Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology – Madras

Lecture – 50
Dynamic Panel Data Model – Part 13

We are discussing about a specific example from the UK context. Where we are taking 140 firms' data to understand what are the factors that determine the firms employment and we assumed that the ith firm's employment in a particular year if it is denoted by y_{it} then that depends on its own lag value along with other factors.

Like what was the prevailing wage rate, labour, capital, and the aggregate output which was used as a proxy for the demand. we said that this is a dynamic model. what is the reason? because hiring and firing is costly for the firm, so what the firm will employ in a particular year that depends very much on its previous value that means how much the firm have already employed in the previous period.

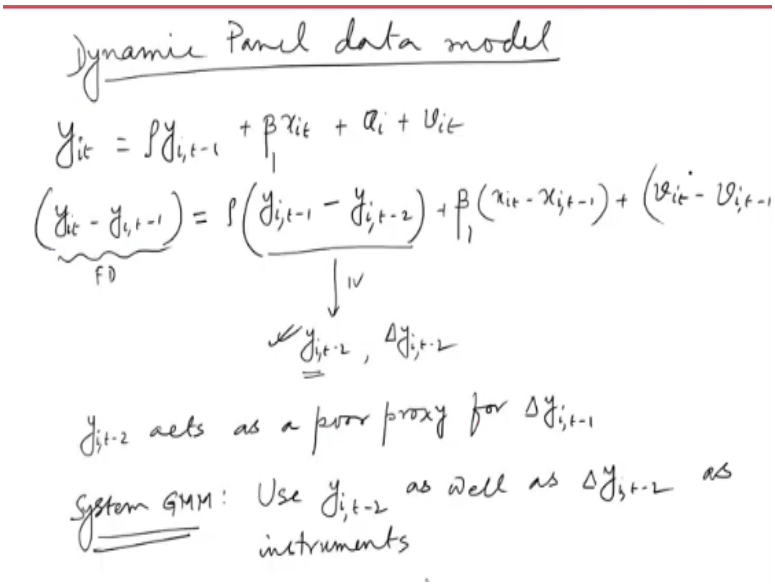
And if you recall, previously we learned basically how to conduct the post estimation checkup for this dynamic panel data model. We said that once we estimate any dynamic panel data model then basically the next step would be to check whether my estimates are reliable. How do you know that just by conducting two post-estimation checkups, one is the presence of higher order autocorrelation or not.

Because the construction itself says that autocorrelation of order 1 must be there, if it is not there then only there is a problem because it indicates the presence of the lag dependent variable is redundant and secondly since in this generalized method of moments which is the major technique to use to estimate the dynamic panel data model it involves so many instruments, so we need to check whether all those instruments are basically satisfying over identifying restrictions.

So, we learned only how to check for the first estimation checkup that is presence of higher order autocorrelation, but we have not said anything about the overidentifying restriction null, probably will talk about that overidentifying restrictions how to check later on. Before we come back to today's discussion I will quickly go through one more important concept that we learn

which was basically the FOD transformation that means followed orthogonal transformation, forward orthogonal deviation basically.

(Refer Slide Time: 03:21)



If we write this is let us say the dynamic panel data model

$$y_{it} = \rho y_{i,t-1} + \beta_1 x_{it} + a_i + v_{it}$$

then we said that the estimation technique basically requires

$$(y_{it} - y_{i,t-1}) = \rho(y_{i,t-1} - y_{i,t-2}) + \beta_1(x_{it} - x_{i,t-1}) + (v_{it} - v_{i,t-1})$$

this is called first difference transformation.

And then we said that this first difference transformation it has a major drawback, what is the drawback? If you are working with a panel which is unbalanced in nature, then some of your observations will be missing. So if y_{it} is missing for any particular form then both delta y_{it} and delta $y_{i,t+1}$ would be missing. So that means the first difference transformation basically will magnify the gaps in an unbalanced panel and that will lead to huge loss of observations, then we say that what is the solution?

Solution was the forward orthogonal deviation, instead of deducting the previous value from the contemporaneous one what we actually do we can subtract the mean value of all the future available information from y_{it} and in that in that case even if some of your observations are missing, we can always compute the average. So this average value is basically available for all the periods except the last one and in that way you can minimize the loss of observations.

Interestingly we have showed yesterday that if you use that a 4-D transformation in unbalanced panel then by reducing or minimizing the observations it can actually help improving the quality of the estimates. How? we showed that if we use a 4-D transformation then our estimates that means coefficient of $y_{i,t-1}$ it goes in that range that means within the upper and lower bound of the estimate set by the FE and OLS estimations.

We also learned how to use xtabond2 instead of xtabond to estimate a dynamic panel data model. basically we discussed about the difference GMM. Here we were using two IV's. Two IV's are suggested one is $y_{i,t-2}$ and next one is $\Delta y_{i,t-2}$. If you use $y_{i,t-1}$ then it is called difference GMM, if you use $y_{i,t-2}$ as well as $\Delta y_{i,t-2}$ then it is called a system GMM.

We will quickly estimate a dynamic panel data model using this $y_{i,t-2}$ as this instrument that means a difference GMM and we will quickly see the property of that estimate and we will apply then system GMM delta $\Delta y_{i,t-2}$ and then we will see is there any improvement in the quality of the estimates. So, we will once again use the same data set. **(Video Starts: 07:53)** This is Arellano and Bond's original data set.

The command that we are going to use is xtabond2 then my dependent variable n and then all my independent variables which is lag of this nL1 and nL2, then wage rate, then capital and capital of 2 kL1 and kL2 and then we will be using ys, then ysL1 and ysL2. This is how we have and then we have year dummies. This is how we have specified our model and then we need to specify what is our endogenous variable and we were using only one endogenous variable that is nL1.

Then we need to specify what is our exogenous variables, exogenous means those variables will be used as their own IV. Then we have w, wL1, k, kL1, kL2, ys, ys1, ys2 and we have this is also yr star as our IV style variable. Since it is a difference GMM we do not need the level equation no level and we need robust standard error and we need small sample correction.

This is how we can specify a dynamic panel data model in its difference, this is basically a specification of difference GMM, is no level equation. We need to put no level equation. This is the dynamic panel data model estimation. what we are having here? We are having the estimated difference GMM.

We have estimated difference GMM using 41 instruments, but the problem here you see the coefficients is coming out to be 0. 25, so much lower than the theoretically set lower limit given by your fixed effect transformation. Now how do you improve on that? One solution is basically to estimate a system GMM. **(Video Ends: 13:12)** That means in terms of our model what we are discussing here we are using $y_{i,t-2}$ as the instrument for a variable which is basically in difference form.

It means $y_{i,t-2}$ acts as a poor proxy for delta $\Delta y_{i,t-2}$. Since the variable is in difference form, we must use the $\Delta y_{i,t-2}$ as well as instrument. So this is a poor proxy and the problem as we discussed earlier the severity of the problem increases when the variables are in other explanatory variables they follow random work. So solution is the system GMM that means use $y_{i,t-2}$ as well as $\Delta y_{i,t-2}$ as instrument.

What we are doing here? We are considering both the level equation as well as the difference equation as a system of equations. Then $y_{i,t-2}$ lag of the level is used as instrument for the difference equation like earlier and $\Delta y_{i,t-2}$ is used as instrument for the level equation. This way when we use a system GMM, then we will see whether the quality of the estimated coefficients is improving or not. How do you check quality?

First of all, it should lie within that interval. Let us see now. **(Video Starts: 15:54)** So here we will put the same command but instead of no level equation I will remove that. So, if I remove no level equation that means I am asking state to estimate a system GMM. I will only add one thing here, so let me estimate this model first. This is let me see what we have estimated? This is dynamic panel data one step system GMM model we have estimated and by doing so what is? It is my coefficient 0. 78.

It means the coefficient is lying within the interval. System GMM basically is an improvement over the difference GMM because of the type of instruments what we are using. While difference GMM could not guarantee our estimates to lie within the interval, system GMM actually guaranteed that yes, our estimates are lying within the interval. With xtabond2 we can easily estimate two types of dynamic panel data model.

One is difference GMM another one is system GMM and as we can see that we are not only getting the estimates but also, we are getting whether my post estimation checkups are also satisfied or not. What is my AR1 test? Look at this AR1 P value is 0.000 which means that I am able to reject my null hypothesis. What is my is null? Null is there is no autocorrelation, there is no autocorrelation which is rejected.

It means first order autocorrelation AR1 is present in this particular model. But here what is happening? The second order autocorrelation AR2 is not there, those two tests are also getting. These are in-built within this xtabond2 command. Now throughout this approach, throughout this discussion if you look at what type of estimates we are getting in a dynamic panel later model? **(Video Ends: 18:47)**

Please keep one thing in mind that so far we have assumed that when you are modeling employment of a particular year $y_{i,t}$ which is again a function of $y_{i,t-1}$ that will lead to endogeneity right that we have, explained earlier how inclusion of $y_{i,t-1}$ in the model it leads to endogeneity. But we have assumed that there is only one endogenous variable and that is the lag dependent variable $y_{i,t-1}$.

And we assumed that wage, capital and other factors, forget about other factors for the time being, the question is should we consider wage and capital also as exogenous variable? So that means the question that we are raising over here are wages and capital really exogenous that is the question that we are asking. Now why do you think that wages and capital can also be indigenous? Look at our dependent variable.

Our dependent variable is $y_{i,t}$ which is employment, $y_{i,t}$ is basically employment and it is actually a function of I am saying let us say wage and capital and so far we have assumed that only wage will lead to $y_{i,t}$, capital we will also lead to $y_{i,t}$, but there is no reverse causality running from $y_{i,t}$ to wage and capital and what happens if $y_{i,t}$ also determines wage and capital? Now if we think logically what amount of wage that we are including, what amount of employment that we have $y_{i,t}$.

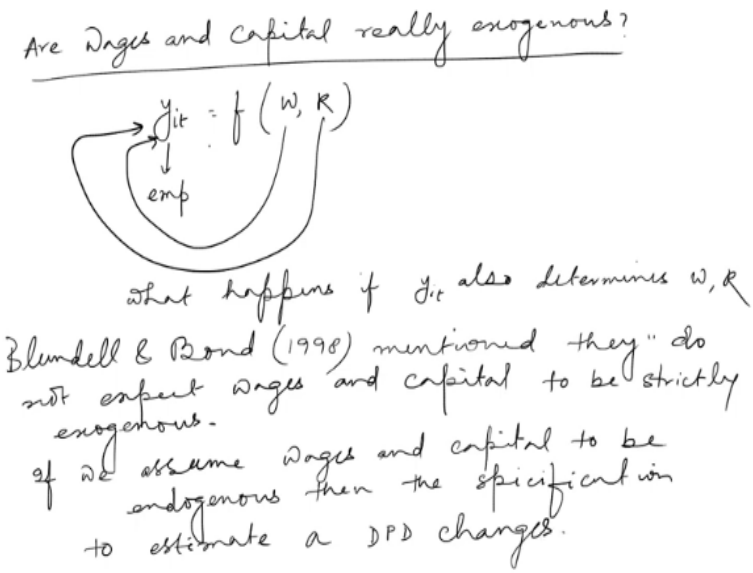
That may also determine what amount of capital should be employed by the firm. Because the production depends on a particular combination of labour and capital. So if we assume capital will determine employment, employment may also determine what should be the capital amount

because of their joint determination. So that means we can understand easily that there might be cases where y it also leading to the determination of capital.

Similarly, $y_{i,t}$ the employment in a particular year will it determine wages, it means what would be the wage that would be prevailing in the market that also depends on the level of employment. It means level of employment may also determine wage, so this relationship could be very well simultaneous in nature and this simultaneity will lead to autocorrelation, so it will lead to again endogeneity. This is endogeneity due to the reverse causality. There might be other reasons for endogeneity as well.

There might be some omitted variable bias and other things, that means even if we do not discuss all those things in detail, we can understand easily that these w: wage rate and k: capital which we have assumed so far as exogenous in the model, these factors could be endogenous. So when you will be working with your data to estimate a dynamic panel data model apart from the lagged dependent variable, you should be carefully looking at other factors if at all there are other endogenous variables also in the model.

(Refer Slide Time: 23:56)



What Blondell and Bond in 1998 mentioned, actually they do not expect wages and capital to be strictly exogenous. If we assume wages and capital to be indigenous then the specification to estimate a DPD changes. **(Video Starts: 26:06)** That means earlier if you look at our command what is the command we have given? Look at this. This is xtabond2 n nL1 all these and then I said within GMM I have given only nL1 lag of the dependent variable I assumed as endogenous variable and I have assumed all other factors as exogenous.

Now the moment w and k also become endogenous, this particular specification does not work. So that means those factors will also come now within this GMM. So how will you write your command then when you assume wages and capital to be endogenous? Look at this, I will put again xtabond2 and then n nLl nL2 wL1 and then I am using precisely first and second lag of capital and industrial output.

This is the compressed way of writing the same command and then I am giving yr* the year specific dummy and then what I am doing this is my specification xtabond2 n nL1 nL2 w wL1 and then first and second lag of capital and output, after that I will put GMM style and within that GMM style what I am including? Lag of n and then w and k and so L dot n w and k so this will be my GMM style instruments and then I will put IV style.

And in IV style I have L and then within bracket again 0/2 ys and yr* year and then what I need let us say I am using same difference GMM no level equation and robust and small. Look at what is happening? So what do we have estimated? One step difference GMM. Now when I am using one step difference GMM, earlier what we saw that difference GMM are inefficient compared to the system 1.

Because difference GMM could not guarantee the estimates to lie within the theoretically determined bound. But once we check the specification properly then even within the difference GMM also see our estimate is now 0. 81 which is lying well within the interval of 0. 74 and 1. 04. So how did this happen? Difference GMM improved the quality of the estimates because we checked for the specification.

Probably earlier what we were assuming that wage rate and capital they are exogenous, the moment we change them to endogenous then our model performed much better than what it was doing earlier. So that is why when we estimate the model one thing we have to keep very carefully in our mind and we have to be very careful about whether we are specifying our endogenous and exogenous variable properly.

If we fail to do so, then that will lead to this type of problem. So just by specifying our model that means just by bringing the endogenous variable from the set of IV style to GMM style earlier we were considering only employment as indigenous variable that means employment

means lag of its employment, but the moment I put that wage and capital also as endogenous variable the quality of my estimates it has improved significantly.

And then we will see that my test for autocorrelation of first order is again rejected null hypothesis that means there is first order autocorrelation, but second order autocorrelation is not there. **(Video Ends: 33:07)**