[Academic Script]

Panal Data Analysis

Subject:	Business Economics
Course:	B. A. (Hons.), 5th Semester, Undergraduate
Paper No. & Title:	Paper – 531 Elective Paper Q1 – Advanced Econometrics
Unit No. & Title:	Unit – 5 Panel Data
Lecture No. & Title:	Lecture – 1 Panal Data Analysis

Academic Script

1. Introduction

Hello friends nice to meet you. Today we are studying what is known as Panel data analysis. Panel data consist of both time as well as space dimension. As for example, literacy ratio, gender statistics in different states of India during certain years. So we have complex data which is really bulk in nature. So we cannot apply our standard model in this case, so we have to consider other types of models, they are known as fixed effect model than random effect model etc. So some of them you will study here in detail, let us now go with the lecture.

We had started our study with three types of data

(i) Time series data, where we observe the values of one or more variables over a period of time (e.g. Industrial production of India during the period 2001to 2016)

(ii) Cross Section data, where we have data for same fixed periods concerning one or more variables (e.g. Population census of India for the years 2001 to 2011 GDP data during the five years plan periods etc.)

(iii) Panel data have both space as well as time dimensions. In panel data we have the same cross sectional unit (family or firm or state) which is surveyed over time.

(e.g. Literacy ratio of different states of India during last 10 years etc.)

Some more illustrations for panel data are as under

(1) Data on total cost of production and total output for 5 leading companies in India producing steel during last 10 years)

(2)Data on labor productivity and wages for different states in India during last 5 years

(3) Income expenditure pattern for different groups of people in a state during last 10 years etc.

Panel data are called Pooled Data, micro panel data, longitudinal data etc.

The regression models based upon panel data are called Panel data regression models.

In economic research such models have substantial applications and hence they are becoming very famous these days. There are many surveys done by government set up which collect panel data.

e.g. In India, Economic Census data, National sample survey (NSS),Centre for Monitoring Indian Economy (CMIE), Annual Survey of Industries (ASI) etc. are a few examples for Panel data.

Broadly speaking, the topic for panel data is vast and techniques (mathematics and statistics) used are quite complicated. There are some user friendly packages such as STATA, SAS, SHAZAM, E views etc. which can help in dealing with the analysis of panel data.

2. **Some Special features of Panel data**

If we observe and highlight panel data for certain organizational set up, we find the advantages and disadvantages.

(A) Advantages

(1) They increasing the sample size considerably.

(2) By studying the repeated cross section observations, panel data are better suited to study the dynamics of change.

(3)Panel data enable us to study more complicated behavioral models.

(4) By combining time series of cross section observations, panel data gives more information on data, more variability, less

colinearity among variables, more degrees of freedom and more efficiency.

(B) Disadvantages

(1) Due to the nature of the panel data, we experience several estimation and inference problems.

(2) Since panel data involve both cross section and time dimensions, there is heteroscadasticity in cross sectional data and autocorrelation in time series data.

(3)Panel data analysis becomes very complex to understand and it also need sufficient care for the implementation.

(Note: (1) Panel data is called balanced panel if each subject (e.g. firm, individual etc.) has the same number of observations

(2) Panel data is called unbalanced panel if each of the entity has different number of observations.)

3. Panel data Regression Models

Very broadly, we can classify the regression models for panel data into 4 categories

(1)Pooled OLS model (constant coefficient Model)

(2) Fixed Effect Least Squares Dummy Variable model (LSDV)

(3) Fixed Effect within group Model

(4)Random Effects Model (REM)

out of these 4 Categories of panel data regression models, we want to study LSDV model and REM.

4. Practical Application

Let us suppose that there are six airline companies in the country. We want to built up the total cost function for these companies. We also assume that we have data for 20 years for each of the airline companies. These data are pertaining to the following factors.

(1)Total cost denoted by variable C

(2) Output in terms of revenue passenger miles (denoted by Q)

(3) Fuel price denoted by P

(4)Load factor which is the average capacity utilization factor of the fleet denoted by L.

We can frame up the format for total cost function for these airlines in functional form as $C = f(Q, P, L)$

Totally we have 120 observations for estimating the above total cost function. We can have a linear or non-linear form for the above cost function. Write a linear function as under

$C_{it} = \beta_1 + \beta_2.Q_{it} + \beta_3.P_{it} + \beta_4.L_{it} + U_{it}$ …….. (1)

Here C_{it} = Total cost for i^{th} airline for period t

Q_{it} = Output for i^{th} airline for period t

P_{it} = Fuel price for i^{th} airline for period t

L_{it} = Load factor for i^{th} airline for period t

U_{it} = Disturbance term for i^{th} airline for period t

$(i = 1,2,3,4,5,6\ , j = 1,2, \ldots 20)$

The above model equation given in (1) is called Constant Coefficient Model. Here β_1 is the intercept which is the same for all airlines assuming that all of them have uniform level of services rendered to their customers.

We can assume that the explanatory variables are non stochastic, and if they are stochastic, they are uncorrelated with the error terms. We may run this regression and find OLSE of Betas, so that cost function can be estimated.

Let us now extend this model further and study Fixed effect model.

2. Fixed Effect Least Squares Dummy Variables model

We rewrite equation (1) somewhat differently as shown below

$C_{it} = \beta_{1i} + \beta_2.Q_{it} + \beta_3.P_{it} + \beta_4.L_{it} + U_{it}$ …….. (2)

$(i = 1,2,3,4,5,6\ , j = 1,2, \dots 20)$

Here only change is the intercept term β_{1i} in place of β_1. This shows that services rendered by different airlines can be different by means of special features of the airlines. The model given in (2) above is called Fixed Effect Model. It is called fixed effect in the sense that intercept term is not dependent upon time. Thus Fixed Effect Model is time invariant. Similarly the slope Coefficients of the regressors do not vary across individual or over time.

Next question is –How can we allow for the fixed effect intercept to vary among the airlines?

This is accomplished by using dummy variables technique. we can write

$\alpha_{1i} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i}$ .....(3)

Since there are 6 airlines, we choose 5 dummy variables to avoid dummy variables trap.

Here $D_{2i} = 1$ for airline 2

$= 0$ otherwise

$D_{3i} = 1$ for airline 3

$= 0$ otherwise

$D_{4i} = 1$ for airline 4

$= 0$ otherwise

$D_{5i} = 1$ for airline 5

$= 0$ otherwise

$D_{6i} = 1$ for airline 6

$= 0$ otherwise

Here we treat airline 1 as the base or reference category (In the similar way we can take any airline as a base.)

Hence on substituting (3) in (1) we get

$C_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i} + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + U_{it}$ ... (4)

Thus $C_{it} = \alpha_1 + \sum_{k=2}^{6} \alpha_k D_{ki} + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + U_{it}$ ..... (5)

This is <u>differential intercept dummy technique.</u>

Equation (4) or (5) above is called LSDV model. Here $\alpha_1$is the intercept value for airline 1, $(\alpha_1 + \alpha_2)$ is the intercept value for airline 2 and so on. Under usual assumptions, we can estimate the model given in (5) above and find OLS estimators. They are called <u>Fixed Effect Estimators.</u>

Note that equation (1) above is also called <u>one way fixed effect model</u>. However if we allow time variable as also effective in the intercept form (e.g. changes occurring due to government policies as per passage of time – thus considering β_{1it} instead of β_1 or β_{1i}) the model is called <u>two way fixed effect model.</u> Also note that fixed effect estimators are always consistent.

<u>Comments on the LSDV model</u>

When we deal with estimating of the model as shown in equation (5) above, there are certain points which should be remembered as a caution.

(1) By introducing too many dummy variables the degrees of freedom reduces and thus we lack sufficient observations for making a meaningful statistical analysis.

(2) By using many dummy variables in the system, (both individual as multiplicative) there is always a possibility for multicollinearity. This makes difficult for precise estimation of the parameters.

(3) LSDV model may not be able to identify the impact of time invariant variables.

(4) Consider the disturbance term. We assume that $U_{it} \sim N(0, \sigma^2)$ Here index i stands for cross section observations and index t

refers to time series observations, assumption for U_{it} needs to be modified to take care of heteroscadasticity and also for autocorrelation.

Random Effects Model (REM)

Let us start with the model equation for fixed Effect Model for our application which is as under

$C_{it} = \beta_{1i} + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + U_{it}$ ..... (1)

$(i = 1,2,3,4,5,6 , j = 1,2, \dots 20)$

Now we assume that β_{1i} is not fixed but it is a random variable with mean value β_1 . Thus the intercept value for an individual company can be expressed as

$\beta_{1i} = \beta_1 + \epsilon_i$ ..... (2)

Where ϵ_i is a random variable with mean value zero and variance σ_ϵ^2.

(i.e. $E(\epsilon_i) = 0, V(\epsilon_i) = \sigma_\epsilon^2$)

From (1) and (2) above, we get

$C_{it} = (\beta_1 + \epsilon_i) + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + U_{it}$

$= \beta_1 + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + (\epsilon_i + U_{it})$

Hence $C_{it} = \beta_1 + \beta_2 . Q_{it} + \beta_3 . P_{it} + \beta_4 . L_{it} + W_{it}$ ..... (4)

Where $W_{it} = \epsilon_i + U_{it}$ ..... (5)

This equation given in (3) above is called Random Effects Model (REM) or Error Components Model (ECM).

Here the composite error term consists of two (or more) components.

The model is given name Error components Model because the composite error term consists of two (or more) error components. Here ϵ_i is the cross section or individual specific error component and U_{it} is the combined time series and Cross section error component.

The usual assumptions made by ECM are that

$\epsilon_i \sim N(0, \sigma_\epsilon^2)$

$U_{it} \sim N(0, \sigma_u^2)$

$E(\epsilon_i \cdot U_{it}) = 0,\ E(\epsilon_i \cdot \epsilon_j) = 0 \quad (i \neq j)$

$E(U_{it} \cdot U_{is}) = 0 \qquad (i \neq j, t \neq s)$

Thus the individual error components are not correlated with each other and are not auto correlated across both cross section and time series units. Also note that W_{it} is not correlated with any of the explanatory variables included in the model.

There is a best named Hausmantest which can decide whether ECM is the appropriate model or not.

From the assumptions given in (5) above, we find that

$E(W_{it}) = 0$ ..... (6)

and $V(W_{it}) = \sigma_\epsilon^2 + \sigma_u^2$ .....(7)

If $\sigma_\epsilon^2 = 0$, ECM is identical to pooled OLS regression model.

In equation (7), the error term is homoscedastic. However it can be shown that W_{it} and W_{is} $(t \neq s)$ are correlated.

In fact $\rho = Corr(W_{it}, W_{is}) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_u^2}$ .....(8)

If we do not take this correlation structure into account, estimating ECM By OLS method gives inefficient estimators, Here the most appropriate method is the method of Generalised Least Squares (GLS),

Note that there are many statistical software packages which can estimate ECM as well as FEM.

We have discussed above in brief (by means of an application) about FEM and ECM. While dealing with Panel data analysis, a question may arise as to which model (i.e. FEM or ECM) is considered to be better.

There are different ways for dealing with this problem by means of the error term patterns concerning the respective models. A

working rule can be given to make a choice between FEM and ECM as shown here under:

(i) If it is assumed that ϵ_i and the $X's$ are uncorrelated, ECM may be appropriate.

(ii) If ϵ_i and $X's$ are correlated, FEM may be appropriate.

There are tests like BPL Multiplier test and Haussman test which can decide about the choice between FEM and ECM.

Note: There are many interesting studies relating to panel data analysis.

This is a current topic for making advanced research in the econometrics field for all higher studies, very complex and complicated tools are to be used which need exhaustive information and advanced techniques.

The latest developments are also connected with some of the following listed topics here

(1) Heteroscadasticity and Autocorrelation in ECM

(2) Unbalanced Panel Data

(3) Dynamic panel data model with lagged values of the variables

(4) Qualitative dependent variable and panel data

(5) Unit roots in panel data

(6) Simultaneous equations involving panel data

(7) Data Mining

(8) Big data analysis etc.

3. Ten Golden Rules for Econometrics Applications

After learning econometrics theoretical background, we have to put our knowledge for its applicability. That is a practician's approach for applied research work in this field. We should give due respect to the ten golden rules for econometric applications.

They are given by Prof.Kennedy of S.F.University in Canda – calling them as Ten Commandsments.

Let us quote these golden rules in brief as under

(1) Use common sense and economic theory

(2) Always ask the right questions (i.e. put relevance before mathematical elegance)

(3) To visualize the context. (i.e. Do not perform ignorant statistical analysis)

(4) Always inspect the data

(5) Do not use complexity, that is keep it stochastically simple

(6) Look long and hard at the results

(7) Beware about the costs of data mining.

(8) Always have a willing to compromise (i.e. Do not worship always the textbook prescription)

(9) Do not confuse statistical significance with practical significance.

(10) Confess in presence of sensitivity (i.e. anticipate criticism).

The above stated golden rules by themselves can speak about the situations that one has to face and also show a way to the milestone of landmark.

8. Concluding Remarks

We have started this lecture series on 'Advanced Econometrics' subject and learnt many things in this course work by means of 12 lectures designed as per the scheduled course.

We started with a brief introduction of the subject with classical linear model and its variations. We have learnt about homoscadasticity, heteroscadasticity, multicollinearity, autocorrelation, time series models, specification errors, simultaneous equations models, lagged variables models,

dummy variables technique and related models and finally panel data analysis.

Yet this is not the end, but it is the beginning! The reason is that there are tremendous advancements in the subject and we have just made a humble approach to understand about this beginning. Truly speaking, Knowledge has no ends. This subject has extremely become popular by its varied applications in many ways. Thus it necessary to have some preliminary knowledge about the subject matters and its application areas. It is our quest for knowledge and pursuit for perfection.

At the end let us see the commonly used and very popular softwares which are as listed here.

9. APPENDIX

LIST OF SELECTED ECONOMETRIC SOFTWARE PACKEGES

* IBM-PC And Compatibles

(1)BMDP/PC (2) ESP (3) ET

(4) DATA-FIT (5) GAUSS (6) LIMDEP

(7) MATLAB (8) MICRO TSP (9) MINI TAB

(10) PC-GIVE (11) PC-TSP (12) RATS

(13) SAS/STAT (14) SHAZAM (15) SORTEC

(16) SPSS/PC+ (17) STATA (18) STATAPRO

(19) STATGRAPHICS (20) SYSTAT

APPLE MACINTOSH

(1) MAT LAB

(2) PC-TSP

(3) RATS

(4) SHAZAM

*Ref. Damodar Gujarati (Third Edition) (1995) MC Graw Hill International Editions (New York)