lot of degrees of freedom. And it isn’t very practical.

In place of that, often the fixed effects estimator is used. And this is another way of sort of removing time-invariant factors from our model. This works by first determining the time-mean of each component, which we will call Y bar, X bar, U bar and αbar.

So the time-mean is simply the average value of each component for the *i*th unit over T years. So for the panel data we showed earlier, this would be like taking income for person 1 over the three years we observed, say 2010, 2011 and 2012, and taking the average income over those three years.

Now you’ll note that in the case of α *i*, our time-invariant individual specific characteristic, the timing is equal to the constant α *i* value. because *a*i is constant, and therefore the same in every time period.

So we then complete what we call a “within” transformation where we create a time-demeaned data set. This means that we subtract the time-means so that what we calculated on the previous slide from the actual values of Yit, Xit, Uit and αit and then run a regression on these time-demeaned values. So we’re running the OLS regression using the transformed data, which is now in the form of deviations from the time-means.

Now the important thing to note here is that the regression will not include any constant terms because the time-means of these factors will be the same as the actual values. So because α *i* is constant and therefore equal to α *i* bar, this nets out to 0. So the least squares dummy variable estimator and the fixed effects estimator will give you the same coefficient estimate. Though if you were doing the fixed effects estimator by hand, you’d have to be careful to adjust the standard errors appropriately. Most statistical packages will do this for you.

You’ll also note that unlike the least squares dummy variable estimator, there has to be some additional math to obtain estimates of the fixed effects parameters. If you had an interest in understanding the unit level effects and their effect on the outcome variable. So we’ll note here that if we were to try to run fixed effects with Stata then, for instance, we could add individual level dummies to our regression model, like the least squares dummy variable estimator. Or we could use the function xtreg, fe, which uses the time-demeaned approach. And they would both give you the same parameter estimates.

So one of the biggest assets of the fixed effects estimator is that it generally can be used to produce unbiased coefficient estimates when the time-invariant omitted factors are correlated with your explanatory variable. And this is a big pro because the reality is that this is likely to be the case the vast majority of the time. And this is also why you’ll see fixed effects favored in the economics literature, which is often focused on modeling causal effects. But those estimates can be subject to high sample-to-sample variability when there are few observations per unit and when there’s limited variability within each unit relative to the variation in your outcome variable Y.

So the latter implies that when most of the variation in the outcome variable comes from variability between subjects, as opposed to within subjects, simulations often show that fixed effects estimators don’t perform well. And the standard errors can be very high in these cases.

So an example of this is if you were examining the effects of, say a state’s median income on crime rates, you may find that while income varies significantly across states, it changes very slowly over time within states. In such a case, standard errors in fixed effects estimators may be very large. This can be contrasted, for instance, with average opinions across states, which may vary little by state but have substantial variation within a state.

So another con of the fixed effects estimator is that if you have an interest in understanding the effect of time-invariant explanatory variables or variables that change very little over time, it will not be possible to determine their effect, as these will be netted out when you time-demean your data.

And finally, fixed effects is optimal for estimating effects for individuals in the data set being used, as the unit effects of unobserved units are unknown. This isn’t necessarily the case with random effects, as we’ll see in C.

So to demonstrate the fixed effects model in action using an example from the recent literature, we’ll briefly look at Oberg et al. 2016, which is a study where the authors assessed the association between labor induction and Autism Spectrum Disorder. In the study we have Swedish register data from 1992-2005 and the authors look at 1,300,000 births and conduct a within-siblings comparison.

The authors are able to control for pretty impressive array of observable factors including birth year, parity and a number of maternal characteristics. What’s missing here though are key unobserved factors, such as environmental factors at the household level and genetic factors shared within families. [inaudible 21:20-21:35] At the maternal sibling level the underlying hazard to [bear? 21:38] between mothers. So the comparison is within siblings only.

And we can see the impact this has on the outcomes. So in the first three models, which include increasing numbers of covariates, we see a positive and significant association between labor induction and autism. We see that this association is attenuated the more covariates we add. So decreasing from a hazard of 1.32 to a hazard of 1.19 from the most parsimonious model, model 1 to the sort of kitchen sink model, model 3, which includes many more covariates. But the effect is still significant and positive.

However, once we include the maternal siblings fixed effects, labor induction is no longer associated with offspring Autism Spectrum Disorder. And this is an indication that unobserved time-invariant factors shared by maternal siblings, like maternal genetic factors, may have been unaccounted for in models 1-3 and may have been biasing the coefficients in these models.

So we’ve seen that fixed effects will tend to produce unbiased estimates in situations where the time-invariant omitted factors are correlated with regressors in our model. But what if this isn’t the case? What if we have reason to believe that the omitted factors are not related to any other independent variable in our model, or that this correlation is very low? Would we still have to run a fixed effects model, which may be soaking up time-invariant terms we want to look at, or which may inflate our error terms?

In the economics literature this is where the random effects estimator comes into play. If you can assume that there’s no correlation between X and your unobserved heterogeneity α, you do not need to use a fixed effects estimator. But you also can’t simply run a OLS estimator either. Because you have repeated observations for each unit, it’s likely that you’re going to have serial correlation. That is the error term at one point in time will be correlated with the error term at another point in time.

And a quick aside on serial correlation, what this means is that what the OLS estimator will not necessarily be biased but it will be inefficient, and the standard errors can be underestimated. Basically, if we fitted a regression line, not taking into account the fact that we had repeated observations of the same individual, the standard errors would be smaller than those with the true regression line. We would be underestimating our variance here. And this can have important implications for hypothesis testing, like leading us to conclude a covariant is significant when in reality it isn’t.

So instead you could use a technique that’s more efficient than fixed effects and doesn’t eat up so many degrees of freedom, but unlike pooled OLS, still accounts for the observed heterogeneity in our model. With fixed effects, we simply got rid of the α *i* term through a within transformation. And that is by time-demeaning our data. While pooled OLS simply ignores the α i term and lets it get subsumed by the error term, random effects is a compromise between these two models. And it maximizes efficiency by calculating β as a weighted average of the fixed effects estimator and the pooled OLS estimator.

So there’s a bit of Greek here. But I just want to show the two components that come into play with the random effects estimator and how we transform the fixed effects system to make it more efficient. Random effects essentially transforms the fixed effects system with an inverse variance weight, λ, where λ is one minus the square root of the variance of u*it*, which you’ll recall is our idiosyncratic error over the variance of u*it* plus T times the variance of α *i*.

So we use λ to, what they call quasi-time demean the data system, the system. Basically we take off a fraction of the time-demeaned value instead of the full time-demeaned value, as we would with fixed effects.

So let’s get rid of the equations and just talk a bit about the intuition behind random effects. λ is typically between 0 and 1. When λ is equal to 0, the system is equal to a pooled OLS regression. And this makes sense, since in order for λ to be 0, sigma α, our variation in our unobserved heterogeneity term, would have to be 0. And this means that variation in α does not comprise a significant portion of the error term and could be ignored.

Meanwhile, if variance in α is very large, then λ will be equal to 1. And the random effects estimator is essentially equal to the fixed effects estimator. And this again makes sense, since α would then comprise a significant portion of the error term. And it can’t simply be ignored. The random effects estimator would try to remove as much of this effect as possible, much like the fixed effect estimator does.

So essentially what happens is that groups with outlying unit effects will have their α *i* shrunk back towards the mean α, which brings the estimated β closer to the pooled OLS estimate and further from the fixed effects estimate. And the effect will be greatest for units containing fewer observations and when estimates of the variant α *i* are close to 0.

So to get an idea of what’s going on when we use the random effects estimator in a statistical package like Stata, let’s break down what’s happening. The random effects estimator is operationalized in two stages. First, we would obtain an estimate of λ or λ hat by determining the variance of u and α. And these estimates are usually obtained by estimating a fixed effects or OLS regression.

We then substitute λ hat for λ into our transformed system which is what we refer to as quasi-demeaning our data and then just run an OLS on the transformed data. And in Stata we can do this using the xtreg, re command.

Now there are a lot important advantages to random effects models. First, as we noted, estimates of β will have less variance as outlying unit effects will have their α shrunk back towards the mean α. And this brings β closer to the pooled OLS estimate and leads to estimates that are closer on average to the true value in any particular sample, not just the sample in question. So a lot of researchers may favor random effects because the inferences from random effects estimates are more generalizable beyond the sample in a given analysis.

Another major advantage is that we can include time-invariance covariates in the model. With fixed effects these are washed away by the estimator, which is a disadvantage if we want to know the effect of that variable on our outcome.

A third advantage is that random effects can be used when we have small samples within units, which for a fixed effects model may result in large variation or even an inability to estimate the model.

So what are the cons? The main reason economists and others interested in causal inference do not favor random effects is because it almost inevitably, you’re going to break the assumption that there is no correlation between α *i* and other covariates. And this will introduce bias in estimating β. And the greater the correlation between covariates in the model and α *i*, the greater the bias in estimates of β.

Finally, a con may be that we don’t actually estimate α *i* in random effects for each individual, which may be something that’s of interest to researchers.

Now to recap briefly, I’d like to ask a quick question from an econometrics standpoint. I’d like to ask, when is it appropriate to use random effects instead of fixed effects? Is it when the unobserved unit-specific factors, α *i*, are not correlated with covariates in the model; when the unobserved unit-specific factors, α *i*, are correlated with the covariates in the model; or the models can be used interchangeably?

Molly: Thank you Dr. Jacobs. So for our attendees, you can now see the poll up on your screen. So please select answer option 1, 2 or 3. And do you verbally want to go over those again real quick, Joe?

Dr. Jacobs: Sure, yeah. I know it was a lot of text, so we had to sort of clip this down to three. So we’re asking: is it when the unobserved unit-specific factors, α *i*, are not correlated with the covariates in the model? That’s answer one. When the unobserved unit-specific factors, α *i*, are correlated with the covariates in the model? That’s two. Or when basically the models can be used interchangeably? That’s three.

Molly: Thank you. So we’ll give people a few more seconds. Responses are still coming in. I’m going to give everybody a chance to think through this real quick. Okay. It looks like we have a pretty good response rate. So I’m going to go ahead and close this out and share those results. So it looks like 40% selected answer 1, 46% answered option 2, and 10% say they are interchangeable. Thank you to those respondents. And I’ll turn it back to Dr. Jacobs.

Dr. Jacobs: Okay. So the question was when we could use random effects. And from an econometrics standpoint, it would be the first answer: when the unobserved, unit-specific factors α *i* are not correlated with the covariates in the model. And this is sort of one of the major takeaways from an econometrics standpoint that sets these two models apart.

And it’s a key assumption in economics that distinguishes the use of the two models. We will see in a bit though that this is not necessarily a hard and fast rule in applied research. So one thing I’d like to touch on before discussing choosing between the two models is clustered data. We’ve considered fixed effects and random effects in the context of panel data up until now, which is observing the same units over time.

Another way to apply fixed and random effects is when we have clustered data, when observations are clustered into groups. For instance, health facilities in a geographic region, patients in the hospital, individuals in a family or individuals with a certain health status.

And the intuition here is somewhat similar to panel data. If we believe unobserved common group level characteristics can affect our outcomes and we don’t account for these factors, we can have bias estimates due to unobserved heterogeneity at the group level. So in place of our *i* individual or unit term nt time subscripts, we would instead have subscripts in our equations for the group and observation number within the group. So Dieleman & Templin 2014, which I provide a reference to at the end, they give a nice overview of how the same intuition we’ve talked about so far, or similar intuition can be applied in cluster data.

So, choosing between fixed effects and random effects. To determine which model to use, something called the Hausman test is often used. And this is a measure of the difference between fixed effects estimates of the β coefficient and random effects estimates of the β coefficient.

The null hypothesis here is that the coefficients estimated by the random effects estimator are the same as the ones estimated by the fixed effects estimator. We’re basically testing whether the covariance between our unobserved heterogeneity term, α *i*, and the covariates, X, is equal to 0.

If this were the case, then we could use the random effects estimator. A rejection of the null hypothesis then indicates that the two models are different, and we should reject random effects in favor of fixed effects.

So as widely used as the Hausman test is, there are some drawbacks to it. And again, Clark & Linzer, which I provide a reference to at the end, use simulations to demonstrate a number of scenarios where the Hausman test can lead to erroneous conclusions. They note that a rejection of the null hypothesis may not just be due covariance between α *i* and covariates in the model. But it may be because the test doesn’t have statistical power to detect departures from the null. Because of this potential they note that careful consideration should go into the choice of the estimator. And there’s always a trade off between bias reduction and variance reduction. And the Hausman test doesn’t really take this into account and help you evaluate this trade-off.

So instead the authors suggest three considerations when choosing between fixed effects and random effects. First you have to consider the extent to which variation in the explanatory variable is primarily within unit, as opposed to across units. So they warn about a sluggish case where independent variables change very gradually over time, relative to changes in the dependent variable. In cases where this happens, correlation between the unit fixed effects and the sluggish variables can destabilize estimates.

The second factor to consider is the amount of data that you have, the number of units and the observations per unit. When a data set contains many units or is organized according to a complex data structure where observations are grouped into more than one unit or at more than two levels, random effects models can be less complicated to specify and interpret.

And third, you have to consider the goal of the modeling exercise. With fixed effects, for instance, out of sample predictions wouldn’t be possible because the unit effects for unobserved units aren’t known. With random effects we estimate the distribution of the unit effect, including the mean effect in the broader underlying population. That’s the idea. So even if the observed units are fixed, this could be a reason to choose random effects over fixed effects.

In that article the authors also provide a nice rubric for how to decide between the models. And this can serve as a guide to researchers. They base this advise on evidence from their simulation models. So when there is variation in X that’s primarily within units, they find there’s rarely a difference between fixed effects and random effects estimators.

You’ll recall an example of predominately within unit variation was average opinion across states. For the average opinion across states may be quite similar, but the individual variation of opinion within states is likely very high. In these cases, only when there’s little data and correlation between the regressors and unit effects is exceptionally high, did the authors find that fixed effects out-performed random effects.

In this specific case they find that any bias in the slope parameter estimate is compensated for by the increase in efficiency. So ultimately they advise in this case that researchers should use whichever model better serves the purpose of their research. If you wanted to make predictions about unobserved units that are not in the data set, or if you’re concerned about collinearity between the regressor of interest and the unit effects, then random effects estimators might be used in place of fixed effects estimators.

It’s not so straightforward when there’s variation primarily across units. And you’ll recall an example of this is that when we think about change in median states income within states over time. Over time median income will change slowly. So most of the variation in the independent variable may be explained by differences in the wealth across states instead of within states. In these situations the choice of estimator depends on a number of different factors. It depends on the amount of data and the underlying correlation between the unit effects and the regressors.

So in this scenario, random effects is only preferred when there are few observations per unit and few units. When there are either few observations per unit or few units, random effects is only preferred if the correlation between α and your covariates is low. Otherwise fixed effects is preferred.

Finally, if there are many observations per unit and many units, fixes effects is preferred unless the correlation between α and your covariates is close to zero. So what does this tell us? It tells us we have to think very carefully about specifying our models. We have to think about what’s included in the model and very importantly, what’s excluded from the models and from our covariates, and how both of these factors are related to each other when we’re choosing whether to use fixed effects or random effects approaches.

So we’ve mapped out some detailed advice on choosing between fixed effects and random effects. But these criteria aren’t necessarily commonly applied. A lot of the time choice of model comes down to the method most favored by the discipline. In a review of health, economic and political science literature Dieleman & Templin showed that while health sciences tended to heavily favor random effects, economics and political science researchers tended to most often employ fixed effects.

Some of this comes down to the fact that health sciences researchers may be working in the framework of an RCT in which unobserved are truly uncorrelated with the variable that is indicating treatment. Economists meanwhile mostly work with observational data and are often more interested in the causal inference as opposed to prediction.

And some of it comes down to terminology. We’re going to conclude by noting that some of these differences may also come down to terminology used across different disciplines. Gelman, which I’ve referenced on the next slide, provides a nice overview of different ways that fixed effects and random effects are defined. In some cases the terms are applied to different concepts. So it can be very confusing when you’re in an interdisciplinary context to figure out what the researcher is referring to.

One thing in particular is that in some contexts fixed effects is a population averaged effect and random effects is the subject-specific effect. The best advice I can give here is asking people to clarify what they mean when they use these terms as it’s very easy to talk past someone, assuming you’re both referring to the same type of model.

One reference that I don’t include on the next slide but that goes over some differences in terminology between fixed effects and random effects from a economic standpoint versus from the, say, hierarchical linear modeling literature, is called Chaplin. And if anyone’s interested in that, you can contact me. And then they do a very good job of breaking down the differences between fixed and random effects from an economics approach. So what we discussed here versus in other modeling approaches.

And I just want to point out some references that were used to put this together, and that could be useful resources. For those who want more detail about how to think about what we’ve discussed in the context of clustered data instead of panel data, Dieleman

provides a very accessible overview of this. It is similar to the Clark & Linzer paper and runs a number of simulations models to demonstrate when random effects versus fixed effects are most appropriate for clustered data. And the Chaplin reference I mentioned previously also talks about the terminology of fixed effects and random effects in other disciplines.

So that concludes the lecture part of the portion. And if there are any questions?

Molly: Thank you. We do have several pending questions. I’ve just given Wade access to those. So Wade just let me know when you can see the questions, and you can just jump right in.

Wade: Okay. For some reason I couldn’t see the questions. Let me see, the questions…

Molly: Directly below the attendee section of your control panel you should see the word “questions”. You can click that.

Wade: Yeah. The question was hidden. I think the first question is asking you to repeat the correct answer to, the last question, what is the relationship between – I couldn’t read it. Molly, can you read the question? Maybe you can help me here.

Molly: Sure. Yeah, no problem. So the first question that came in: is serial correlation similar to auto-correlation?

Dr. Brown: Yes. Yeah, similar. The terminology is different but the idea is similar. So there’ll be correlation in the error terms over time. The error term in one time period will be correlated with the previous. So, yeah.

Molly: Thank you. Can you give us an example or examples when we assume that α is not correlated with the error term?

Dr. Brown: Yeah. And that’s a great question. And that’s why you’ll see in literature from economics and political science where causal inferences opposed to prediction is more important, random effects won’t be used as often. Because it’s actually very hard to come up, if not impossible, to come up with an example where it would not be related to the covariates in your model. And I definitely welcome anyone to try and think of one and send it in. And I can use it for the next lecture. But I actually struggled a lot to come up with an example where that’s the case. Because it’s very rarely the case, and that’s why random effects aren’t used as frequently in studies where you’re trying to get a causal inference.

Molly: Thank you. Wade if you want to jump in and answer any of these, just let me know if you have something to add before I move on to the next one.

Wade: Yeah, you can go ahead and move on to the next question.

Molly: How can we interpret the coefficient of interest when we use fixed effects versus random effects?

Dr. Brown: So that’s a good question. It’s if you’re assuming – which I previously said is a hard assumption to make – that there’s no correlation between your α *i* and your covariate of interest, then both should be quiet similar, in that if there’s no other omitted variable, no other sources of endogeneity, it could be the effect of X on Y. But if you’re using random effects for other reasons, such as to get around collinearity or not about causal inference, I’d say the random effects coefficient is further from the true or unbiased coefficient than the fixed effects some of the time.

And we talked about a lot of exceptions to this. And in both cases it’s difficult to say that it means it’s a causal effect as well, I’ll note, because there are potentially time-varying omitted factors that we’re not accounting for.

Molly: Thank you. The next person writes, they would like for you to please repeat the answer to the last poll question that was asked.

Dr. Brown: Sure. And that’s an important one from the econometrics standpoint because it’s the distinguishing, sort of theoretical reason, behind why we choose fixed effects versus random effects. So the question again was: from an econometrics standpoint, when it is appropriate to use random effects in place of fixed effects? And the correct answer was when the unobserved unit-specific factors, so α *i*, our time-invariant omitted factors, are not correlated with the covariates in the model.

And this is, I guess, one takeaway from an econometrics standpoint for sort of any lecture on this topic, that would be the takeaway. That from a theoretical perspective this is when you could choose random effects over fixed effects. But as we talked about in the last several slides, in applied research or in reality there may be different factors to consider. Which again, I encourage people to read Clark & Linzer. It’s a very accessible overview of stats done and criteria and sort of a rubric for deciding between the two models and trading off this bias in your coeffecients versus the changes or variance in limiting variability or variance in you model with our random effects.

Molly: Thank you. What is the relationship between what’s covered here and multi-level modeling?

Dr. Brown: Yeah. And that’s a good question, too. So if we’re thinking hierarchical in your models and those sorts of things, I actually don’t have the reference here. But if you want a very good break down of what the relationship is and how they’re quite different, I’d suggest something by Duncan Chaplin called “Hierarchical Linear Models: Strengths and Weaknesses”. And I can send a reference for that. And they do give a breakdown of what some of the differences in terms are with respect to random effects and fixed effects in those models versus panel linear regression models.

Molly: Thank you. For our attendees, we do have just a few minutes left. If you have any remaining questions, otherwise I- ah, here it is. Please discuss how to random effects and fixed effects for prediction at the person observation unit.

Dr. Brown: So from an applied standpoint, if you were using a statistical package, like Stata, you would sort of set your data up so that the unit of observation is the person. So there’s an option, I can send sort of code and sources on how to do that, depending on what your statistical package is. But you sort of preset your data so that your unit of observation is the person and you identify a time variable, time or year, whatever your variable is. And then you can run these regressions using xt set commands. If that’s sort of what the question is getting at? Basically you set up your data so that it is in a panel format and the unit of observation is at the person level.

Molly: Thank you. In a meta-analysis of several studies, should random effects or fixed effects be used for a summary estimate?

Dr. Brown: So that’s an interesting question, a little out of the scope of this. But I have come across some references, if you want to provide contact information, about which to use in a meta-analysis context. Contact me directly if there’s questions. My email is Josephine.jacobs@va.gov. And if there’s any questions or if you want follow-up references to questions like that, I can certainly provide them.

Molly: Thank you. So again, that is the final pending question at this moment. So I will toss it back to you for any concluding comments, Dr. Jacobs.

Dr. Jacobs: Great. Yeah. So just thanks everybody for attending today. I hope that this was a good introduction, or a new introduction or a different way, perhaps, to look at the terms fixed and random effects. And if there’s any follow-up questions, I’ll just reiterate please feel free to contact me at Josephine.jacobs@va.gov and I’d be happy to follow up with that or references to anything if I don’t particularly have the answer.

Molly: Thank you. Do you have your email address up on any of the slides? Is it on the last one by chance?

Dr. Jacobs: I actually don’t, which I just realized as I was saying that.

Molly: No, that’s fine. Can you spell it out and I’ll write it down for these people?

Dr. Jacobs: Sure. Actually I’ll just go to the first one. It’s Josephine.jacobs@va.gov.

Molly: Well thank you so much for coming on and lending your expertise to the field. And I also want to thank the attendees for joining us. At this point I am going to close out the meeting. Please wait just a minute while the feedback survey populates on your screen and take just a moment to provide us some feedback. We do look closely at your responses and it helps us to improve our presentations as well as the program as a whole. So once again, thank you everybody. And have a great rest of the day.

[END OF AUDIO]