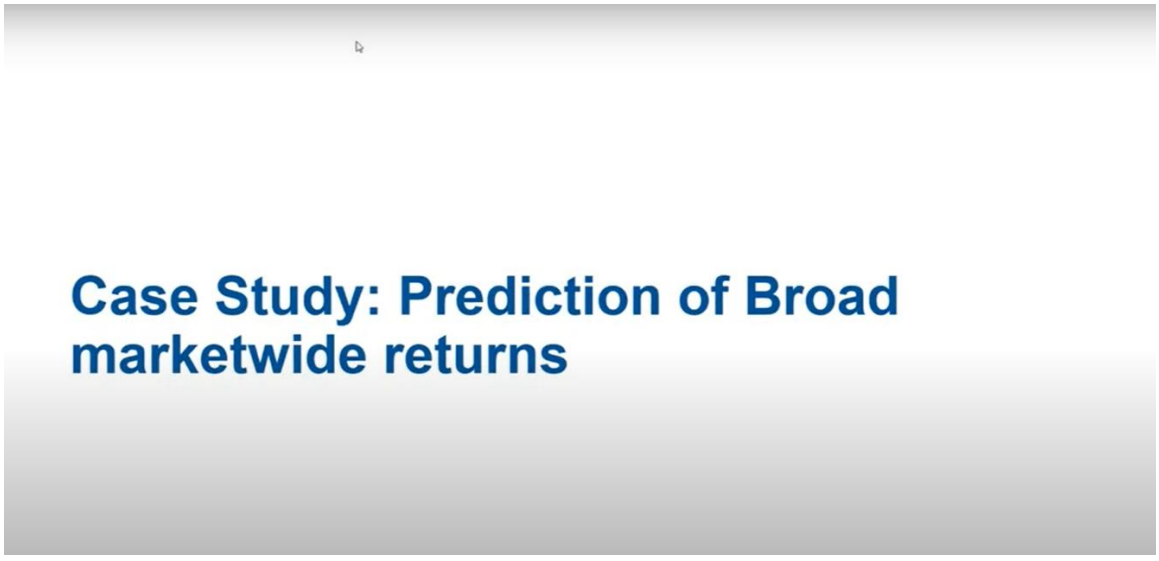
Advanced Algorithmic Trading and Portfolio Management
Prof. Abhinava Tripathi
Department of Management Sciences
Indian Institute of Technology, Kanpur
Lecture 23, Week 6

In this lesson, we will examine the application of panel data methods through a case study implemented with R. First, we will introduce the case study, and then we will load the data, visualize it, and explore it for various properties. We start the analysis by fitting the OLS and discussing the issue with the OLS estimation of panel data. Next, we apply the least square dummy variable approach and see how estimates improve. Subsequently, we will apply time and individual fixed effects and compare the results. We will also see how to extract these fixed effects.

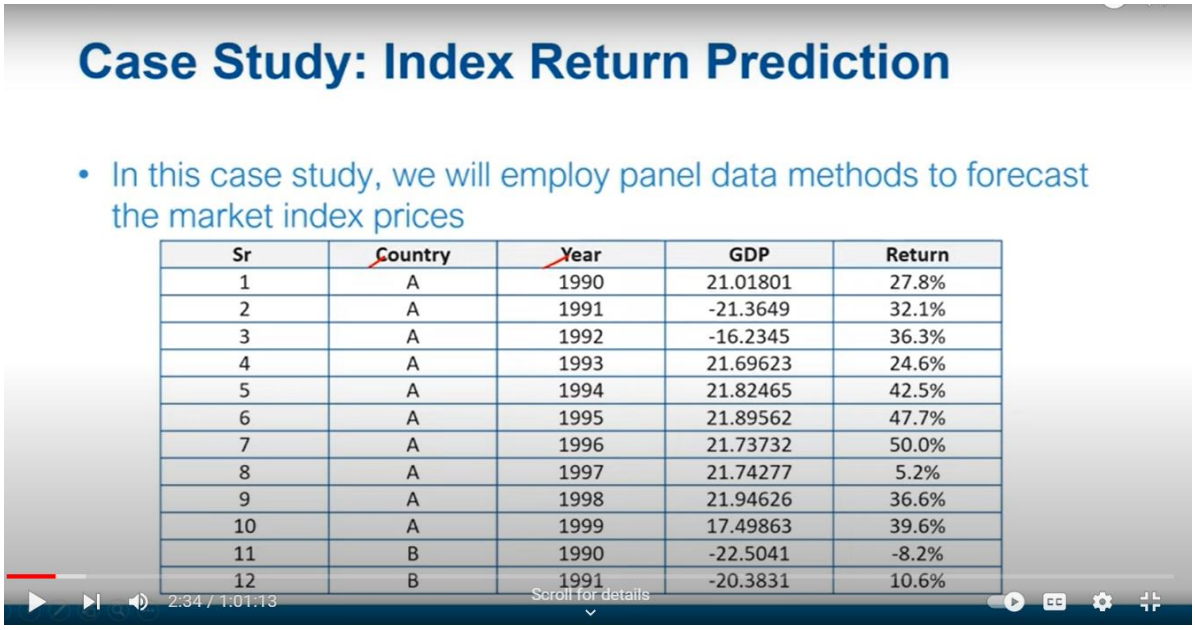
We will also compare the first difference estimates with fixed effects estimates. Next, we will fit the model with a random effect. We will also examine whether random effect or fixed effect is the suitable model using the Hausman test. Next, we will conduct error diagnostics and lastly, we will perform statistical inference using robust standard errors.



In this video, we will introduce our panel data case study on the prediction of broad market wide index returns. Broad market-wide indices such as nifty-50 are one of the most advanced and sophisticated forms of portfolios, very well diversified portfolios. These broad market-wide indices are known to reflect the growth of the economy and are often correlated with macroeconomic factors such as GDP. Therefore, this strategy to invest in the market based on forecasts related to factors such as GDP has become a very prevalent strategy known as factor investing which is investing in the market based on predictions related to GDP. The idea is that GDP has a fundamental relationship with

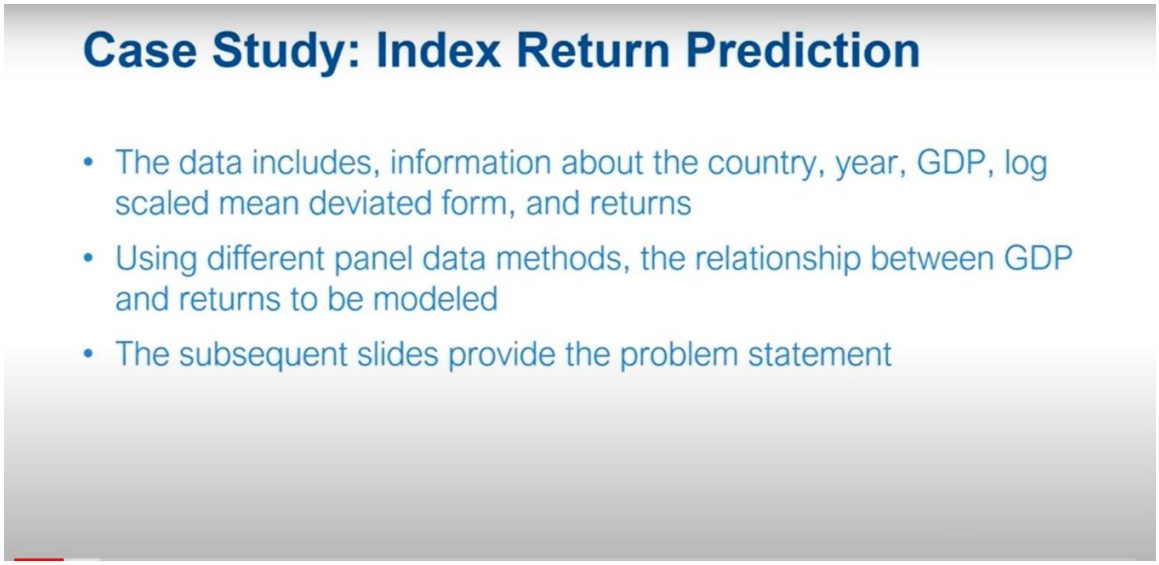
these broad market wide indices which essentially reflect the economic activity in a country.

However, this exercise may be vitiated if the data is of multi-country panel data then this exercise may be vitiated by the unobserved heterogeneous as we discussed in the theoretical discussions such as country specific factors. The data employed in this study is of the panel form and is used to forecast market index prices or returns. So we have serial numbers, we have countries, year, GDP and return values. These values are scaled so they do not reflect exactly the GDP or return values but they are scaled to suit the case study. Using different panel data methods, we will examine the relationship between GDP and returns.



Sr	Country	Year	GDP	Return
1	A	1990	21.01801	27.8%
2	A	1991	-21.3649	32.1%
3	A	1992	-16.2345	36.3%
4	A	1993	21.69623	24.6%
5	A	1994	21.82465	42.5%
6	A	1995	21.89562	47.7%
7	A	1996	21.73732	50.0%
8	A	1997	21.74277	5.2%
9	A	1998	21.94626	36.6%
10	A	1999	17.49863	39.6%
11	B	1990	-22.5041	-8.2%
12	B	1991	-20.3831	10.6%

We will try to model it and in subsequent slides, we explore the problem statement. We will try to perform the following set of tasks. First, we will try to visualize the data. We will plot the year-wise market index returns for each country. We will also plot year-wise GDP for each country and we will try to make use of a box plot to show the heterogeneity in GDP and returns across countries and years and we will try to infer whether indeed the distribution of returns and growth of GDP over the years differ across country to country or carries heterogeneous across countries. We will try to model the relationship for GDP with broad market-wide index returns first using simple pooled OLS and we will try to infer what are the problems with this kind of approach. First, through a visual approach, we will try to see how the fitted line appears vis-a-vis actual data. Then, we will follow the least square dummy variable or LSDB model approach and model this relationship between GDP and index returns after adding country-wise dummies. Next, we will convert the data into panel format in our environment with country and year as data identifiers or index variables. Then, we will model the data using fixed effects using individual time and two-way both individual and time effects.



We will perform the test of poolability to see whether the fixed effects are indeed significant, or pool model is a decent model. Next, in case we find the fixed effects are indeed significant, we will try to extract these time and individual fixed effects from the modeled fixed effect object from the panel data regressions. We will also model the data with random effects, and we will examine the output and comment on individual and idiosyncratic heterogeneity. We will also comment on whether random effect transformation is closer to pool or fixed effect. If you recall whether lambda is equal to 1, then in that case it is closer to fixed or lambda is equal to 0, then in that case it is closer to the pool model.

We will also conduct tests such as the Hausmann test to examine whether a random or fixed effect model is more appropriate. Lastly, we will examine the cross-sectional and serial correlation in errors for different models such as pool fixed effect and random effects models. Then we will also examine the homoscedasticity and heteroscedasticity-related issues in the errors that is we will perform error diagnostics lastly as a solution we will use robust standard errors to estimate coefficients that are robust to error autocorrelations and heteroscedasticity in the model. To summarize, in this video, we discussed the case study, we discussed how first we will visualize the data and see whether panel data methods are appropriate or not and then empirically. We will examine different models such as pooled fixed, and random effects and also first differences and see their goodness of fit. We will examine and perform residual diagnostics such issues such as cross-sectional and serial correlation errors and heteroscedasticity and lastly, we will use robust standard errors and try to perform statistical inference for these models.

Let us start our implementation of the panel data case study. First we will set the working directory. As usual we will go to the session command set working directory choose directory and we will choose the appropriate directory where our data is available. I will click on open. You will notice a command has appeared on my console window.

I will take up this command and paste it here. So, notice as soon as I set my working directory you will find some of the data files appearing on this file tab and I can read data from here. Now as a starting point, first let us load the relevant packages. So we will load the load the relevant library packages. In this particular session, we will be going using the car library then we will also be needing ggplot2 we will use plm and we will also we may also use the t-series package. These are some of the packages that we may require. Now we need to load the data. So we will read the data we can already see the data file so the easiest way to do is to click on these data files. So for example we have a file in CSV format I can click on import. The data file appears on my screen, and we can see how it appears.

We can see the shape country year GDP return. We have already discussed the data. Let's name it panel. We will name the data as panel data and I can click on import to read the data. That is one way to do it. I can read the data. A command appears on my console window I can for future reference I can copy and paste this command and also the brief view of data appears in this window so I can close it. I can paste these commands. Also, I have put the data in my file in RDS format. so I can click on this RDS format and then also I can click on OK to read this data. So that is also another way to read the data I can select the file name as I wish, and a command is run which I can again copy and paste for reference and this is also. Please recall for R this arrow symbol is the same as equal to our assignment operation so I can very well use this equal to symbol. So I have the commands as well ready for me. If I have read the data correctly then I can just have a quick look at it. For example, I can see the summary of this data through the summary command that is one. I can also run the brief command to have a look at the brief of my data so I can check what is inside data so I can check the brief command some of the initial variables. So first I have the country variable so the country is a factor variable then I have years so it's panel data I have GDP so GDP data is slightly scaled to suit this exercise so you may not find it as a regular GDP number so it's slightly scaled number and similarly I have return number as well. To summarize this video, we read the data we loaded the relevant packages we have set the working directory and now we are ready for basic exploration of our data through visualization and some other exploratory exercises.

Next, in a series of videos, we will try to visualize the data through various important visualizations. Let's start with a simple scatter plot command this is available in our package car package so we use this scatter plot command first will visualize return with the year and the moment of returns with the year and we will distribute the observations across countries so we will write this country and command will be with this pipe symbol country. We will write data as panel data because our data is named is a panel and we will put this smooth argument as F so that there is no confidence band provided.

So now we will run this command and the plot should ideally appear on the plotting window here as I run it you can see the plot window. The interesting thing about this diagram is that the lines show the relationship between returns over the years for different countries and for each country the scatter points are there you can see different shapes and also the line is given in the form of different colors. Notice the interesting thing that for each of the countries we can see the pattern of GDP and relationship the relationship has over the years behaved differently for different countries it's not the same for all the countries and that creates a kind of heterogeneity in this relationship across countries. So, this is the first step. Next, now that we have seen the return moment of return across countries, we have seen that it's heterogeneous in the same fashion, and very same fashion we will also see the behavior of GDP as well.

So, we saw the returns and now we will see the behavior of GDP. I simply paste the command I will not rewrite it and I will run the command. Again, very interesting thing we can make an interesting observation that GDP has also behaved in a very heterogeneous manner for example for some countries many of them it is increasing but there are few countries where it is decreasing as well. So, there is quite a heterogeneous behavior of GDP across countries. Let us also use the ggplot command to visualize in the box plot form, box plot is a very usual sort of very interesting and useful way to visualize univariate series. So we will make use of box plot for this ggplot command.

We use the ggplot command will supply the data which is panel data and aesthetic argument we need to provide which is on x axis we have GDP on the y axis we have so because this is univariate series there is no y axis we will group the data across countries so I will use this country as a grouping variable so very nice in a very nice way the data will be grouped across countries. And fill will use different colors so we will use the color according to country and now let us add the geom argument while the data will be printed now itself I can print my data here itself however let me show you first what exactly I mean by that so for example two layers of my code is complete so I can very well print the data will not discuss it now you can see here there will slightly make the plot more aesthetic with some more arguments as we will see. So first I will add the axis labels first is the title, title I will use heterogeneity across countries so this is my main title heading I will give the heading for x axis let me just minimize the plot window little so that I have some more space so x equal to GDP so on x axis I have my GDP this is third layer I have added next layer I will add theme so I will just specify some more properties of the plot so for example first the title so as there is a very interesting command element underscore text which will help me choose the font and size for example adjustment horizontal adjustment of the title size is 0.5 so it will be sort of in middle I will put the font as bold so font is dark bold and I will also decide its size the size may be 18 so it is slightly higher than the axis titles. Similarly in a very similar

manner I will specify the axis properties for example I will use this axis dot text and again the same element underscore text argument so I will just copy it and choose some argument for example I will use the size of 16 so I will use the size of 16 font I will keep as bold so it looks more sharp and the last layer the last layer I want to add is sorry not the last layer but the last argument here I want to add is axis dot title so I can first I change the properties of axis labels and now title also I will use the element text same element text again and I will give the argument same arguments I will use so now I am ready to plot so let me just slightly enlarge the plot window let me show you how it appears when I not so we can see a very interesting plot will appear now so you can see here so now in this plot box plot is a very interesting and usual way each box if you notice each box is a set of certain parameters for example the upper and lower ends of this box is basically the 25 and 75 percentiles Q1 and Q3 the median line is like 50 percentile or median point which is at the mid of this box the ends are viscous these viscous represent sort of range sort of range you can say for example from 75 percentile onward you will find a whisker and 25 percentile below you will find a whisker these are the range of the data and then you have certain extreme outliers pointed in the form of these dots so these dots are extreme outliers generally these whiskers are those points that are within certain range of 75 percent and above and 25 percent and below and what is this range this range is basically called interquartile range which is equal to the distance between the 75 and 25 percentile so this is the interquartile rate and this interquartile range is extended from the Q3 point here on the upside whisker and Q below Q1 which is the lower side whisker and beyond this you have extreme outliers in the form of these dots as you can see here.

If we examine this box plot we can see that across countries we can see our data there is a lot of heterogeneity in the GDP data so there is a lot of heterogeneity in the GDP data similarly I can also just add the instead of GDP I can add my return variable here and see the distribution of return as well I will just type here return I will again enlarge my plot window so the plot is appearing nicely and again I will not describe the box plot properties again as we have done it already but you notice the distribution of return across different countries is very different very very different. So, this gives some idea that there is a considerable amount of heterogeneity across countries and this gives us some intuition although we will see empirically as well this gives us some intuition that if we model them simply by pooling all the observations together the results may not be accurate. Now for your report writing another purpose is you can use this export command to save the data as an image you can change the format to PNG, JPEG you can change the width and height or you can do it by dragging through the mouse cursor here you can drag the chart get it in the right shape or also you can use this export command to save it as PDF or copy to clipboard. For a digital examination, you can zoom in as well so that a facility is also available to zoom it and have a good look at the plot. To summarize in this video, we saw how to visualize the data the relationship and evolution

of variable return and GDP variables and we also saw their distribution through the box plot.

In this video, we will perform ordinary least square regression by pooling the data or rather pooled OLS regression. So, we will conduct the pooled OLS fitting or regression of the data and we will see what are the issues with it and how it can be misleading while working with this kind of panel data. So, to perform this we will create an OLS variable that will store the output of the linear model LM for linear model regression. Return is our dependent variable GDP is our independent variable and the data that is used is panel data. Now the output is stored here so if I summarize this OLS variable as we will see you will get the output and in this output, as you focus you will find that the GDP coefficient appears as insignificant which is very counterintuitive because generally, you would expect the GDP to have a significant role to play in the broad market wise index returns.

However here it seems that GDP is not very effective. First and foremost, let us see the problems with this. So, I will use this plot OLS command to see some basic diagnostics. So, I will run this plot OLS and after that, we will see a more comprehensive analysis but first, we will plot this OLS and it gives us into a center. So, the first plot it sort of plots residuals versus fitted values and you can see the very poor fit and also the variance of residuals is not very same across observations so it is a kind of indication that there is a lot of heteroscedasticity in the data and also the fit is very poor.

Also, you can see the QQ plot although the QQ plot suggests that many of the observations in the QQ plot. you compare the two distributions. For example, here we are comparing the distribution of standardized residuals with theoretically normal quantize or normal distribution as you can see here at some places there is deviation. So, for example, it suggests that residuals are not exactly normally distributed and if you remember one of the important assumptions in regression OLS regression of this kind is that residuals are normal distribution, normally distributed. The higher the deviation from normality, the poorer the regression model estimation. Also, you can press enter and you can check the scale location plot. The scale location plot suggests the indicate you can see some of the outliers you can identify the outliers for example 44, 31, 51 these are some of the outliers that may affect your estimation and vitiate your estimation. So, if you want to make a proper estimation you probably would do well by removing these observations. Although to highlight that in this case the magnitude of this there is not much in this thing. So, for example, this plot also gives us some idea about homoscedasticity or heteroscedasticity rather than heteroscedasticity of the data because error variance is not the same. So, you can notice the distribution of residuals is not the same across the fitted value.

So, this is another way to have a look at the sort of heteroscedasticity in the data. The last plot is very important because it gives us an idea of which observations are truly influential. Well, there may be some extreme outliers as we could identify in some of the previous plots. But here you can directly see the influential observations and although it appears that observations number 22, 23 and so on 11 are slightly more influential but the nature of this plot is we do not see the cook's distances. So generally custom read this plot you will find the lines corresponding to cooks' distance of 1 and 0. 5 and those points that are beyond those lines would be slightly extreme. In this plot, I do not see those cook's distances that are 1 and 5 so probably the cook's distances are rather small. Let me have another way to examine the cook's distance probably to get more idea why we do not see those two lines. So let me summarize the cook distance for this particular model as we can see it or the maximum cook distance itself is like the order of 0.1 so that is why probably we could not see the cook distance of 1 and 0.5. As a practice whenever you see those lines those lines will be shown here in a sort of dotted form like this and when you see those lines of 1 and 0.5 you gather that those are points that are beyond those lines that are extremely influential and affecting your regression. The last step in this video will be, to plot a GG plot and see how the average aggregate plot appears as compared to the other individual country plot. So, we will as a first step data aesthetic command X equal to GDP, so we want on the X axis we want GDP on the Y axis we want returns. We will group our observations according to country so for that this grouping according to country and we will also color the observations the individual lines according to country so that we get a nice color this is the first layer of our GG plot the second layer is the point so we will also plot the points through this geom point command and the next the third layer to this is geom smooth so this geom smooth will ensure that we have a line so we will get the line and I am specifying the method as LM so this method equal to LM will ensure that the line is straight and I will also put a standard error as false so this S equal to false a false argument ensure that there is no confidence band that slightly vitiates the graph creates hazy graph which is difficult to visualize. So now we have added the individual country lines and now we will add this ab line command and I will add the OLS model here the aggregate line so, how to do it I have already fitted the aggregate OLS I will make use of its first its intercept which is this first and then also its slope which is equal to this so this is my slope 2 and I will put a line width of maybe 2 and color of let so it comes out sharply so this is my remaining next step. Now while the plot will be appear I can show you a brief glimpse of this plot but in order to make it as technically more appealing I need to do some addition to it so so let me run this plot let me see if it is working and we will just have a glimpse of this plot but I want to add some aesthetics so the plot is there but still I need to add some from aesthetic purposes I need to add some more layers as you will see so first and foremost I will add the this labs command and in the labs command I will give the main title, title of the plot which is hetero-genity cross country so this is my central title and my y axis I have my y axis as

returns so the next layer I would like to add is the theme layer and probably I will pick up from the previous set of code and not rewrite the code in the interest of time I will just pick up from here and same set of probably I will do with the same set of arguments so I will not change them much let me run the full command and you will see a nice plot appearing here so I will just enlarge my plot window and now it is there so in this plot you can very nicely and clearly see the individual lines for countries A, B, C, D and their scatter points also which shows that there is lot of hetero-genity for example the green one country C the green one is on the upper end while the lowest end is the D country the dark green is on the lower D and then you have some countries in between and also the relationship between GDP and returns are not homogeneous across countries there is lot of hetero-genity in the relationship and the average fit if I would have done the average fit it would have come as red line so obviously average fit would have indicated poor sort of vitiated because of this individual hetero-genity across countries so probably it is not the right way to estimate the relationship. So in this video, we examined the OLS fitting or the pooled OLS fitting of the panel data we found not only visually but empirically as well that due to unobserved heterogeneity or individual heterogeneity across countries the relationship estimation is vitiated and probably not a very good fit and accurate estimates not very accurate estimates are obtained so we need to do probably we need to make use of panel data methods that we have discussed earlier. In this video we will start with a very simple version of the panel data model which is the least square LSDV least square dummy variable method we have already discussed the theoretical aspect of it to implement this method the least square dummy variable method let us call it LSDV model. In the LSDV model all we will do is simply use this LSDV let's store this object here LSDV and we will use the linear modeling not much change there so we will regress our return variable on GDP plus now the only addition in the simple OLS we will do is we will add the factor as country and then I will add the data as panel and now when I summarize this model notice something interesting will happen when I summarize this model R will automatically add the dummy variables if you recall we need n minus 1 dummies so the country A R chooses the dummy alphabetically so the country A is loaded on this intercept variable and we can see now in the output we will be focusing on the coefficient of the GDP variable which has now become significant at 90 percent confidence so it has become significant at a significance level of 10 percent or what we call as confidence level of 90 percent which shows that when we explicitly accounted for the individual heterogeneous country specific heterogeneous then our model shows some clarity or some clarity has emerged with this GDP variable being significant at 10 percent.

So, with this, we find that this LSDV method of panel data estimation gives us some intuition that probably these individual heterogeneous were indeed creating some vitiating effect on our simple pooled OLS estimation and therefore it is more desirable to look at panel data methods to account for the unobserved heterogeneous. In this video,

we will continue with our examination of the panel data relationship. First, we will declare our data formally as panel data with our environment and second, we will see whether these fixed effects are significant or not. So, we will start with the examination of panel versus pooled. First, let us explicitly tell R that our data is a panel data set, and for that, we will use this pdata frame command in the PLM package.

We will tell that our data panel index is equal to C country. So, first is the country's individual index and then time. So now we have told let us see and whether your data is converted into panel data or not you can run a head command and you can see here now there is a specific dimension that is added it is not a column it is just a property of your data itself like a timestamp you have a time and individual dimension. You can also add a summary command to see the more clearly panel set. So, you can see your data has been summarized properly across the country and your dimensions. Now let us model the fixed simple fixed effect model. Now we will model fixed effect with individual heterogeneity. So only individual heterogeneity is accounted for in this fixed effect we will use PLM command return GDP tilde so very similar format data will describe as panel.set and model within by default runs individual effect. So, by default, it runs individual effects if you do not specify any effect. If I run the summary of this fixed. individual notice in the output that the GDP coefficient is very similar to what we got with the LSDV method, and its significance is also similar at 10 percent which is 90 percent confidence which is very similar to what we got with LSDV. So that means the fixed effect has accounted for the heterogeneity and results are similar to LSDV. There is a very simple and interesting way to perform pooled OLS also within the panel data framework PLM package I can write pool and it will give me the let me call it pool and this pool model will be very nicely run with PLM rather than earlier we were using LM but now we are using PLM, and it will be very nicely run. Now all I need we can see the output again with the pool as we have already seen the results are not the same and that we have already seen but more importantly with this pool object in the PLM environment we can compare we can perform a poolability test in the PLM package for fixed effects. So, we have a test of iron pool fixed and fixed. individual effects so we can see the individual effects are significant we reject the null hypothesis this sort of F test says that in both the models are there is no incremental benefit of adding these fixed effects but here we can see that we reject the null and we say that indeed there are significant fixed effects by rejecting the null. So, there is another way to conduct this test which is the PF test which will also help us so we will again type fixed pool and we fixed. LM here this should be fixed.individual and again we get the significant effects we reject the null hypothesis of no individual effects indeed there are individual effects we also so there is also something called PW test for the same and we can perform PW test of significant effects and for all these tests we get the significant effects unobserved individual effects. Now as of now, we have tested for individual effects but we can also test for time effects for example the way to do it very similar set of commands we will

apply for time effects how to do that so we will run the command fixed.TM so I will use since we are using time effects I will use fixed.TM and here all I need to do is put effect as time effect equal to time and when I run it and I can check the summary. So this model probably is not so good we get PR equal to insignificant value is very high so it is not significant as we can see so which means time effects are not very significant in fact you can further cross-check it with the same set of tests here instead of this I will use TM so we will check no so I can run all sort of test on TM so it is not here as well so I can see that time effects are not significant.

Now that we have tested with time effects let us test with two ways both time and individual effects. So when I run the two ways effects it is also not very difficult I can run here instead of the time I can put two ways sorry so I need to put it I will put it TW so let us give it a name TW and let us see what is the output so when I run with two ways I can see the output you can see again so GDP is not turning to be significant so that is slightly problematic which means time somehow individual effects all the best-fit model. Let us compare it with the pool model so we see that with the pool model, it says that significant effects in both cases we are able to reject the null and give these p values which means we are saying that effects are significant, but we have already seen that time effects are not significant so the significant is coming from the individual effects. There is also a very interesting way to simply test this with PLM test so I can simply use PLM test and it is a very nicely conducted PLM test and testing for individual testing for fixed effects. So I can simply type here pool and then you can test for the effect which you want for example you can specify the individual and you can run it yep they are significant the version of the test here is Honda you can change the test you can type question mark PLM test and you can select the relevant test here, for example, there are multiple tests available Honda BPGHM and so on so you can test multiple tests as well there is nothing wrong with that so individual I can test for time let us see what is the result for time not significant So now you know that when I run two ways without even running it we should know that it may be significant but the significance is derived from the individual effects two ways are a sort of combination of time and individual because we are getting significant effects they most probably they are coming from individual effects and time effects are not so significant.

To summarize this video, we examined the fixed effects for individual time and both individual and time two ways effects we found that there are indeed significant effects, and these significant effects are coming from the individual unobserved heterogeneity aspect of the data. In this video we will examine how to extract the fixed effect that we have modeled so we will examine how to extract the fixed effects both time two ways and individually. So let us start by extracting with this fixf command fixf we will use this we have already in the fix. individual effects and we will specify that we want to extract

the type level. So, with this level command, we will get the fixed effects at level something interesting we have already extracted the LSDV model.

I will show you that we have already done the LSDV model, and I will show you that there is a relationship between LSDV dummies and the levels of fixed effect that we have extracted here you can see that the intercept is loading the variable as you can see here. Similarly, I can add up this fixed effect at level by adding this intercept to this fixed effect which is at b to get this number. So basically, at every individual country, it is a difference between the intercept and the level. So, for example, at c the effect is nothing but if I sum this delta c 0.91 with 0.29. I will get this number 1.2. So, these are the differences. So, these are the differences with intercept also R allows you the flexibility to extract fixed effect at first difference. So, there they are in the form of the first difference that you can obtain you can see they are exactly sort of identical. So, a is not produced which is loaded on the intercept and other fixed effects are sort of identical here. In addition to the first difference, you can also add the mean difference of the mean and there you can get the mean number average subtracted from the averages of fixed effect. This is how you extract the individual fixed effects. Now if I want to extract the time-fixed effect again the command is pretty simple. Also, there is one more thing you can check the significance of this fixed effect by summary command I can put these fixed effects and inside the summary command to get you can see we get the fixed effect as well as their summary number the significance. Similarly, I can get for time so all I need to do is supply the variable tm and I can get the fixed effect also I can get the summary number here. So here we are extracting the time-fixed effects. So that also we can obtain as you can see here, we can see their significance and other things all I need to do is revise the command a little bit and I can give you the summary and their significance also we can obtain.

In a very similar manner, I can also obtain the two-way fixed effects. All I need to do is just specify that these are two-way fixed effects, and we can find the two-way fixed effects we can also summarize them, so it is pretty easy. I can just summarize the two-way fixed effects. All I need to do is put tw. All I need to do is this tw and summarize and we can get the two-way fixed effect. I need to specify that the effect is two-way so I will specify the effect as two-way. So here I will get the two-way effects. You can see here the combined effects two-ways or rather I will put the effect here as two-ways so I will get the two-way effect here so we can obtain the two-way effects also. so it is pretty easy. So in this video, we saw how to extract the fixed effects for both individual time and two-way model and we also saw the individual effects were pretty similar to the LSDV model least square dummy variable model that we run earlier which is also intuitively thinking about the model which should have been there as well.

In this video, we will learn how to implement the first difference model. First difference

panel data model we have already discussed the theory so I will not rewrite the code I will simply borrow from the previous code. All I need to do is I will give it the name fd fd for the first difference and all I need to do is I need to specify instead of within I will use fd and the same command can be used so I will run the first difference model. Let us summarize this and examine the output of this first difference model. Let us see how it appears. So notice the output is pretty similar to what in terms of its coefficient of GDP and its significance it is very similar to the LSDV as well as the fixed effect model.

So we can see the significance is around 10 percent which means a 90 percent confidence level and a coefficient that is also very close to we had around close to 0.039 or something with the fixed effect and the LSDV model it is very similar. So, we get a very similar result from the fd fixed effect and LSDV models. This is very intuitive and similar to what was discussed in the theoretical part. In the next video, we will discuss the random effects model. In this video, we will see how to implement a random effect model. Again, I will not write the full code, I will just borrow it from the fixed effect model, and we will implement the random effects. We already discussed the theoretical part. The code is pretty simple.

Let us name it random and I specify here the random. Now let me state at the beginning that this is individual effects because we have not specified any particular effects, so it is individual effects summary random. So, I will just run it. Let me run the model. So now interestingly in this model, there is one additional aspect in the output, and you can choose the effect particular model that is run. For example, as of now, summary random is the default version but there are different models available to check.

Notice the effects because this is individual you can see only individual effects it is pretty high as compared to the idiosyncratic part of the variation the individual effect is pretty high around 62 percent. So, it indeed seems to be important. The lambda we discussed in the theoretical discussion is the same as theta here in that means almost it is closer to 1 which means it is the random model is closer to the fixed effect rather than to the band or spectrum of effects. It is closer to fixed effect rather than pool which is also confirmed by this theta parameter which is the same as lambda in our discussion notice the GDP coefficient which is also very close in terms of significance and its effect is very close to what we had with fixed effect and LSDV and FD.

So, it gives us some idea about the consistency of our results. Now in a very similar way, we can include time-fixed effects for example if I want to let us name it an individual random individual rather than make it make a distinction, I will name it a random individual. So, this is the random individual model that we have run and then I will run with time effects. Let us see how it appears. So instead of so I need to add one more parameter effect as time and here I will specify as pm. So, it will give me random time effects. Let us see what the nature of output is. Now here in the previous part

notice the time effects are explicitly loaded and it seems when I load the time effects time effects are not significant and even theta parameter is 0 which says that our random model is closer towards pooled OLS. This is due to the fact that this time parameter is not contributing much to the explanatory power of the model and the model feels that better would have been if you had used pooled. This was expected already because we have already seen in the fixed effect results that time effects are not contributing much. Let us see what about two ways. So, we will use the two ways effect, two ways and it is very easy to implement. I can write these two ways commands here and let us call it tw, random tw, and interesting output will be generated. As you can see now another dimension apart from idiosyncratic you have individual and time and as we expected the individual dimension is pretty large in terms of contribution to variation around 63 percent. Theta is also important when we have individual effects. It is getting diluted with time and two ways. So that means time indeed in fact time and time effect is diluting our overall effect part and therefore it is better to keep only individual effects where theta parameter is pretty high and closer to the fixed rather than pooled.

Again here because the time effects are vitiating our model has slightly lost its power. So, as a model summary, we will stick to our individual effect. Now if you recall whether to use random or fixed, we relied on our Hausman test. Hausmann test and the way to perform the Hausman test, here we have already had a great detailed theoretical discussion about the Hausman test. So here we will only talk about implementation which is very simple in ours. So, I used pH test for Hausman. If you want to know more about what is are the parameters and arguments, you can simply type the question mark pH test and you will see a lot of details appearing here. So, a lot of details you can test here but for simplicity all I need to do is supply fixed effect and supply fixed dot individual and also random because both of these models are good random and fixed individual models. So I will run them and I am not able to reject the hypothesis. Now this is slightly interesting. The model suggests that we are not able to reject the null that both of the models are consistent. When both the models are consistent you go ahead with the random model and in this case, it seems random with individual effects is a good model. So, we go ahead with the random model but if you recall random model had a parameter of theta equal to 0.76 which was indeed very close to a fixed effect. So, whether we are using random individuals or fixed individuals broadly we are getting similar results which is good for experiment that confirms or rather in research we use something called the weight of evidence approach. So, we try different models and wherever the evidence is pointing out more we tend to choose that. So let us say if you have 10 models and 8 models pointing towards a particular result then probably that result is more accurate. And in this case, both fixed random and indeed LSDV and first difference models are pointing towards the coefficient of GDP in this relationship which is closer to 0.003 or 0.0035 and a significant level which is around 0.01, sorry 0.10 which

is 10 percent. Let me show you the outputs in fact. So, if I summarize the outputs you will recall that are fixed individuals this 10 percent significance level and 0.0039. Lastly, if I summarize random, so I get around similar results I summarize LSDV, I also get a very similar result and lastly if I summarize FD first difference then also, I get, so it is a very resound kind of very solid result we can say that almost identical for all the models around 0.004 or 0.4 percent coefficient. So, it is a very sounding and we should be happy about this that we are getting a very consistent sort of result across different models. So, to conclude we saw that results while the overall Hausman test suggests that it should be the random model which is more accurate but looking at the output we find that all the results are consistent pointing towards a GDP coefficient of 0.004 percent and a significance level of 10 percent that means a confidence level of 90 percent. Also, we found from the theta parameter or what we also call as lambda parameter the model is the random effect model is closer to the fixed effect rather than pooled OLS on the spectrum of effects. In this video, we will examine problems of serial correlation, cross-sectional correlation as well as hetero plasticity in error terms. So, we start with the serial examination of serial correlation error terms, and serial correlation error terms and this will also help us decide between fixed effect and for different model. For example, to start with you can try the simple PWFD test of serial correlation in FD model. So, if I supply the model which is the FD model it tells me that FD model errors are indeed serial correlated which suggests that the FD model is probably not so good model if errors from the FD model are serially correlated then it is better to go with FE fixed effect model but still let us test if errors from FE model are serially correlated or not as expected they are not. The other way to test this is called the PWAR test where you can supply your fixed effect model and fixed effect model errors are not serial correlated. So overall these results suggest that between FD and FE as we discussed in the theoretical part also because the errors of the FD model are serially correlated it is better to go ahead with the fixed effect model.

Next let us look at the cross-sectional dependence in error terms, cross sectional dependence. So, if there is any cross-sectional dependence it will be highlighted by the cross-sectional correlation in error term. The way to perform this test is also quite simple. So let us start with our fixed effect model. There is something called the PCB test and we will supply our fixed dot end effect model which is the best fit model, and we will use the test as LM so we will use the LM test. Let us start with the LM test. There seems we reject the null and we find that indeed there is some cross-sectional dependence in the fixed effect model. There is another apart from LM we also have a CD test so patient CD test for earlier we had Bruesh pagan LM test now we can also perform the CD test but the CD test sort of rejects the cross-sectional dependence, so we have a mixed sort of result. We can also perform this on the random model so for example we can also perform this test on the random model let us say there is a random model as per LM there is some correlation let us perform the same on the CD test as well CD test says that there

is no correlation. So, here we get a sort of mixed result in terms of cross-sectional dependence. However, in our previous examination, we have already well settled that the model to go ahead is all the models are giving similar results, but the two best models are the fixed effect with individual effects and the random effect model with individual effects. The last test that we wanted to conduct on our error was about heteroscedasticity, so we wanted to conduct heteroscedasticity whether our errors are heteroscedastic or homoscedastic. So let us see very simple BP test is available that would require a library LM test if we have not loaded, we need to load the LM test library and in the LM test library we will perform this BP test and we have written GDP as a variable and let us add factory.

Let us add the effects so effects are country-specific effects we will apply and after applying for country-specific effects let us see how our errors behave. In our details panel, we will put student equal to f and our BP test suggests that significant effects hetero drastic effects are there, so it seems hetero drastic is there in our error after accounting for these effects explicitly. So how to deal with this? So, to summarize in this video we saw that there is some evidence of cross-sectional dependence and also hetero drastic. So how do I correct for these correlation error terms and heteroscedasticity in my coefficients and correct the coefficient matrix for proper hypothesis testing? So, in the next video, we will see robust standard errors that will be employed to compute appropriate hypothesis testing, t-test, and statistical influence.

In the previous video, we conducted residual diagnostics and we found that our error terms were affected by problems of correlation, cross-sectional correlation as well as heteroscedasticity. So, in this video, we will examine robust standard errors a very interesting and ingenious way to deal with these error problems. So, this would require a library LM test and what we will do is let us start with a simple co-if test function. We will supply our random model first let us see the random individual model and the error we will choose is VCOV, VCOVHC. So, this is the corrected matrix for a random model which is corrected for heteroscedasticity-related correction. You can also be in the same model, and you can see this is the correction for hetero drasticity. Similarly, you can correct the fixed effect model as well. Let us correct the fixed effect model results. So, coefficients will not change only the significance and t-values because these robust standard errors will collect for standard error. So, and t-value estimate is divided by the standard error so the t-value that you get is corrected because of the standard adjustments to standard error, and accordingly, the probability value of the p-value will also be adjusted.

So let us do the same correction for fixed effect. This is so there is a slight decrease in the power of the test as you can see because probably we were getting some more favorable results because of heteroscedasticity and so on. So that is now corrected.

Similarly, we can run the pool thing as well. Let us see. Pool. So the pool is all the pool is not so meaningful in this case we are working with panel data. There are some more advanced versions of robust standard errors. Let's start one by one. Let's focus on fixed effects and similarly, you can apply random, but we'll start with fixed.

So, let's say we have a fixed individual model. Then we provide a slightly more advanced version of this variance covariance error term. So, for example, VCOVHC is our variance-covariance heteroscedasticity consistent matrix but here the method that we will use there are different methods available. For example, you can choose white-to-white correction, and you can specify the type of the kind of type that you want to use is HC3 and you will get a slightly different modified result.

So here I can run this. Let's run this. You see we get the result. Yes and now with this adjustment significant level has improved a lot. Similarly, there are different versions for this. For example, you have apart from this you also have R-Lano method.

You can apply and then there are different versions. For example, instead of HC3, I can apply HC0 or maybe HC1. So different ways to correct for this thing. Similarly, you can change for R-Lano also.

I can change this to is white method I can change for R-Lano also. Let's see if there exists HC0 for R-Lano. Yes, it exists. We can check that. Also, if you want to check what are the options you can run this question mark and you will get in the help window you can see what options are available to you. So, there are R-Lano white ones so the options available to you are R-Lano white one by two and there are different types of HC0. If you want to read more about these, you can just have a look at what exactly these terms mean to you.

You can read in more detail about all the literature. But the point that I wanted to highlight here there is also a way you can cluster. For example, you can cluster across groups or time. For example, we realized in our case time is not such a major problem but clustering across groups is a problem.

So, the way we can adjust for it I will not rewrite the code. I will make a small adjustment. So what we can do here is we can specify the method as R-Lano maybe type as HC3. You can choose a different type but then you can do clustering and you can change the clustering methods. For example, you can cluster and you can cluster across groups. So because in our case we realized cross-sectional dependence was a problem so you can cluster across groups but also although time was not a problem but you can also cluster across time as well. So that also you can do. And now another thing you can change these methods. For example, instead of R-Lano, you can use maybe white1 and so white so we have in white we have white1 white2 so we will use white1 maybe HC0 and you can read all the interpretation about these white methods and so on in the

description help menu. So we can see the coefficient significance. Broadly our results are pointing to the same coefficient. We have a number of permutation combinations so broadly our results point to a coefficient value closer to 0.4% with a significance level that ranges within 10% or better. So around 10% and in some models it has improved as well. So broadly we get consistent results. So to summarize in this video we examined the robust standard errors and how to implement them with R. We saw very different versions of robust standard errors and we also saw how to cluster them across time and groups.

So that is also what we saw. In this lesson, we examined the relationship between GDP and broad market-wide index data. For example, NIFTY50. Since the data has multi-country dimensions it carries unobserved heterogeneity. We started the estimation with a pooled OLS approach which offered counterintuitive results. That is there is no relationship between GDP and broad market-wide indices such as NIFTY50 and S&P500. This leads to suspicion that unobserved heterogeneity may be affecting our estimation. We start our panel approach with the least square dummy variable method. The results from the least square dummy variable method show that there is a relationship between GDP and market-wide returns. We corroborate these findings with the fixed effects method. We find that indeed individual effects are significant but not the time effects. We also extract these effects and find that the effects observed with the least square dummy variable approach and fixed effects approach are similar.

Next, we employ the first difference approach, and results similar to that from the fixed effects approach are obtained. Lastly, we perform the Hausman test which suggests the application of random effect is more appropriate. The random effect approach also shows that the individual effects are more significant. The results from the random effects approach are similar to those obtained from the fixed effects method. Lastly, we conduct error diagnostics of the model. We correct issues related to errors such as heteroscedasticity and serial correlation by using robust standard errors. Thank you.