**Advanced Algorithmic Trading and Portfolio Management**

**Prof. Abhinava Tripathi**

**Department of Industrial and Management Engineering**

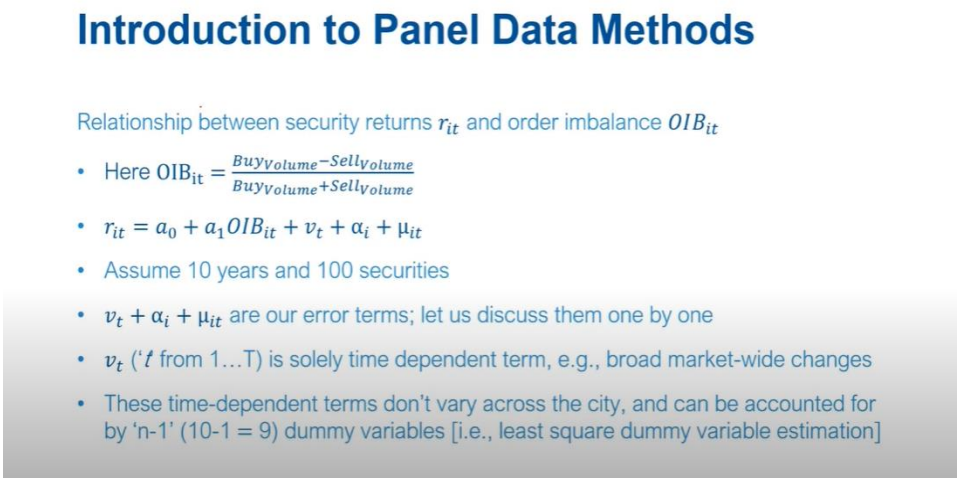
**Indian Institute of Technology, Kanpur**

**Lecture 22, Week 6**

In this lesson, we will introduce panel data algorithm. We will start by highlighting the properties of panel data. In particular, we will discuss the problems associated with pooled-OLS estimation of panel data. Next, we will discuss the least square dummy variable approach to panel data estimation. Then we will discuss the first difference estimation of panel data. Then we will discuss fixed effects and random effects estimation of panel data.

We will also discuss the scenarios in which these fixed effects and random effects indicators are more suitable. We will also discuss the model diagnostics with these different panel data approaches and under what conditions each of these approaches is more appropriate than others.

In this video, we will provide a brief introduction to panel data methods and the background and setting to the problem statement.




Order imbalance is a very important variable in financial markets. Its construction is done by Buy volume minus Sell volume divided by Buy volume plus Sell volume. The value of Sell here can be either dollar volume or number of shares or number of orders. This

measure is a very important measure of information arrival. It varies from minus one to plus one. A very low value of minus one would indicate all selling.

That means all the orders are sell orders while a value of plus one would indicate a lot of buying, almost absolute buying in fact. And a value of zero would indicate equal number of buy and sell orders whether in terms of dollar, number of volume or number of shares or orders. Let's think of the simple equation where we are trying to examine the predictability of returns from this OIB measure as independent variable and returns as dependent variable. Alpha nought is the constant term. But for us, the remaining three more terms, vt, alpha and mu_it are very important.

$$OIB_{it} = \frac{Buy_{Volume} - Sell_{Volume}}{Buy_{Volume} + Sell_{Volume}}$$

$$r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$$

Let us assume a period of maybe 10 years and 100 securities. So it's a panel data kind of setting where multiple securities are being tracked over multiple time periods. Let's discuss these terms vt, alpha i and mu_it one by one. These are sort of our error terms which we are not accounting for in the model. vt here is solely time dependent.

$$r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$$

It varies from one to three and up till time T and it only captures those influences that do not vary across individual securities, but rather vary across time such as broad market wide changes. For example, changes in monetary policy, maybe changes in policy rate like, repo rate in India or Fed rate in US. These time dependent terms, they do not vary across city. And those of us who have done some kind of regression course and dummy variable analysis would recognize that simply accounting for n minus one or rather t minus one which is in this case if there are 10 years, nine dummy variables can easily account for these vt factor or vt. But what is this vt? This vt is the average of all those influences that are not varying across individual securities, but they are varying over time.

So sort of trending with time and it is the average effect of all those changes for a given period. The next term of interest is alpha i. Alpha i is a security specific term and if there are n securities, it moves from one to three and up till n. Factors like firm size, firm beta, industry that are not changing with time, but rather changing with security to security. These are time invariant security specific terms.

What are these terms? Generally, these alpha i are like we discussed these factors, the aggregate sum of all those factors will be loaded on this alpha i. All such influences that

are security specific will be loaded on this alpha i term. In general and also in this example, like we discussed, this t period is rather small, while the number of entities like securities in this case is rather large. So it is easy, quite easy to model this t through dummy variable approach. But imagine, for example, if you have 100 to 1000 securities and you put n minus one dummy, for example, 99 or 999 dummy variables, that would make model extremely un parsimonious.

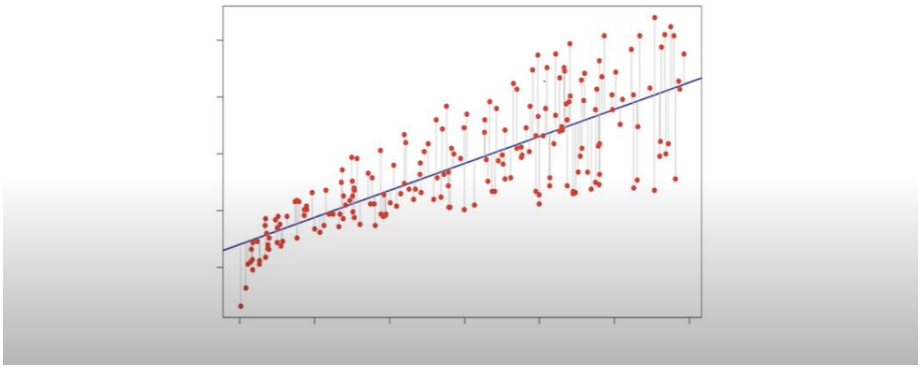
And therefore, accounting for this alpha i or what we also call in technical terms unobserved heterogeneity through dummy variable method can make the model extremely inefficient. Although we will discuss this dummy variable method as a part of LSDV least square dummy variable method later on as well. But I hope we have some intuition by accounting for, let us say we have only two periods t equal to zero and t equal to one. And I want to account this through a dummy variable or other period t equal to one or t equal to two. Then I can place a dummy variable d, which takes on value of zero when period t equal to one and d equal to one when dummy variable is two.

Usually the dummy variables are coded in zero one format, which has its own nice and useful mathematical properties. And also, depending upon your categories, for example, time periods, you tend to use n minus one dummy. For example, if you have 10 time periods, you would employ 10 minus one that is nine dummies and the constant term will automatically load the properties or attributes of the missing 10th period here on alpha nought. So, we have understood that we can account for this vt, but what if we do not account for this unobserved heterogeneity alpha i. To summarize this video, we discussed how a simple model like written predictability of order imbalance can create problems this OIB can create problems with rit when it is in the form of panel data.

Due to this unobserved heterogeneities that are time invariant and security specific and security invariant time specific. We also noted that many times in practical applications, we can account for this time changing term vt, which is changing with time but not with security through dummy variables when time dimension is small. But accounting for this individual dimension, alpha i is not as easy and it can create problems as we will see next set of videos.

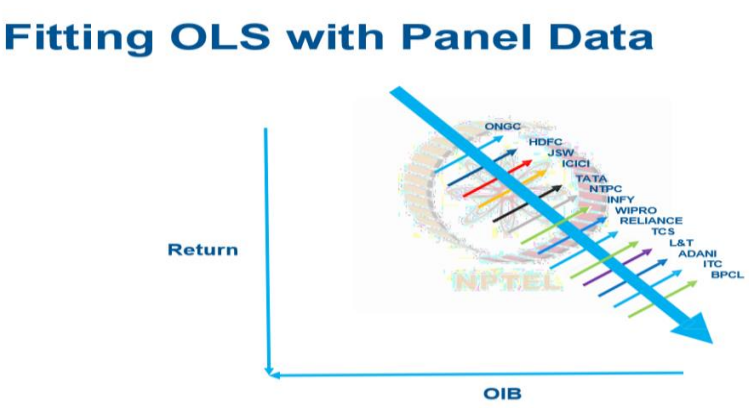
In this video, we will try to understand visually as well as econometrically the problems with OLS estimation of panel data. More specifically, we will understand how that unobserved heterogeneity would affect the OLS estimation of panel data.

## Fitting OLS Through Scattered Data Points

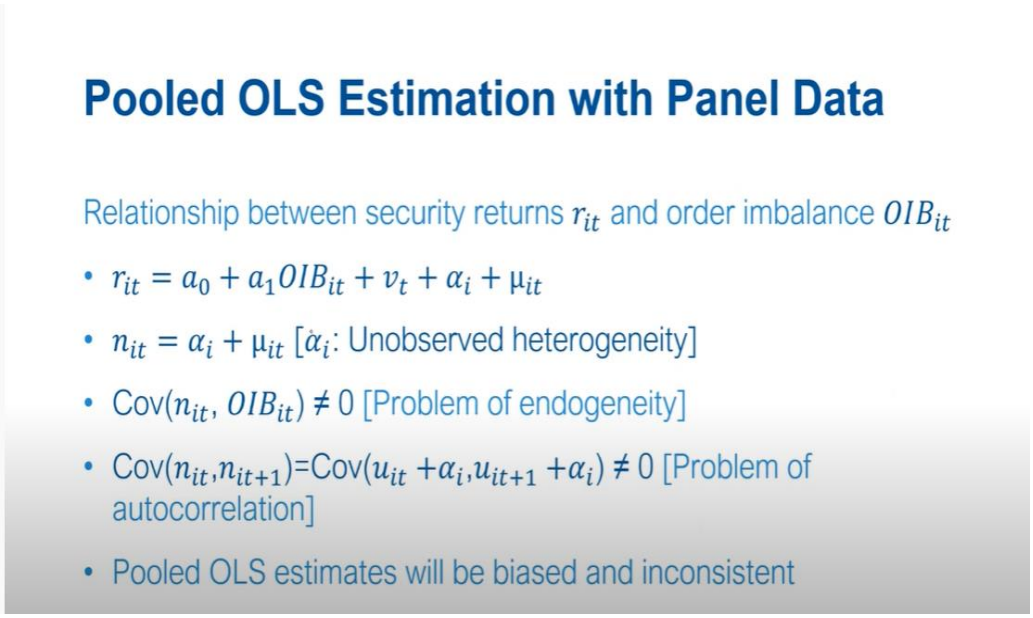


Let us recall the OLS fitting procedure. You would have number of scattered point on X, Y, axis like this and you would fit the line of best fit such that this error term, this error mu i and the summation of squares of these mu i terms across all the scattered fit points is minimized. That is what you do with OLS fitting, ordinary least square fitting of X, Y points. However, when we do this kind of fitting with panel data, the following problem may arise. Recall our relationship between rit return when it is regressed on our OIB variable, order imbalance variable.

Intuitively you would expect a positive relationship that is when OIB is negative, you would expect returns to be negative following prices or OIB to be positive and therefore rising prices or positive returns. So therefore a positive coefficient for alpha 1. Now imagine, let us say we have time, two time periods t equal to 1 and t equal to 2 and only two observations for each security. While the relationship is positive like this, increasing like this for all the securities and commonsensically when you would perform this regression, you would look at the slope of all these lines individually first and then average them out to find the average correlation, this average impact of OIB on returns. In this setting, you will find the average impact of all these securities by averaging these positive slopes and to get the relationship between OIB and returns.




However, the OLS may not think in this way. What OLS fitting would do is look at these points like this and may feel that average fit line with a negative slope like this may be more relevant and therefore may, it may turn out that this alpha 1 is negative and even significant, which is totally counterproductive and spurious. While this is an extreme case or rather more extreme case or spurious case, but it drives on the point very well, how panel data may get affected by such vitiating issues. Let us also look at econometrically what happens with panel data estimation when done through OLS. Remember this equation, which captures the impact of OIB on returns.




Here, we noted that alpha i, the unobserved heterogeneity is not accounted for by any variable. It may be size of the security, may be beta of the security. For the time being, we will ignore this vt term, which is security invariant term, which is moving with time. We may assume that we have modelled it through dummy variable or something like that. So we will ignore it for the time and we will focus on this unobserved heterogeneity alpha i, which is specific to securities.

This term, because it is not being explicitly modelled through any variable will get mixed with this error mu I t and therefore the resulting error eta I t, let us call it eta i t is summation of alpha i and mu i t. So this unobserved heterogeneity gets mixed up with the error term, which was supposed to be purely random, purely random term mu i t, which varies with i and as well as t. So the subscript i t purely random. Usually, if you recall the regression modelling, this would be modelled through some kind of assumption about distribution like normal distribution. Now the problem is this new error term, which is eta i t carries alpha i, which is more specifically a property of security i.

$r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$

$n_{it} = \alpha_i + \mu_{it}$ **[$\alpha_i$: Unobserved heterogeneity]**

**Cov($n_{it}$, $OIB_{it}$) ≠ 0 [Problem of endogeneity]**

**Cov($n_{it}$,$n_{it+1}$)=Cov($u_{it} + \alpha_i$,$u_{it+1} + \alpha_i$) ≠ 0 [Problem of autocorrelation]**

And therefore, with the high probability, it is correlated with order imbalance term, because order imbalance also varies, may vary with security. For example, size. Now alpha i may represent the effect of size on return, but also depending upon the size, the OIB may also differ across security, security for example, a particular amount of trade volume for a large security may have a different interpretation in terms of OIB while for a very small security, the same volume of trade may have a different interpretation. Same goes for beta and various other similar parameters. And therefore, there is a high probability that this alpha i which is mixed into now the error term and therefore this revised error term eta i t may have some correlation with OIB.

From our knowledge of regression, we already know that this may result in issue of endogeneity when your error term is correlated with your independent variable. This may further result in issues such as biasness and inconsistency of the estimate alpha 1. Another problem which is not as significant as problem of endogeneity, but still important problem is that now your new error terms like eta i t and eta i t plus one will have a common term which is alpha i and this is true for all the error terms that is eta i t minus one, eta i t minus two, eta i t plus one and so on. Because of this common alpha i, this covariance of mu i t and mu i t plus one despite the fact that in the original errors, there may be no serial correlation. I repeat, even if there is no serial correlation, the original error terms mu i ts still this alpha i would have some common correlation and therefore this covariance would not be zero non zero.

It will be some finite quantity and therefore this also introduces the problem of serial correlation in error terms. And to summarize overall the pooled OLS estimates will be biased and inconsistent of this panel data. To summarize this video, we learned visually as well as econometrically that if we do not specifically account for the panel data

unobserved heterogeneity alpha i specifically and we try to estimate it through pooled OLS, the resulting regression may be the estimates from the resulting regression that is alpha one are biased as well as inconsistent and therefore vitiate the estimation.

In this video, we will discuss in a formal manner the least square dummy variable approach to panel data estimation. Recall our earlier example of return being regressed on order imbalance with a formal model like this where alpha is the unobserved heterogeneity which we did not account for specifically the model.

$$r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$$

## LSDV Estimators

Relationship between security returns $r_{it}$ and order imbalance $OIB_{it}$

- Assuming that time-varying effects can be modeled using time-dummies
- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$ (1)
- Include 'N-1' dummy variable for 'N' securities $(S_2, S_3, \ldots, S_N)$ as follows
- $r_{it} = a_0 + a_1 OIB_{it} + \sum_{n=2}^{N} a_n S_n + \mu_{it}$ (2)
- Here, $S_2$ is a dummy variable that takes a value of 1 for security 2, and 0 otherwise; and so on for securities 3, 4,…, N
- Thus, we are explicitly accounting for the unobserved heterogeneity for each security individually

Now as per the dummy approach, let us say there are n securities s1, s2, s3 and up till sn. So, we can use n minus 1 dummy variables like s2, s3 and so on up till sn. And the resulting model will appear like this rit equal to alpha naught plus alpha 1 into y bit and so on summation an into sn plus mu it, mu it is the error term here. Notice that we have not accounted for s1, the first dummy which will be specifically loaded on this alpha nought. So, the effect of s1 dummy will be loaded on alpha nought.

$$r_{it} = a_0 + a_1 OIB_{it} + \sigma_{n=2}^{N} a_n S_n + \mu_{it}$$

Now, how to interpret these dummy variables? For example, the first dummy here is the s2 and its coefficient alpha 2. So, s2 is the dummy variable that will take a value of 1 for security 1 and 0 for all the other securities and similarly s3 will take a value of 1 for security 3 and 0 for all the other securities and so on. The coefficient corresponding to s2 which is alpha 2 reflects the impact of this particular dummy variable or unobserved heterogeneity corresponding to security 1 on returns. Similarly alpha 3 would represent the unobserved heterogeneity impact of security 3 on returns and so on. While alpha nought here would compute or account for the impact of unobserved heterogeneity

corresponding to security 1 which is not here in this model, but by design will get loaded here.

$$r_{it} = a_0 + a_1 OIB_{it} + \sigma_{n=2}^{N} a_n S_n + \mu_{it}$$

If you specifically account by putting all the n variables that is n dummies s1 to sn, then what you will fall in what we call as dummy variable trap or the issue of perfect multicollinearity and therefore your model will not run. So ultimately, we are explicitly accounting for the unobserved heterogeneity corresponding to each security and the advantage of this model unlike the previous model that we discussed or some of the panel data models and OLS approach as well that you get to estimate the impact of unobserved heterogeneity explicitly here. So, you measure that. Let us discuss some of the assumptions and conditions that facilitate estimation through LSDV estimator. So, one point to remember that once you have accounted for this model through dummy variables, now you estimate this through OLS only.

$$\mathbf{Cov}(u_{it}, OIB_{it}) = 0$$

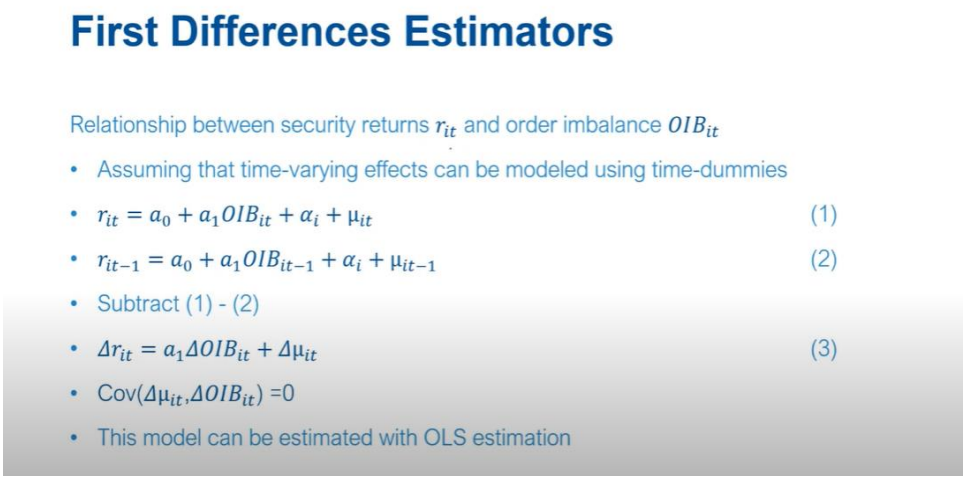
So, you have accounted for the an observed heterogeneity through dummy variables, then you use OLS approach to estimate the model. So, this OLS approach or rather LSDV pooled model through OLS approach, the coefficients here alpha i's are consistent under the following conditions. If the original error term which is mu it the original error and the independent variable order imbalance, they have no serial correlation and also there is no correlation or covariance between original terms and OIB which is 0, there is no serial correlation in errors that is mu it and mu it minus 1 or previous serial, previous terms are uncorrelated, this correlation is 0. So, when I say correlation covariance, I essentially mean the same thing. So, if the correlation or covariance is 0 and also there is no homoscedasticity in error term that means the variance of error term is constant across different values of independent dependent variables, it is not systematically varying.

The variance of error term is not systematically varying which is called homoscedasticity. And under these conditions, we can take our estimates of alpha i's as consistent. Later, we will also discuss another fixed effects approach. Under these assumptions, the estimates from LSDV are exactly identical to fixed effect estimates if these conditions are held in a theoretically identical sense. And also the advantage of this approach or FE approach which we will later discuss is that in this dummy variable approach, we can estimate the unobserved heterogeneity alpha i explicitly unlike FE estimator, fixed effect estimator that we later discuss, where we will tend to eliminate them from the model through a small and intuitive procedure we will eliminate these alpha i's in the fixed effect approach.

But here, we can explicitly measure it well, which is important if your objective is also to measure the impact of alpha i on the dependent variables or return in this case. However,

the problem of this model is that model is not parsimonious as n tends to increase the model become less and less parsimonious and with all the negative or adverse effects that come with a not parsimonious model, where so many variables or dummy variables are introduced. To summarize in this video, we examine the LSDV approach to paneel data methods. We noted how to incorporate or explicitly introduce dummy variables to account for the unobserved heterogeneity alpha i in the model. We also saw that if with certain assumptions of endogeneity, for example, the original error terms and independent variable not correlated, no serial correlation errors and homoscedasticity in error terms, then the estimates from LSDV method are consistent. However, with the introduction of more and more dummy variables, the model may not be parsimonious.

In this video, we will discuss the first difference approach to panel data estimation. The first difference estimators, let us start with the introduction. So again, we will recall our relationship between returns and order imbalance. We assume that the time varying effects such as vt are explicitly accounted through dummy variables.




So our focus will remain on individual unobserved heterogeneity, which is security specific that is alpha i. Now, the original model we had this alpha i here. Assume in the original model, we take a time lag that is for all the terms with subscript t, we take it one time before or lag it. So we get the resulting, for example, return as rit minus one. Similarly, order imbalance i t will become order imbalance for previous period for security i y bt minus one and so on.

$r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$

$r_{it\text{-}1} = a_0 + a_1 OIB_{it\text{-}1} + \alpha_i + \mu_{it\text{-}1}$

For the error term mu i t becomes mu i t minus one, which is one period before. So we have equation one which reflects the dynamics at time t equal to t and equation two

which reflects the relationship at time t equal to t minus one. Let us subtract one from two and we will get the resulting term here. So this delta rit is nothing but the difference of these two returns. Alpha 0, alpha 0 will cancel each other out or rather we can call it a0 or alpha 0.

$$\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta \mu_{it}$$

$$\mathbf{Cov}(\Delta \mu_{it}, \Delta OIB_{it}) = 0$$

Here the second term would be the delta OIB it and the difference in error. Now as long as the covariance or correlation between delta mu i t and delta OIB it equal to zero, the model can be estimated with OLS and the estimates of alpha would be consistent. Just a quick diversion, I often refer covariance and correlation interchangeably. Correlation is just a standardized version of covariance which is covariance divided by standard deviation of entity one into standard deviation of entity two. So covariance between one and two divided by standard deviation of one and standard deviation two which reflect correlation between one and two.

Now let us see what are the issues with this kind of first difference estimator. So once you have this kind of model which is change in returns as a relationship between change in OIB where coefficient is a1 plus the error terms which is delta, you artificially introduce correlation in error terms. For example, the covariance or correlation between change in error terms at t versus change and change in error terms at t minus one has a common term which is mu i t minus one. So there is some kind of correlation even though the original errors may not have any kind of correlation but still there is some correlation driven by this first differencing process, FD process or FD system, FD transformation will lead to some kind of serial correlation errors.

$$\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta \mu_{it}$$

$$\mathbf{Cov}(\Delta u_{it}, \Delta u_{it-1}) = \mathbf{Cov}(u_{it} - u_{it-1}, u_{it-1} - u_{it-2})$$

Also the original variables ri t and OIB were at levels. However, when we do this first differencing transformation, the changes in variables are maybe they may be much smaller and therefore the variation in these variables, delta ri t and delta OIB which may be very small and therefore the relative values of error and what we call as standard error of estimates become relatively larger as compared to the variation in these terms, the original variables because now we are looking at these variables as changes not as levels. So the variation in the resulting variable may be relatively very small as compared to the variation in error and therefore it affects the power of test. Secondly, you also have loss of observations due to differencing. If you have n, let us say if you have for each security, there is loss of first observation at t equal to one and therefore if there are n securities,

you have loss of n observations. So here I cannot account for time specific or those are the terms that are time independent.

$$\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta \mu_{it}$$

For example, alpha i is any term which is time independent will be eliminated in this procedure like we saw that alpha i and a0 the constant terms were eliminated which were time independent. And therefore in this resulting equation, because all those terms with no variance across time will be eliminated, we ought to have sufficient variance and the independent variable, they ought to have sufficient variance across time and also for a proper estimation, they should also have some variation across city otherwise the estimates of this alpha 1 will be very spurious in nature. So we need variation of terms not only across time, but also across city as well. To summarize in this video, we saw how first difference estimation can help us accounting for or eliminating the unobserved heterogeneity, the vitiating effects of unobserved heterogeneity. However, we also saw that it comes across, it comes at certain price and cost. For example, loss of observations and serial correlation error terms which may vitiate the estimation process to some extent.

In this video, we will introduce fixed effects estimator, a very important class of estimators for panel data. Let us recall the original model where we were examining the relationship between order imbalance and returns, where alpha i was the unobserved heterogeneity which was vitiating our estimation. Now consider a time-demean equation. When I say time-demean that means for every security i, let us say the return variable average is taken for all the time.

## Fixed-Effects Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$ (1)
- Time-demean equation (1) $\frac{1}{T}\sum_{t=1}^{T} r_{it}$ $\forall$ i's = 1, 2, 3…N
- $\bar{r_i} = a_0 + a_1 \overline{OIB_i} + \alpha_i + \bar{u_i}$ (2)
- Substract (1) - (2)
- $r_{it} - \bar{r_i} = a_1 (OIB_{it} - \overline{OIB_i}) + (\mu_{it} - \bar{u_i})$
- $\tilde{r}_{it} = a_1 * \widetilde{OIB}_{it} + U_{it}$;
- Here, Cov($\widetilde{OIB}_{it}, U_{it}$) = 0, and pooled OLS estimates will be consistent

$$r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$$

For example, for security i equal to 1, we take average for all the times from t equal to 1, 2 and so on up to time t. This time averaging is reflected with this term rit summation t equal to 1 to t, rit divided by t for all i. Now for all the variables including returns and order imbalance, we can perform this time averaging and resulting time average equation is represented by equation 2. So this is time averages of variables. If we subtract this equation 1 from 2, this is called time-demeaning process or time-demeaning transformation where I subtract each observation with its corresponding time-demeaning value.

For example, for security 1, rit minus ti bar and therefore we get this kind of transformed equation. Notice the set of terms that is time-invariant like alpha i, the time-invariant or time average is nothing but alpha itself because it is not changing with time and therefore in this subtraction process such terms like alpha and alpha nought that are time-invariant will be eliminated. And therefore the resulting equation this rit minus ri bar equal to a1 into OIB it minus OIB i bar plus error term that is mu it minus mu i bar can also be represented as tilde. So we replace this with tilde and write rit tilde equal to a1 into OIB tilde it plus error term it. This estimation, this resulting transformed system can be estimated with pooled OLS as long as this the covariance between the resulting independent variable which is OIB tilde it and error term mu it.

If there is no covariance or correlation between these, then the resulting system or the transformed system can be estimated with pooled OLS and the estimates of a1 will be consistent in nature. However, please note that fixed effects also remove time constant terms like alpha i and they are also costly because now you have transformed the system and you are not estimating the original variables at level but rather this transformed or time-demeaned equation the transformed fixed effect system. To summarize this video, we discussed the fixed effect estimators and how to obtain the fixed effect transformed system that is to be estimated with pooled OLS.

## Random Effects (RE) Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$
- Recall that the model would have an issue of endogeneity if the unobserved heterogeneity ($\alpha_i$) is correlated with one of the independent variables: $Cov(OIB_{it}, \alpha_i) \neq 0$
- Thus, pooled OLS is not effective, and we used FD/FE methods to remove $\alpha_i$ from the model
- However, if $Cov(OIB_{it}, \alpha_i)$ is reasonably close to '0' then, we need not apply FD/FE as they involve a heavy transformation in data
- E.g., FE leads to loss of observations (T-1 periods instead of T)

In this video, we will introduce a very important class of estimator that is Random effects estimator. Recall in the previous discussions while examining relationship between order imbalance and return, we said that there is a high probability or with the issue of endogeneity if the unobserved heterogeneity that is alpha i is correlated with one of the independent variables that is in this case OIB.

$$r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$$

$$\text{Cov}(OIB_{it}, \alpha_i) \neq 0$$

So, if this covariance or correlation is non-zero, then there is a possibility of heterogeneity which has a high likelihood because there are a number of variables like size, beta we have ignored in this model and they may get aggregated or accumulated in this alpha i resulting in a covariance or correlation between these two variables. And therefore, we said the pooled OLS estimation may not be effective because of these issues related to endogeneity, the estimates may be biased and inconsistent and therefore, we needed heavy transformations such as first difference or fixed effect methods to eliminate this alpha i from the model. However, if you have a strong reason to believe that this correlation or covariance between unobserved heterogeneity and or independent variable is close to zero, then you need not go ahead for such heavy transformations as FD or FE as they require lot of heavy transformation data such as for example, FE transformation leads to loss of one observation, it involves variables at rather not at levels, it drastically transforms the original model. For example, FD leads to differences rather than level. Such drastic radical transformation in the original model are not very desirable in that sense and leads to less efficient estimators as we will see later.

However, if you have a reason to believe in this relationship that this covariance between OIB and alpha i is very close to zero, this sometimes is a reasonable assumption in some of the cases for example, you may have included all the relevant variables and accounted for them and therefore, you have less reason to believe for a large value of alpha i. Similarly, a priori from theory you believe that relationship is like this only and the presence of alpha i is likely to be very small relative to variables like OIB and therefore, this correlation may be almost zero. In this scenario, you may believe that pooled OLS or something which is closer to pooled OLS may provide consistent estimates. However, the error may still be serially correlated for example, this issue of covariance or correlation between revised error terms that is eta it which was original error terms mu it plus alpha i, this one may still be there. So, error terms may still be correlated across time and therefore, the serial correlation may exist but please note that the serial correlation can be easily corrected through random effects transformation without putting a heavy cost on data with transformation such as first difference or fixed effect.

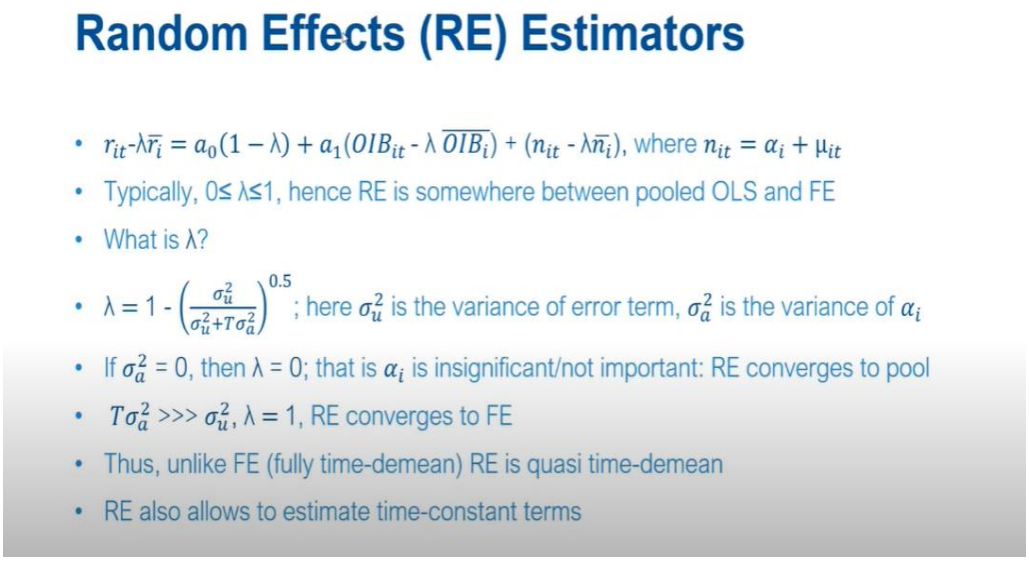
In general, it is noted that random effect is found to be more efficient than pooled or FE or FD. Why because it is believed that standard errors of random effect estimates or rather I should write beta RE estimates, the standard error of these estimates is often found to be lower than pooled OLS or FD or FE estimates. This is a very useful property which means the efficiency of this estimator is higher than FD FE or pooled OLS. And therefore, if you believe that either you have model very well specified so that sufficient variables have been entered in the model. So this covariance this endogeneity problem has been resolved that means essentially this is equal to zero.

So this problem has been resolved or for that matter you have reason to believe that alpha is very small then RE is better than FE and OLS and FD as well. And therefore, you can use a model like this for example, instead of completely fully time demeaning the model you can simply demean with the multiple of lambda which is between zero and one that is less than equal to one. And therefore, your resulting question is rit minus ri bar only that in the original FE transformation instead of completely time demeaning I am demeaning time demeaning with the factor of lambda. So this term is not eliminated you notice all the time all the terms which are time invariant they are not eliminated but they are still there like a nought into one minus lambda a one into this and eta it minus lambda times eta i where eta it is equal to alpha i plus mu it. Now, this is what we call as random effect system and this system is estimated with pooled estimation.

Please note in this if in this model if lambda equal to one this model transformed into what we called originally as fixed effect and if lambda is zero then it transformed into simply the pooled OLS and therefore, random effects is a sort of in between the spectrum of a two extreme ends which is pooled OLS and FE. So neither it is as extreme as FE nor as simplistic as pooled model. To summarize this video we noted that if we have reason to believe that unobserved heterogeneity alpha which was leading to all the problems such as endogenity and so on. If that is very small in magnitude and we have reason to believe that it has very low correlation with the independent variables and therefore, issue of endogenity is very not so major in our model only the issue of serial correlation may exist we can correct for the serial correlation with a rather less drastic transformation of the system what we call random effect. So this is sort of quasi time demean unlike the fixed effect this is like partial time domain or quasi time demean kind of transformation and this transformation is less drastic and the estimates from random effect therefore are more efficient as compared to fixed effect or first difference and pooled OLS estimates.

And the model is remains consistent the pooled OLS estimates of this random effect system transformation are consistent as long as we have reason to believe that this endogenity problem is not very severe.

In the previous video we discussed how random effects transformation is sort of in between two extremities that is pooled and fixed effect. In this video we will examine this issue in more detail in an empirical manner. Recall in the previous discussion we said this is the transformed random effect system in this system this parameter lambda determines the position of random effect between pooled and fixed effect estimator. For example, if lambda is closer to it varies between 0 and 1 if it is closer to 0 then the model is pooled and if it is closer to 1 then the model is fixed further to fixed effect.




Let's see how it works. So the formula for lambda is 1 minus sigma mu square upon sigma mu square plus t times sigma square where t are the time periods whole raise to the power 0.5 where sigma mu square is the variance of error term and sigma square is the variance of alpha i. Now recall if we have a reason to believe that this sigma square alpha that is the variance of unobserved heterogeneity is very very small relative to the error term. Let's say believe it is relatively very small or close to 0 then in that case we can ignore this and resulting value of lambda is 0. So, in the case where alpha and its size and variance is insignificant not very severe then ri converges, random effects converges to pooled model.

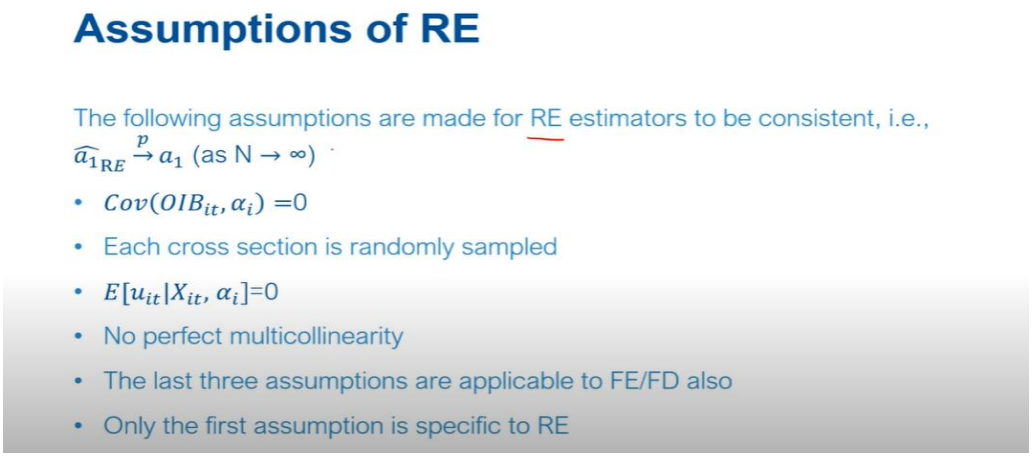
Similarly, if relative to error variance if the size and variance of this unobserved heterogeneity is quite large and therefore 2 into sigma alpha square is quite large as compared to sigma square mu then this term is 0 and lambda becomes 1 close to 0 asymptotically if close to 0 then this lambda becomes 1 and then random effects converges to fixed effect. Also, we noted that unlike fixed effect which is called fully time demeaned random effect is sort of quasi time demeaned so we are time demeaning but not with a factor of 1 but something which is less than 1 but greater than 0. And another advantage of random effect is that it allows us to estimate the time constant term

for example alpha 0 into 1 minus lambda so it allows us to estimate those time constant term that is one great advantage of this method. The way to estimate random effects model here is to estimate this system of equation 1 and 2. However the problem is that in the first stage it requires estimation of lambda which is not directly observed and therefore first you estimate equation 1 through fixed effect for different or OLS method so you need to estimate this so that you can find estimates of error terms as well as you try to estimate the unobserved heterogeneity.

Once you have these terms you can estimate lambda. Once you estimate lambda then only you are able to estimate the system of equation 2 and then that is whole model. So this combined estimation is the random effect method of estimation and because of this sort of transformation where you first estimate lambda and then in the second step you estimate the original model this is often referred to as feasible generalized least square methods. To summarize in this video we discussed the estimation of lambda a very important parameter that empirically evaluates how large is the magnitude of unobserved heterogeneity. If unobserved heterogeneity is quite large then you have to go with model a random effect model which is closer to fixed effect and if this unobserved heterogeneity is very small then you go for a model which is closer to pooled OLS.

In this video we will talk about some of the assumptions pertaining to random effect estimator and also the estimation of the time invariant terms.

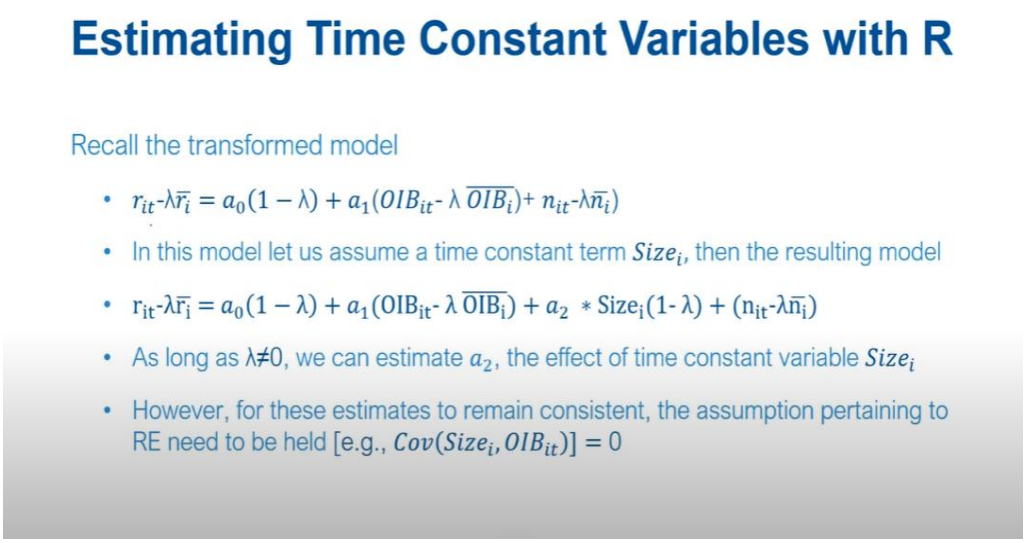
When we talk about consistency of RE random effect estimates asymptotically the following assumptions need to be held. So asymptotically random effect estimates have to converge to the population parameter a1 then first and foremost unobserved heterogeneity needs to be uncorrelated with the order imbalance term so that endogeneity properties avoided. So there is no endogeneity. Each cross section is randomly sampled. Third the original error terms mu i t are not are their expected value is 0 given your independent variable and unobserved heterogeneity.



So, for given alpha i for each security i given alpha i and independent variable X i t their expected value has to be 0. No perfect multicollinearity. Please note the last three assumptions this cross sectional to be randomly sampled expected value of error term to be 0 and no perfect multicollinearity are also applicable to Fe and Fd. So, this first assumption is specific to random effect which states that there is no correlation or covariance between independent variable and unobserved heterogeneity.

$Cov(OIB_{it}, \alpha_i) = 0$

$E[u_{it}|X_{it}, \alpha_i] = 0$

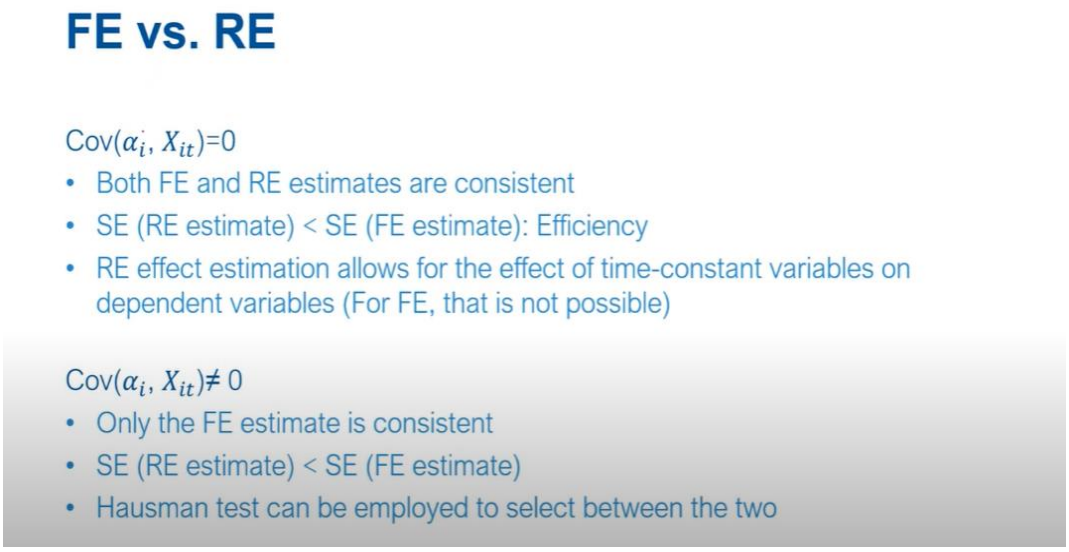



Now how to estimate time constant term with random effects in R. So recall our transform system of random effect model. This was our transform system. In this transform system let us say there is a time constant term size which is specific to individual entity i needs to be estimated and introduced in the model. Now because this term is time constant as long as the transform model will add another term which is size i into 1 minus lambda which is similar to this term and as long as the lambda estimated lambda is not equal to 0 that is model is not pooled OLS we can estimate its coefficient a2. As long as this lambda is not 0 we can estimate the a2 and we can estimate the effect of time constant variable size.

However please note again for these estimates a1, a2, a0 to remain consistent the assumption pertaining to random effects that is covariance between this time invariant term size and OIB needs to be 0. To summarize in this video we discussed some of the assumptions that are required for random effect estimators to be consistent some of the main assumptions and also we saw how to estimate time constant or time invariant terms with random effect model in our environment.

In this video we will discuss a comparison between fixed effects and random effects estimators. We will see what are the conditions that are more suitable to fixed effects vis-a-vis random effects and we will also examine the Hausman test as a selection criteria between fixed effects and random effects. In particular the most important condition while comparing fixed effects and random effects is the covariance or correlation between the unobserved heterogeneity alpha i and the independent variable xi whether it is equal to 0 or not.

If it is equal to 0 maybe for different reasons for example probably, we have accounted for all the relevant variables so that the magnitude and variance of alpha i is very small and in that case both the fixed effects and random effects are consistent because all those problems related to endogenity and so on are not there and therefore both the estimates fixed effects and random effects are consistent. However, as we noted earlier the efficiency of random effect is higher because the standard error of the random effect estimate is lower than the fixed effect estimate. In fact, it is lower than first difference and well as pooled estimate as well if this condition is true and therefore in this case, we choose random effect over fixed effect because these estimate both the estimates are consistent but the efficiency of random effect is higher. Also, one advantage is that random effect estimation allows for the estimation of time constant terms on dependent variables.



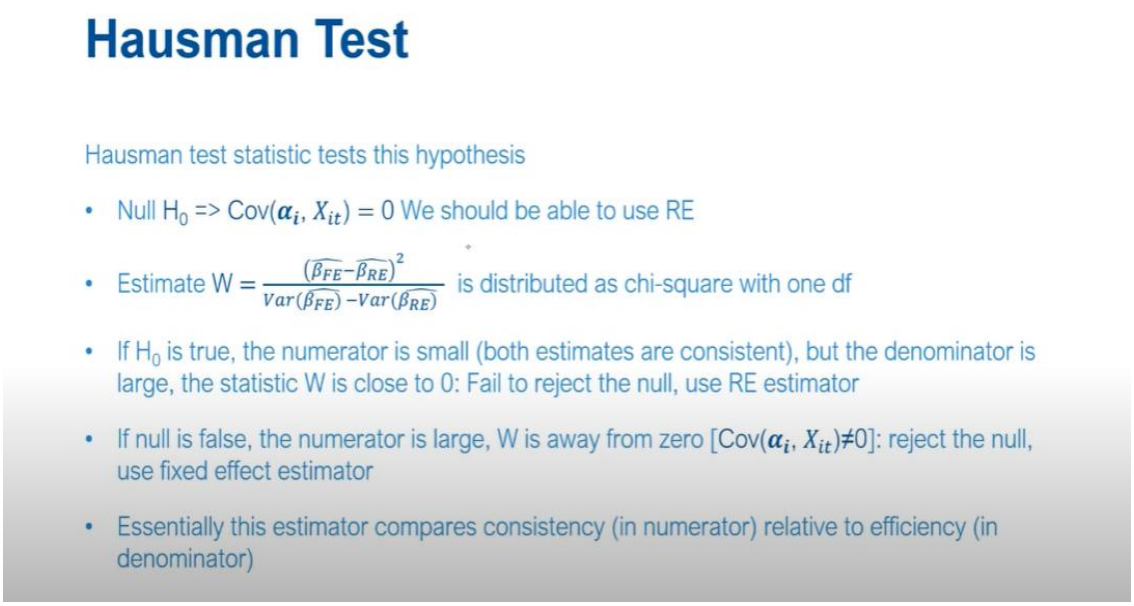

For FE that is slightly more tricky and not so easily possible. Also, in the second condition where this correlation or covariance between unobserved heterogeneity and independent variable is not necessarily 0 then all those problems related to endogenity may appear and estimates may not be consistent. In that case only fixed effect estimates are consistent and RE estimates are not consistent. Although still the efficiency or the

standard errors of random effect estimate is lower but because of this consistency property because that is for more desirable for us we will go ahead with fixed effect estimator. This particular dynamics can be tested with the help of Hausmann test.

**Cov($\alpha_i$, $X_{it}$)=0**

**Cov($\alpha_i$, $X_{it}$)≠ 0**

The Hausmann test statistic can help us in the selection criteria. Let us see how. Let us examine the Hausmann test statistic and its hypothesis. The null hypothesis here is that this covariance between alpha i, unobserved heterogeneity and X it equal to 0 and therefore both the models RE and FE are consistent and because of its higher efficiency we should be able to use RE. So, RE is better than Fe in this particular null hypothesis. The test is designed in a manner that on numerator you have the difference between fixed effect and random effect estimates and denominator has their variances.




So, this statistic is distributed as a chi-square with 1 degree of freedom. So, it appears something like this, the chi-square statistic with 1 degree of freedom. If the null hypothesis is true that is both the estimates are consistent and therefore both beta FE and beta RE they converge to the two population parameter beta and therefore the numerator is approaching 0 while the denominator because we already know that this number is lower than the variance of beta FE because random effect are more efficient, this number approaches to 0 because of the numerator and therefore we fail to reject the null that is we are somewhere here closer to 0 and we choose the random effect estimator that is both the model we conclude that both the models are consistent and we choose the random effect. However, if we reject the null and when we will reject the null in the case

where the both the estimators are not consistent and therefore there is a substantial difference between them and the difference is so large that not only it is much larger than their differences in their variances, it is so large that we are able we are in the rejection region, we are able to get the hypothesis, null hypothesis and therefore we are in this rejection region which leads us to believe that this covariance is not 0 because in that is the only that is the case when this both of these will not be consistent, fixed effect will anyway be consistent, random effect will not be consistent and there will be substantial gap between them. So, when we are rejecting the null, we assume that there is a lot of gap between these two in case we are in the rejection region and therefore not only random effect is not consistent, it is way further apart from the fixed effect.

So, we choose the fixed effect. So, essentially if you look at this test statistic, the numerator is sort of consistency property and denominator is sort of efficiency. So, here we are looking for a sort of trade-off between consistency in numerator for efficiency in the denominator. To summarize this video, we discussed the condition in which we should use random effect or fixed effect. We noted that this covariance between unobserved heterogeneity and X it independent variable, is what determines whether we go ahead with random effect and fixed effect. We also saw the construction of Hausman test and how it helps us in determining depending upon acceptance or rejection of null to choose random effect or fixed effect.

So, if we fail to get the null, then we consider both the statistics fixed effect and random effect as consistent and therefore we go ahead with random effect for its efficiency. However, if we reject the null, then we believe that not only fixed effect is consistent, random effect is inconsistent and way out of the mark and therefore we go ahead with the fixed effect as consistent.

In this lesson, we discussed panel data methods. We started with a brief discussion about panel data properties. We highlighted that the unobserved heterogeneity associated with panel data makes OLS estimation of our choice. The estimates are biased and inconsistent.

We started with the least square dummy variable approach where dummy variables provide an easy and simple solution to deal with this unobserved heterogeneity. Next, we discuss how first difference approach can be very useful in tackling the problems associated with this unobserved heterogeneity by simply eliminating time invariant terms. Next, we discussed fixed effects approach which is effectively time demeaning the properties of the data. This approach also leads to eliminating time invariant terms and leading to unobserved heterogeneity reduction. Next, we discussed random effects approach which is somewhere between two extremes that is pooled and fixed effect approaches.

Random effects approach, quasi-demeans the data when the issue of heterogeneity is not as severe and only serial correlation in others is the main issue, the fixed effects is rather more extreme treatment. In this case, random effects can account for serial correlation and do the job satisfactorily. We also discussed the intuition behind the Hausmann test to select between random effects and fixed effects methods. Essentially, Hausman test evaluates the trade-off in estimation across consistency and efficiency. While RE estimates are more efficient and if the unobserved heterogeneity issue is not that severe, then both random effects and fixed effects are consistent.

Hence, random effects become more suitable. However, if the unobserved heterogeneity is a major issue, then only fixed effect is consistent and suitable. Thank you.