BLOCK 3
PANEL DATA MODELS

INTRODUCTION TO BLOCK 3

Block 3 is on Panel Data Models. It has two units (Units 8 and 9). **Unit 8** is on Introduction to Panel Data. It first outlines the distinction between pooled cross section data and panel data. Advantages of panel data over pooled data is then explained. The unit discusses Chow Test used in situations of Linear Static Panel Data Models. A discussion on differentiating between 'fixed effect models' and 'random effect models', with a reference to Policy Relevant Inference, is then presented in the unit.

Unit 9 is on Estimation of Panel Data Models. Apart from explaining the estimation methods, the unit discusses a model selection test, for deciding between the fixed effects and random effects models, viz. the Hausman Test.

UNIT 8 INTRODUCTION TO PANEL DATA*

Structure

8.0 OBJECTIVES

After reading this unit, you will be able to:

- differentiate between panel data and pooled data;
- state the distinctive features of time series data, cross-section data and panel data;
- explain, with illustrations, the panel data models;
- indicate the problem of 'endogeneity' in a 'static panel data model';
- distinguish between the concepts of 'self-selection' and 'endogeneity';
- illustrate the framework of a 'pooled cross section data';
- elucidate the features of 'panel data';
- contrast the advantages of 'panel data' over 'pooled data';
- define the terms 'Static Panel Data Model' and 'Dynamic Panel Data Model';
- specify the steps involved in performing a Chow Test;
- discuss the features of 'fixed effect (FE) model' and 'random effect (RE) model' in panel regressions;

* Dr. Poulomi Roy, Jadavpur University.

- present a comparative profile of the relative contexts in which the FE or the RE model is appropriate to adopt; and
- outline, with illustration,' the 'policy relevant inference' that could be drawn from panel data models.

8.1 INTRODUCTION

In applications, econometricians often use either pure cross sectional data or time series data. A cross sectional data is one which is collected for different sample units for a same point of time (e.g. NSSO's 5-yearly data on manufacturing firms). A time series data, on the other hand, is collected over different time points for the same set of sample units (e.g. GDP). Such sample units, in a pure time series data, could themselves be cross sectional in nature.

A data set that has both cross sectional and time series dimensions are nowadays very common in empirical research. Such data sets, often used for policy analysis, could be pooled to form a panel data set. Note that an **independently pooled cross section** can be obtained by a random sampling of a large population at different points of time (usually, but not necessarily, for different years). From a statistical standpoint, such data sets have an important feature i.e. they consist of *independent* sampled observations. Such independently sampled observations play a key role in our analysis of cross-sectional data, where, among other things, it rules out correlation in the error terms across different observations. An independently pooled cross section data differs from a single random sample. This is in the sense that sampling from the population at different points of time likely leads to observations that are not identically distributed. For instance, distributions of wages and education have changed over time in most countries.

A panel data, or longitudinal data set, thus consists of time series data for each cross sectional unit in the data set. Panel data is collected on the same individuals or firms or geographical units over specified periods of time. The key difference between the panel data and pooled data is that, in case of panel data, the same cross sectional units are followed over a given time period. In case of pooled data, different cross section units are observed for a given time period. Thus, the main features of the three types of data are:

a) Time Series Data: Many observations (large T) on as few as one unit (small N): e.g. stock price trends, aggregate national statistics.

b) Pooled Cross Sections: Two or more independent samples of many units (large N) are drawn from the same population at different time periods.

c) Panel Data: Observations of multiple phenomena obtained over multiple time periods (small T) on many cross section units (large N).

Hence, for using methods of panel data analysis, we need information on same cross section units over a given period of time. This is more

difficult to obtain than a pooled cross section data where the sampled units could be different. But in observing the same cross sectional units over time repeatedly, it helps us in controlling for certain 'unobserved characteristics' of the cross sectional unit.

8.2 PANEL DATA MODELS

Let us now illustrate panel data models with some instances. Suppose that the population consists of all manufacturing firms in a country operating during a given three year period. Production function describing the output in the population of firm can be specified as:

$$Log(output_{it}) = \delta_t + \beta_1 \log(lobour_{it}) + \beta_2 \log(capital_{it}) + \beta_3 spillover_{it} + quality_i + u_{it} \quad (8.1)$$

Here, spillover is a measure of foreign firm concentration in a region containing the firm. The term quality refers to unobserved factors (e.g. managerial or work quality) affecting productivity. We must note that quality is a firm specific term and is constant over time. It also has the same effect in each time period, while u_{it} changes across time and firm. Each firm is randomly chosen from the population of all manufacturing firms. Thus, in a panel regression, for a specification like in (8.1), *'i'* is an indicator of cross section unit and *'t'* is an indicator of time. In analysing a panel data set, our aim is to capture this time constant for firms as a specific unobserved effect. The error term *'u'* represents the unobserved shocks in each time period. The presence of the parameter δ_t represents intercepts in each time period, allowing for aggregate productivity to change over time. The coefficients of regressors are assumed to be constant.

Another context we can consider relates to the effect of a training programme on employees (or any other similar programme like providing mid day meals to children in school) in their subsequent performance. Its specification can be considered as:

$$\log(performance_{it}) = \theta_t + z_{it}\gamma + \delta_1 prog_{it} + c_i + u_{it} \quad (8.2)$$

where '*i*' indicates individual, *'t'* indicates time period and θ_t indicates the time varying intercept. z_{it} is the set of observable characteristics that affect not only wage but may also be correlated with program participation. c_i indicates the ability of the individual. Now, suppose at t =1 no one has participated in the programme. It implies $prog_{i1} = 0$ for all i. Then, let us say some individuals are chosen to participate in the programme and their subsequent performance are observed for the two groups (i.e. the group which did not undergo training and the group which underwent the training). The sub-group that participates in the training programme is defined as the 'treatment group' and the other one as the 'control group'. In period t = 1 none received treatment but in t = 2 treatment group received training but the control group did not receive the training. The term c_i included in (8.2) stands

for an individual *'i'* who can choose to participate in the programme with his/her own choice i.e. it can be correlated with the inherent ability (or proactive initiative) of the individual. This is identified in the literature as the problem of 'self-selection'. The important issue in a panel model like (8.2) is whether unobserved factors of productivity relevance are correlated with the observable factors? Another issue is whether we can assume at any time point *t*, that the unobserved effect is uncorrelated with the error term of other time periods or not? For instance, the effect of job training on productivity and on subsequent wages. This is known as the problem of 'endogeneity'. In the above example we can see how the self-selection problem can lead to the problem of endogeneity.

8.2.1 Pooled Cross Section Data

Pooled cross sectional data are obtained by collecting random samples from a large population independently of each other at different point of time. Panel data sets have both cross-sectional and time series features (it consists of time series data for each statistical unit in the cross section). For instance, consider two cross-sectional household surveys taken: one in 1985 and one in 1990. In 1985, a random sample of households were surveyed with variables like income, savings, family size, etc. In 1990, a *new* random sample of households was taken using the same survey questions. To increase our sample size, we can form a **pooled cross section** by combining the two years. Pooling cross sections from different years is an effective way for analysing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change. As an example, we can consider the data on housing prices taken in 1993 and 1995 i.e. before and after a reduction in property taxes was effected in 1994. Suppose we have data on 250 houses for 1993 and on 270 houses for 1995. One method of arranging such a data set is as given in Table 8.1. Observations 1 through 250 correspond to the houses sold in 1993, and observations 251 through 520 correspond to the 270 houses sold in 1995. A pooled cross section is analysed much like a standard cross section, except that we often need to account for secular differences in the variables across time. In fact, in addition to increasing the sample size, the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time. With large N and small *T* one may introduce separate intercepts for each time period.

Table 8.1: Pooled Data on Houses Sold

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
-	-	-	-	-	-	-
-	-	-	-	-	-	-

-	-	-	-	-	-	-
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
520	1995	57200	16	1100	2	1.5

Source: Woolridge (5th Edition)

In India, many surveys of individuals, households and firms are repeated in the NSSO's [National Sample Survey organisation (NSSO)] periodic surveys conducted on individuals and households at regular intervals. For these surveys, NSSO randomly samples households at every five year interval. If a random sample is drawn at each time period, pooling the resulting random samples gives us an independently pooled cross section. One reason for using independently pooled cross sections is to increase the sample size.

8.2.2 Panel Data

The unique characteristic of panel data structure is that each cross section unit is followed over a certain period of time. Panel data sets are fairly easy to collect for districts, cities, states, and countries. Hence, policy analysis is greatly enhanced by using panel data sets. For the econometric analysis of panel data, we cannot assume that the observations are independently distributed across time. For instance, unobserved factors (such as ability) that affect someone's wage in 2010 will also affect that person's wage in 2011. Likewise, unobserved factors that affect a city's crime rate in 2015 will also affect that city's crime rate in 2020. For this reason, special models and methods have been developed to analyse panel data.

In using panel data in an econometric study, it is important to know how the data should be stored. We must be careful to arrange the data so that the different time periods for the same cross-sectional unit (person, firm, city, and so on) are easily linked. For instance, let us suppose that the data set is on cities for two different years. For most purposes, the best way to enter the data is to have *two* records for each city, one for each year. The first record for each city corresponds to the early year, and the second record is for the later year. These two records should be adjacent. Therefore, a data set for 100 cities and two years will contain 200 records for each of the variables. The first two records are for the first city in the sample, the next two records are for the second city, and so on.

The above method of data arrangement makes it easy to obtain the differences in the two records for each city and store them in a pooled cross-

sectional manner for an analysis of the differencing estimation. Most of the two-period panel data sets are stored in this way. We use a direct extension of this scheme for panel data sets with more than two time periods. A second way of organising the two periods of a panel data set is to have only one record per cross-sectional unit. This requires two entries for each variable, one for each time period. Creating the differences from T_1 to T_2 is then easy. Placing the data in one record, however, does not allow for a pooled analysis by using the two time periods on the original data. Also, this method of organisation does not work for panel data sets with more than two time periods. Table 8.2 presents a two-year panel data set on crime and related statistics for 150 cities. Cities are numbered as 1,2,…,150. Just as in a pure cross section, the ordering in the cross section of a panel data set does not matter. We could use the city name in place of a number. But it is often useful to have both.

Table 8.2: Panel Data on Crime and Unemployment by City

obsno	City	Year	Murders	Population	Unem	Police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Source: Woolridge (5th Edition)

8.2.3 Advantages of Panel Data Over Pooled Data

Because panel data require replication of the same units over time, panel data sets, especially those on individuals, households, and firms, are more difficult to obtain than pooled cross sections. Not surprisingly, observing the same units over time leads to several advantages over cross-sectional data or even pooled cross-sectional data. The benefit that we will focus on is of having multiple observations on the same units which allows us to control for certain unobserved characteristics of individuals, firms, etc. As we will see, the use of more than one observation can facilitate causal inference in situations where inferring causality would be difficult if only a single cross section were available. A second advantage of panel data is that it allows us to study the importance of lags in the behaviour or the result of decision making. This information can be significant because many economic policies can be

expected to have an impact only after some time has passed. It therefore follows from here that the advantage of panel data is that we can observe the 'before and after effects' of receiving a treatment by the same individual. It also provides the possibility of isolating the effects of treatment from other factors affecting the outcome.

Panel data obtained by combining both the cross sectional and time series data capture both the inter-cross-sectional differences as well as the intra cross sectional dynamics. It has several other advantages over cross sectional and time series data. For instance, cross sectional data may be viewed as a panel with T =1 and time series data may be viewed as a cross section with N = 1. Hence, panel data combining both cross section and time series data provides more degrees of freedom and more sample variability than either only the cross sectional or only the time series data. It hence improves the efficiency of econometric estimates.

Evaluating the effectiveness of certain programmes by using a cross-sectional sample typically suffers from the fact that those receiving treatment are different from those without. In other words, one does not simultaneously observe what happens to an individual when she receives the treatment or when she does not. An individual is observed as either receiving treatment or not receiving treatment. Using the difference between the treatment group and control group could suffer from two sources of biases: (i) selection bias due to differences in observable factors between the treatment and control groups and (ii) selection bias due to endogeneity of participation in treatment.

It is also frequently argued that the real reason one finds (or does not find) certain effects is 'due to ignoring the effects of certain variables in a model specification which are correlated with the included explanatory variables'. Panel data contain information on both the inter-temporal dynamics and the individuality of the entities. This therefore allows for one to control for the effects of missing or unobserved variables.

By pooling random samples drawn from the same population, but at different points in time, we can get more precise estimators and test statistics with higher power. Pooling is helpful in this regard only in-so-far as the relationship between the dependent variable and at least some of the independent variables remain constant over time. Using pooled cross sections raises a statistical complication viz. the two populations could have different distributions. To reflect for the fact that the populations may have different distributions in different time periods, we allow the intercept to differ across periods. This is also accomplished by including dummy variables for all but one year i.e. for the earliest year in the sample which is usually chosen as the base year. Sometimes, the pattern of coefficients on the year dummy variables could itself be of interest.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) Distinguish between time-series data and cross-section data.

……

……

……

……

……

2) Differentiate between pooled data and panel data.

……

……

……

……

……

3) State the two main advantages of 'panel data' over 'pooled data'.

……

……

……

……

……

4) What is a 'statistical complication' in pooling two cross sections? How is this complication dealt with in practice?

……

……

……

……

……

……

8.3 LINEAR STATIC PANEL DATA MODEL

Suppose for each cross section unit we collect data on same set of variables for T time periods. Let X be a vector of k exogenous variables which affect Y. At any time point 't', the population model is like:

$$Y_{it} = X_{it}\beta + c_i + u_{it}, \; t = 1,2,\ldots,T, \; I = 1,2,\ldots,N \qquad (8.3)$$

where c_i is the unobserved effect and u_{it} is the random error term. (8.3) is a panel regression with *'i'* as an indicator of cross section unit and *'t'* as an indicator of time. The most commonly used method for estimating the parameters β is the 'ordinary least square' (OLS). The OLS assumes that the explanatory variables are exogenous in nature and they are uncorrelated with the random error term. Primary motivation behind the panel data is to solve for the omitted variables problem. In panel data models, we consider time to account for the unobserved effect (like quality in the example considered above). We **assume** that the 'unobserved effects' are random variables. This is an instance of a linear static panel data model. It is a static model because all explanatory variables are contemporaneous dates corresponding to the value of *Y* in period *t*. In contrast, in a dynamic panel data model, one or more lagged dependent variables are allowed in the models as a 'partial adjustment mechanism'. In this unit, we discuss only the static panel data models. You may however note that a dynamic panel model, with one lagged dependent variable and a single regressor *X*, is defined as:

$$y_{it} = \beta_1 + \rho\, y_{it-1} + X_{it}\beta_2 + c_i + \varepsilon_{it} \qquad (8.4)$$

where c_i is for a specific unobserved effect and ε_{it} is the overall random error term.

8.3.1 Chow Test

Chow Test examines whether parameters of one group of data are equal to those in the other groups. Simply put, the test checks whether the data can be pooled. If only intercepts are found different across groups, it becomes a 'fixed effect model'. Let us consider two groups from a model like $y = \alpha + \beta x + \varepsilon$ as follows:

$$y = \alpha_1 + \beta_1 x + \varepsilon_1 \text{ for } n_1 \text{ observations (group 1)} \qquad (8.5)$$

$$y = \alpha_2 + \beta_2 x + \varepsilon_2 \text{ for } n_2 \text{ observations (group 2)} \qquad (8.6)$$

The null hypothesis is $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. If the null hypothesis is rejected, it means the two groups have different slopes and intercepts and hence the data is not poolable. The Chow test is simply an *F* test used to determine whether a multiple regression function differs across two groups. We can also apply this test to two different time periods. The 'sum of squared residuals' (SSR) obtained from the pooled estimation for both the groups combined is designated as the 'restricted SSR' (SSR*r)*. The unrestricted SSR is the sum of the two *SSRs* obtained for the two groups separately. A Chow test can also be computed for more than two time periods. In such cases, we first estimate the restricted model by doing a pooled regression and obtain the *SSRr* (i.e. the restricted *SSR*). We then run separate regressions (for each of the time periods, say, *T*) to obtain the sum of squared residuals for each time period. The unrestricted sum of squared residuals is then obtained as $SSR_{UR} = SSR_1 + SSR_2 + \ldots\ldots\ldots + SSR_T$. If there are *k* explanatory variables

(excluding the intercept or the time dummies) with T time periods, then we are imposing $(T-1)k$ restrictions for the $(T+Tk)$ parameters estimated in the unrestricted models. Hence, if $n = n_1 + n_2 + ... + n_T$ is the total number of observations, then the 'degrees of freedom' (df) for the F test are '$(T-1)k$ and $(n-T-Tk)$'. We compute the F statistic as usual i.e.:

$$(\text{SSR}r - \text{SSR}ur)\ (T-1)k\ /\ \text{SSR}ur\ (n-T-Tk)] \qquad (8.7)$$

You may simply note at this stage that, as with any F test based on sums of squared residuals, this test is not robust to Heteroscedasticity.

8.4 FIXED EFFECT VERSUS RANDOM EFFECT PANEL MODELS

With panel data, the most commonly estimated models are the fixed effects and the random effects models. Let us therefore focus first on the major differences between these two types of models. Several considerations affect the choice between the two types of models. For this, first of all, one has to identify the nature of the variables that have been omitted from the model. If we have reason to believe that there are no omitted variables, or we believe that the omitted variables are uncorrelated with the explanatory variables in the model, then a 'random effects' (RE) model is probably the best. It will produce unbiased estimates of the coefficients, use all the data available and produce the smallest standard errors. On the other hand, if there are omitted variables, and these variables are correlated with the explanatory variables in the model, then 'fixed effect' (FE) models provide a means for controlling the 'omitted variable bias'. In a fixed-effects model, 'subjects' serve as their own controls. The idea is that whatever effects the omitted variables have on the subject at one time, they will also have the same effect at a later time. In this sense, their effect will be 'constant' or 'fixed'. However, for this to be true, the omitted variables must have time-invariant values with time-invariant effects. By time-invariant values, we mean that the value of the variable does not change across time. Gender and race are obvious instances, but this can also include the 'educational level' of the respondent.

Second, one needs to consider the variability within subjects or cross section of units. If subjects change little across time, a fixed effects model may not work very well. This is because, there needs to be within-subject variability if we are to use subjects as their own controls. If there is little variability within subjects, then the standard errors from fixed effects model could be too large. Conversely, random effects models will often have smaller standard errors. But, the trade-off is that their coefficients are more likely to be biased.

Third, one needs to decide whether one wants to estimate the effect of variables whose values do not change across time. With fixed effects models, we do not estimate the effect of variables whose values do not change across time. Rather, we control for them or 'partial them out'. This is similar to an experiment with random assignment. Though the RE models estimate the

effect of time-invariant variables, the estimates could be biased because we are not controlling for omitted variables. For a more clearer description, let us consider a situation where y and $x \equiv (x_1, x_2, ..., x_k)$ are observable random variables with a linear relationship like:

$$y = \alpha + x\beta + c \tag{8.8}$$

where *'c'* the unobservable random variable. We are interested in the partial effect of the observable explanatory variables x_j while holding *'c'* constant. Our interest is to estimate the vector β. If *'c'* is uncorrelated with *x*, then *'c'* is just another unobserved factor uncorrelated with the explanatory variables. If $cov(x_j, c) \neq 0$ for some *j*, then we cannot consistently estimate β.

8.4.1 Fixed Effect Model

Fixed effect (FE) are thus variables that are 'constant across individuals'. These variables are like age, sex, ethnicity which do not change (or change at a constant rate) over time. FE explores the relationship between the predictor variables (i.e. explanatory or independent variables) and outcome variables (i.e. the dependent variable). The relationship between them is explored within an entity (country, person, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables. For instance, being a male or female could influence the opinion toward certain issue, the political system of a particular country could have some effect on trade or GDP, the business practices of a company may influence its stock price, etc. When using FE, we assume that something within the individual may impact, or bias, the predictor and therefore we might wish to control for this. This is the rationale behind the assumption of the correlation between entity's error term and predictor variables. FE removes the effect of those time-invariant characteristics so that we can assess the net effect of the predictors on the outcome variable. Another important assumption of the FE model is that the time-invariant characteristics are unique to the individual and are not correlated with other individuals' characteristics. In other words, each entity is different and therefore the entity's error term and the constant (which captures individual characteristics) are not correlated with the others. If the error terms are correlated, then the FE model is not suitable. In that case, we need to model that relationship using the RE model.

The FE model allows the unobserved individual effects to be correlated with the included variables. We can therefore model the differences between units as parametric shifts of the regression function. This could be viewed as applying only to the cross-sectional units in the study and not for the additional units outside the sample. For instance, an inter-country comparison may include the full set of countries for which it is reasonable to assume that the model is constant. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across the cross-sectional units.

8.4.2 Random Effect Model

The random effects (RE) model is useful when we have reason to believe that the unobserved effect is uncorrelated with all the explanatory variables. In such a situation, the time constant's unobserved effect is uncorrelated with the explanatory variables and the *parameters* could be consistently estimated by using a single cross section. There is therefore no need for panel data. But using a single cross section disregards much useful information in the other time periods. We can therefore use the data in a pooled OLS procedure i.e. just run the OLS of *dependent variable* on the explanatory variables with the time dummies. This, too, produces consistent estimators of the *parameters under* the RE assumption. But it ignores the fact that the existence of unobserved effect in the error term in each time period is serially correlated across time. We can use the GLS method to solve for the serial correlation problem.

RE assumes that the unobserved effect is uncorrelated with all explanatory variables irrespective of whether the explanatory variables are fixed over time or not. Hence, we can include a variable like education even if it does not change over time. But we are assuming that education is uncorrelated with the unobserved effect. Hence, in applications of FE and RE, it is usually informative to compute the pooled OLS estimates. Comparing the three sets of estimates can help us determine the nature of the biases caused by leaving the unobserved effect. We must, however, remember that, even if the *unobserved effect* is uncorrelated with all explanatory variables in all time periods, the pooled OLS standard errors and test statistics are generally invalid. This is because they generally ignore the often substantial serial correlation in the composite errors. But it is possible to compute the standard errors and test statistic which are robust to arbitrary serial correlation (and Heteroscedasticity) in composite error. Note that the FE approach allows for the arbitrary correlation while the RE approach does not. Hence, the FE approach is widely thought to be a more convincing tool for estimating the '*ceteris paribus*' effects.

To sum up, therefore, if the key explanatory variable is constant over time, we cannot use FE to estimate its effect on dependent variable. In such situations, we must rely on the RE (or pooled OLS) estimate. We can however use the RE approach if we are able to assume that the unobserved effect is uncorrelated with the explanatory variables. Typically, when one uses random effects, many time-constant controls are included among the explanatory variables. However, with the FE approach, it is not necessary to include such controls. RE is preferred to pooled OLS due to its generally higher efficiency.

8.4.3 Policy Relevant Inference

The choice of fixed or random effects should be based on the basis of the background knowledge and the availability of data. Let us have clarity on

what we mean here by the term 'policy-relevant inference'. Ideally, policy-relevant inferences are causal inferences about average treatment effects. Causal inferences tell us what happens if we intervene and change the way the things are being done. Within the regression modelling framework, and in the absence of experimental or quasi-experimental data, many issues can be overcome by making assumptions. But, estimating the treatment effect in an unbiased manner becomes difficult. A realistic goal is therefore to produce policy-relevant estimates that may be biased, but are not too much so, so as to lead to misleading policy recommendations. Recall that the RE approach requires the strong assumption that the unobserved effect is uncorrelated with any of the covariates. An important reason why the random effect assumption fails is that there is usually non-random selection of cross section units. For instance, if each school had drawn its pupils at random from the pupil population, then the random effect assumption would hold. But, in reality, a non-random selection mechanism operates through which parents choose schools and some schools select which children to accept. Thus, the probability of selecting a particular school varies systematically according to a series of factors characterising the child, his/her family, the school itself or the higher local education authority. Some of these factors will be associated with pupil attainment, either directly or indirectly, through a mediating mechanism.

In light of the above, the debate remains inconclusive on whether we should conclude that the FE approach is always preferable? The answer depends on circumstances. The FE estimator for β is robust when a school F is not empty. Therefore, if we have some knowledge about the school selection mechanism, and we can include measures of these factors in the model as 'controls', then we can estimate the average treatment effect using the RE approach.

Check Your Progress 2 [answer within space given in about 50-100 words]

1) Distinguish between Linear 'Static Panel Data Model' and 'Dynamic Panel Data Model'.

……

……

……

……

2) For what purpose is the 'Chow Test' used? What does it basically seek to examine?

……

……

……

……

3) In what contexts, the 'fixed effects' or the 'random effects' panel data model used?

……

……

……

……

……

4) Specify the considerations that determine the choice between the FE and the RE models.

……

……

……

……

……

8.5 LET US SUM UP

The unit introduces the panel data models. Panel data refers to observations on multiple variables obtained over different time periods for the same firms or individuals. It can be understood by the common expression 'the two data sets are drawn from the same panel'. In contrast, pooled cross section data refers to a time series of cross-sections where the observation on each cross section do not necessarily relate to the same units. In India, the surveys of NSSO, conducted on many subjects periodically, usually at an interval of 5 years, are based on independent random samples. They are therefore useful for methods of 'pooled data' analysis and techniques. The unit has introduced you to the concepts and application of two main leading approaches viz. the FE approach and the RE approach. The contexts in which the choice between the two could be made is outlined. Generally, the choice need to be based on the background knowledge on variables and the nature of availability of data.

8.6 KEY WORDS

Pooled Cross Section Data : Refers to data collected on same cross sections at two different points of time but pooled for the purpose of analysis. By combining the two samples, we get increased degrees of freedom or higher *'n'*. Data is pooled to assess the impact of a new government policy. In other words, pooled cross section data helps us in assessing the before/after effects.

Panel Data : Refers to data collected for two time points on same sample units. In other words, unlike in 'pooled cross section', no

new random samples are used in the two surveys. Data is collected on same variables. This is particularly useful for assessing the effect of 'lags' which is usually there in govt. policies introduced.

8.7 SUGGESTED BOOKS FOR FURTHER READING

1) Wooldridge J M (2006). Introduction to Econometrics: A Modern Approach. *Michigan State University. USA*.

2) Greene W H (2016). *Econometrics*. Prentice Hall.

8.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Time series is data is collected over different time points for the same set of sample units (e.g. GDP for states). Cross section data is collected over different sample units for a same point of time (e.g. NSSO's surveys in India on 5-yearly basis).

2) In case of panel data, the same cross sectional units are followed up over different time periods. In case of pooled data, different cross section units (i.e. two independently selected random samples) are observed for a given time period.

3) One, it allows for causal inference. Second, it allows us to study the effect of lags in the behaviour or the result of decision making.

4) The complication is that the two samples might have come from populations with different distributions. The way it is dealt with is by allowing for different intercept terms or by using a 'dummy variable'.

Check Your Progress 2

1) It is like $Y_{it} = X_{it}\beta + c_i + u_{it}$, $t = 1,2,\ldots,T$, $I = 1,2,\ldots,N$ where c_i is for a specific unobserved effect (like quality), assumed to be a random variable. u_{it} is the overall random error term, accounting for all other unobserved factors or effects. When one or more lagged dependent variables are allowed in the models, as a 'partial adjustment mechanism', it becomes a dynamic panel data model.

2) It is used to test for the feature of 'poolability' across data collected in groups. In other words, it seeks to examine whether the parameters in the models for the two or more groups are equal.

3) In general, a panel data model is used for determining the effect of 'omitted variables'. If we have reason to believe, no variable is omitted, then 'random effect model' can be used. If it is not so, applying the

'fixed effects panel data model' helps in controlling for the 'omitted variable bias'.

4) If the key explanatory variable is constant over time, then RE model is to be applied. Alternatively, we can use the RE model when 'we are able to assume that the unobserved effect is uncorrelated with the explanatory variables'. The FE approach allows for the arbitrary correlation while the RE approach does not. Hence, the FE approach is a convincing tool for estimating the '*ceteris paribus*' effects. Therefore, The choice of fixed or random effects should be based on the basis of the background knowledge and the availability of data.

UNIT 9 ESTIMATION OF PANEL DATA MODELS*

Structure

9.0 OBJECTIVES

After reading this unit, you will be able to:

- specify the form of a 'panel data regression model';
- state the assumptions behind the 'random effect estimation models';
- discuss the various features of the 'method of random effect estimation';
- list the assumptions behind the 'fixed effect estimation models';
- explain the features of 'fixed effect estimation model';
- describe the Hausman's 'model selection procedure'; and
- elucidate, with illustrations, the application of 'simultaneous equation models' in panel data contexts.

9.1 INTRODUCTION

In a panel data set up, data is collected on each cross section unit '*i*'. It is done over a period of time '*T*'. Therefore, if *X* is a vector of '*k*' exogenous variables affecting *Y* (the dependent variable), at any time point '*t*', the population model is of the form:

$$Y_{it} = X_{it}\beta + c_i + u_{it},\ i = 1,2,...,N;\ t = 1,2,....,T \qquad (9.1)$$

where c_i is the 'unobserved effect' and u_{it} is the 'random error term'. They are also called as 'idiosyncratic error terms' because they change over 'time and cross sections units' i.e. over *i* and *t*. Equation (9.1) is a panel regression

* Dr. Poulomi Roy, Jadavpur University

where '*i*' indicates a cross section unit and '*t*' is an indicator of time. The primary motivation behind a panel data analysis is solving for the 'omitted variables problem'. In panel data models, we consider c_i as 'time invariant'. We further assume that the 'unobserved effects' are random variables. This is in the case of a 'linear static panel data model'. It is a 'static model' in the sense that the explanatory variables are contemporaneous (fixed or pre-determined) values corresponding to the value of *Y* in period *t*. In Equation (9.1), to enable us to assume that the effect of c_i is with a zero mean, we explicitly add an intercept term. With this, Equation (9.1) takes the form:

$$Y_{it} = \beta_0 + X_{it}\beta_i + c_i + u_{it} \qquad (9.2)$$

9.2 RANDOM EFFECT ESTIMATION METHOD

In the absence of any 'omitted variable', where the omitted variables are assumed to be uncorrelated with the regressors, the random effect estimation technique is preferred. The assumptions required for the estimation of the parameters of the random effect model are the following:

i) For the i^{th} cross section unit, at any time point '*t*', the model is defined as:

$$y_{it} = \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + c_i + u_{it} \qquad (9.3)$$

where c_i is the 'unobserved effect' and u_{it} is the 'random error term' (i = 1,2,……,N and t = 1,2,.....,T). βj's are the parameters to be estimated.

ii) Each explanatory variable changes over time (at least for some *i*). No perfect linear relationship exist among the explanatory variables.

iii) For each *t*, the 'expected value of the idiosyncratic error terms' given the X_i's [for all (∀) '*t*' and 'c_i'] is 'zero' i.e. $E(u_{it} \mid X_{i,} c_i) = 0$.

iv) $\text{var}(u_{it} \mid X_i, c_i) = \text{var}(u_{it}) = \sigma_u^2$, t = 1,2,...,T.

v) The covariance of the uncorrelated error terms, conditional to X_i's and c_i's ∀ $t \neq s$ is 'zero' i.e. $Cov(u_{it}, u_{is} \mid X_i, c_i) = 0$

vi) $E(c_i|X_i) = E(c_i) = 0$ i.e. the expected value of c_i is constant.

vii) $var(c_i \mid X_i) = \sigma_i^2$ i.e. the assumption of homoscedasticity hold for c_i's.

Under the above assumptions, the random effect model (9.3) can be written as:

$$y_{it} = x_{it}\beta + v_{it} \qquad (9.4)$$

where $v_{it} = c_i + u_{it}$ is the 'composite error term' such that $E(v_{it} \mid x_{ti}) = 0$, t = 1,2,3,....,T. Note that the defining of the composite error term (as $v_{it} = c_i +$

u_{it}) implies that the v_{it}'s are 'serially correlated across time'. Hence,

$$Correlation\ (v_{it}, v_{is}) = \frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2},\ t \neq s\ .$$

Therefore, the usual pooled OLS procedure cannot be applied. In such cases, we have to apply the GLS method of estimation with 'auto regressive serial correlation'. Therefore, the model in (9.4) can be written as:

$$y_i = x_i\beta + v_i \qquad (9.5)$$

where $v_i = c_i\, j_T + u_i$ (j_T being the TX_1 vector of ones) and we define the unconditional variance matrix v_i as: $\Omega = E(v_i, v_i')$. We assume Ω to be a 'positive definite matrix' which is constant for all cross section units due to the assumption of random sampling. This means, for the consistency of GLS, we need the 'rank condition' that $Rank\left(E\left(X_i'\Omega^{-1}X_i\right)\right) = k$. Under this assumption, $E(c_i, u_{it}) = 0\ \forall\ t$=1,2,3,...,$T$ and

$$E\left(v^2{}_{it}\right) = E(c_i{}^2) + 2E(c_i u_{it}) + E(u^2{}_{it}) = \sigma_c^2 + \sigma_u^2$$

where $\sigma_c^2 = E(c_i^2)\ \forall\ t \neq s$ |.

We have to now derive the variance-covariance of the elements of v.

$$E\left(v_{it}, v_{is}\right) = E\left(c_i + u_{it}, c_i + u_{is}\right) = E\left(c_i^2\right) = \sigma_c^2$$

Therefore:

$$\Omega = E\left(v_i, v_i'\right) = \begin{pmatrix} \sigma_c^2+\sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2+\sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \cdots & \sigma_c^2+\sigma_u^2 \end{pmatrix}$$

Thus, $\Omega = (\sigma_c^2 i_T j_T + \sigma_u^2 I_T)\ i_T j_T$ where Ω depends on two parameters σ_c^2 *and* σ_u^2 regardless of the size of T. For efficiency of feasible GLS, we assume that $E\left(v_i, v'_i \mid x_i\right) = E\left(v_i, v'_i\right)$. Now the RE assumption $E\left(u_i, u'_i \mid x_i, c_i\right) = \sigma_u^2 I_T$ implies
$E\left(u_{it}, u_{is} \mid x_i, c_i\right) = 0, t \neq s; t, s = 1,2,3,..., T$

With the estimated σ_u^2 and σ_c^2 used in Ω we get $\hat{\Omega} = (\hat{\sigma}_u^2 I_T + \hat{\sigma}_c^2) i_T j_T'$. Applying the GLS technique, the RE estimator is obtained as:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} x_i'\Omega^{-1}x_i\right)^{-1}\left(\sum_{i=1}^{N} x_i'\Omega^{-1}y_i\right).$$

Now, given that: (i) $E\left(u_{it} \mid x_i, c_i\right) = 0$, *(ii)* $E\left(c_i \mid x_i\right) = E\left(c_i\right) = 0$ and (iii) $Rank\ \left(E\left(X_i'\Omega^{-1}X_i\right)\right) = k$, $\hat{\beta}_{RE} \to \beta$ as $N \to \infty$. Deriving the GLS transformation that eliminates the serial correlation in the errors, requires

sophisticated matrix algebra. For this, let us define: $\theta = 1 - \left[\sigma_u^2 / (\sigma_u^2 + T\sigma_c^2)\right]$ which lies between zero and one. Then, the transformed equation becomes:

$$y_{it} - \theta\overline{y}_i = (x_{it} - \theta\overline{x}_i)\beta + (v_{it} - \theta\overline{v}_i) \qquad (9.5a)$$

where the 'over bar' denotes the 'time averages'. This is a very interesting equation, as it involves 'quasi-demeaned data' on each variable. If the model considered had been of 'fixed effects', then the estimator would have subtracted the time averages from the corresponding variable. Since we are here considering the random effects transformation, it subtracts a fraction of that time average, the fraction depending on σ_u^2, σ_c^2 and T. The GLS estimator is thus simply the pooled OLS estimator of Equation (9.5a). The transformation in (9.5a) allows for the explanatory variables that are constant overtime. This is in fact the one advantage of random effects (RE) over 'fixed effects' (or first differencing). This is possible because RE assumes that the 'unobserved effect' is uncorrelated with all explanatory variables irrespective of whether the explanatory variables are fixed over time or not. In many applications, the whole rationale for using panel data is to allow for the unobserved effect to be correlated with the explanatory variables.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) Specify a 'linear static panel data model'. Why is the model called 'static'?

……

……

……

……

……

……

2) Why is it that in the case of 'linear static panel data models', the usual OLS method cannot be applied? Which method is to be applied in such cases?

……

……

……

……

……

……

……

3) State the assumptions required to be made for the 'random effect panel models'.

……

……

……

……

……

4) In Equation (9.5a), what is meant by 'quasi demeaned data'? What is its significance?

……

……

……

……

……

9.3 FIXED EFFECT ESTIMATION METHOD

Let us again consider the 'linear unobserved effects model' [(9.2)] for T time periods:

$$y_{it} = x_{it}\beta + c_i + u_{it},\ t=1,2,....,T$$

The random effect approach to estimate β effectively puts c_i into the error term. This is done under the assumption that c_i is orthogonal to x_{it}. It then accounts for the implied serial correlation in the composite error term: $v_{it} = c_i + u_{it}$ using a GLS analysis. In other words, the FE analysis assumes that c_i is correlated with x_{it}. The T equations in the model are like:

$$y_{it} = x_{it}\beta + c_{it} j_T + u_{it} \qquad (9.6)$$

where j_T is the TX_1 vector of one. The assumptions required for the FE model are the following:

i) For each i, the model is:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + ... + \beta_k x_{itk} + c_i + u_{it},\ t = 1,2,...,T$$

where the β_j are the parameters to be estimated and c_i is the 'unobserved effect'.

ii) The cross section is a random sample.

iii) Each explanatory variable changes over time (for at least some i) with 'no perfect linear relationships' existing among the explanatory variables.

iv) For each t, the expected value of the idiosyncratic error term, given the explanatory variables in *all* time periods, and the unobserved effect, is 'zero'. That is: $E(u_{it} \mid x_i, c_i) = 0$.

[Under these assumptions {(i) to (iv)}, the FE estimator of β_j is consistent with a fixed T as $N \to \infty$].

v) $Var(u_{it} \mid x_i, c_i) = Var(u_{it}) = \sigma_u^2 \; \forall t = 1,2,\ldots\ldots,T$

vi) For all $t \neq s$, the idiosyncratic errors, conditional on all explanatory variables and unobserved error term c_i, are uncorrelated.

[Under the assumptions {(i)-(vi)}, the FE estimator of β_j is the 'best linear unbiased estimator' (BLUE)].

vii) Conditional on x_i and c_i, the u_{it} are 'iid' i.e. independent and identically distributed as N (0, σ_u^2).

If we add the assumption (vii), then the FE estimator is normally distributed with the t and F statistics having the exact t and F distributions. Without the assumption (vii), we can rely on asymptotic approximations. In other words, without making this special assumption (i.e. the viith one), the approximations require large N and small T.

Now to estimate β of Equation (9.6) under the above assumptions, we transform the equation to eliminate the 'unobserved effect' c_i. We obtain this FE transformation by first averaging Equation (9.6) as:

$$\bar{y}_i = \bar{x}_i \beta + c_i + \bar{u}_i \tag{9.7}$$

where $\bar{y}_i = \frac{1}{T}\sum y_u$, $\bar{x}_i = \frac{1}{T}\sum x_u$, $\bar{u}_i \frac{1}{T}\sum u_u$,

Now, subtracting (9.7) from (9.6) we get:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta + (u_{it} - \bar{u}_i) \tag{9.8}$$

Or, $\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it}$, t = 1,2,..,T (9.9)

where, $\ddot{y}_u = y_u - \bar{y}_i$, $\ddot{x}_{it} \equiv x_{it} - \bar{x}_i$, $\ddot{u}_{it} \equiv u_{it} - \bar{u}_i$

Note that in (9.8), the 'time demeaning' has removed the individual specific effect c_i. The 'fixed effects transformation' is also called as the 'within transformation'. The important thing about Equation (9.9) is that the unobserved effect c_i, is eliminated. This suggests that we should estimate (9.9) by pooled OLS. Thus, our 'fixed effects estimator' (also called the 'within estimator'), is a pooled OLS estimator, based on the time-demeaned variables. Now we proceed to estimate (9.9) by the pooled OLS method as follows. Since, $E(\ddot{x}'_{it}, \ddot{u}_{it}) = 0$ t = 1,2,....,T, under a strict exogeneity assumption on the explanatory variables, the fixed effects estimator is

unbiased. Alternatively, at least roughly, the idiosyncratic error u_{it} are uncorrelated with each explanatory variable across *all* time periods. Note that the 'fixed effects estimator' allows for the arbitrary correlation between c_i and the explanatory variables in any time period. Because of this, any explanatory variable that is constant over time for all *i* gets swept away by the fixed effects transformation: $\ddot{x}_{it} = 0$ for all *i* and *t*. In other words, we cannot include variables like gender or a city's distance from a river. Note also that the FE estimator is the pooled OLS estimator from the regression of $\ddot{y}_{it}$ on $\ddot{x}_{it}$, $t = 1,2,....,T;\ i = 1,2,...,N$. In order to ensure that the FE estimator is well behaved asymptotically, we need a standard rank condition on the matrix of time demeaned explanatory variables viz.

$$Rank\left[\sum_{t=1}^{T} E\left(\ddot{x}'_{it}, \ddot{x}_{it}\right)\right] = Rank\left[E\left(\ddot{x}'_{it}, \ddot{x}_{it}\right)\right] = k.$$

This implies that the 'time constant variables' are not allowed in FE analysis unless they are interacted with 'time varying variables' such as time dummies. Hence, when analysing individuals, factors such as gender or race cannot be included among x_{it}. We can now define the FE estimator as:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \ddot{x}_i' \ddot{x}_i\right)^{-1} \left(\sum_{i=1}^{N} \ddot{x}_i' \ddot{y}_i\right) = \left(\sum_{i=1}^{N}\sum_{i=1}^{T} \ddot{x}_{it}' \ddot{x}_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it} \ddot{y}_{it}\right) \quad (9.10a)$$

and $\hat{c}_i = \overline{y}_i - \overline{x}_i \hat{\beta}_{FE}$ (9.10b)

When we estimate the time-demeaned Equation (9.9) by the pooled OLS method, we have *NT* total observations and *k* independent variables. There is no intercept term in (9.9) since it is eliminated by the fixed effects transformation. Therefore, we should apparently have '*NT – k'* degrees of freedom. However, this calculation is incorrect. For each cross-sectional observation *i*, we lose one *df* because of time-demeaning. In other words, for each *i*, the demeaned errors $\ddot{u}_{it}$ add up to zero when summed across *t*. Hence, we lose one degree of freedom. There is no such constraint on the original idiosyncratic errors u_{it}. Therefore, the appropriate degrees of freedom is *'NT–N – k = N*(*T* –1) – *k'*. Note that the significance of FE estimators can be tested by using *t*-statistic. Note also that, unlike here, in case of RE estimation we do not lose any *df* i.e. in the RE's case the *df* is full (*NT – k*). Note further that although the time-constant variables cannot be included by themselves in a fixed effects model, they *can* interact with variables that change over time i.e. with year dummy variables. Let us now consider the following wage equation:

$$\log(wage_{it}) = \alpha_0 + \beta_1 educ_i + \beta_2 \exp er_{it} + \beta_3 \exp er_{it}^2 + u_{it} \quad (9.11)$$

In the above wage equation, education is constant over time for each individual in the sample. However, we can interact education with each year

dummy to see how the return to education has changed over time. But we cannot use fixed effects to estimate the return to education in the base period. This means we cannot estimate the return to education in any period. We can only see how the return to education in each year differs from that in the base period.

In applications of FE and RE, it is usually informative to compute the pooled OLS estimates. Comparing the three sets of estimates can help us determine the nature of the biases caused by leaving the unobserved effect, c_i, 'entirely' in the error term (as done in the pooled OLS for FE) or 'partially' in the error term (as done in the RE transformation). But we must remember that, even if c_i is uncorrelated with all explanatory variables in all time periods, the pooled OLS's standard errors and test statistics are generally invalid. This is because they ignore the often substantial serial correlation in the composite errors, $v_{it} = c_i + u_{it}$.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) In Equation (9.2), what does 'c_i is orthogonal to u_{it}' mean?

……

……

……

……

……

……

2) In what respects, the assumptions made for the FE differ from those made for the RE effect?

……

……

……

……

……

……

3) How is the FE estimate obtained in principle? To what effect?

……

……

……

……

……

4) What is the d.f. in estimating the FE Equation (9.9)?

……

……

……

……

……

5) What is a limitation of the FE estimate?

……

……

……

……

……

9.4 MODEL SELECTION: HAUSMAN TEST

The key consideration in choosing between a random effects and fixed effects approach is whether c_i and x_{it} are correlated. It is therefore important to have a method for testing this assumption. Hausman (1978) proposed a test based on the difference between the random effects and fixed effects estimates. Since FE is consistent when c_i and x_{it} are correlated (whereas RE is inconsistent), a statistically significant difference is interpreted as 'evidence against the random effects'.

If our interest is in a time-varying explanatory variable, then one should use RE rather than FE. But situations in which $Cov(x_{it}, c_i) = 0$ should be considered the exception rather than the rule. In most cases, the explanatory variables are themselves a result of selection processes and would likely be correlated with the unobserved factor, c_i. In practice, one applies both the random effects and the fixed effects, and then formally test for the statistical significance of the differences of the coefficients on the time-varying explanatory variables.

Hausman (1978) first proposed such a test and some econometrics packages routinely compute the Hausman test under the full set of random effects assumptions. The idea is that one uses the random effects estimates unless the Hausman test rejects $Cov(x_{it}, c_i) = 0$. In practice, a failure to reject means 'either that the RE and FE estimates are sufficiently close and it does not matter which of the two methods is used'. A rejection, using the Hausman test, is taken to mean that the key RE assumption, $Cov(x_{it}, c_i) = 0$, is false. In such cases, the FE estimates are used.

Hausman's specification test compares an estimator $\hat{\beta}_1$ to be consistent with an estimator $\hat{\beta}_2$ that is efficient. The null hypothesis it tests is that 'the

estimator $\hat{\beta}_2$ is efficient (and consistent) of the true parameters'. If this is the case, the conclusion made is that, there should be no systematic difference between the two estimators. If there exists a systematic difference in the estimates, we then have reason to doubt the assumptions on which the efficient estimator is based. The Hausman statistic is distributed as χ^2 and is computed as:

$$H = (\beta_c - \beta_e)' (V_c - V_e)^{-1} (\beta_c - \beta_e) \quad (9.12)$$

where β_c is the coefficient vector from the consistent estimator, β_e is the coefficient vector from the efficient estimator, V_c is the covariance matrix of the consistent estimator and V_e is the covariance matrix of the efficient estimator. The 'degrees of freedom' for the statistic is the 'rank of the difference in the variance matrices'. When the difference is positive definite, it is the number of common coefficients in the models being compared.

Hausman test can be used to decide whether to use the RE model or the FE model. It states that: "if $\hat{\delta}_{RE}$ denotes the vector of RE estimates (without the coefficient on the time constant variable or aggregate time variables), and if $\hat{\delta}_{FE}$ denotes the corresponding FE estimates, then:

$$H = \left(\hat{\delta}_{FE} - \hat{\delta}_{RE}\right)^1 \left[A\hat{var}\left(\hat{\delta}_{FE}\right) - A\hat{var}\left(\hat{\delta}_{RE}\right) \right]^{-1} \left(\hat{\delta}_{FE} - \hat{\delta}_{RE}\right) \sim \chi^2_M \quad (9.13)"$$

The null hypothesis tested is H_0: assumptions of RE models hold'. If we reject H_0 then it implies that the FE model must be considered. If you fail to reject H_0, then the implication is that the RE model is better.

Check Your Progress 3 [answer within the space given in about 50-100 words]

1) What is a key consideration in deciding between the choice for the FE or RE model in panel data regressions? How is this achieved?

……

……

……

……

2) How is the test result determined in the Hausman's test?

……

……

……

……

……

3) Outline the procedure adopted in the Hausman's test.

……

……

……

……

……

9.5 SIMULTANEOUS EQUATION MODELS IN PANEL DATA

The need for considering the 'simultaneous equation models (SEM)' arise also in panel data contexts. In any regression modelling, generally, an equation is considered to represent 'a relationship' to describe a phenomenon. Many situations involve considering a 'set of relationships' to explain the behaviour of variables. For instance, consider labour supply for a group of married women already in the workforce. In place of the demand function, we can write the 'wage function' in terms of 'hours worked' and the other 'productivity variables'. With the equilibrium condition imposed, the two structural equations become 'labour supply function' and the 'wage function' equations. In such a case, in addition to allowing for simultaneous determination of variables within each time period, we can allow for the 'unobserved effect' in each equation. For instance, in a labour supply function, it would be useful to allow for an 'unobserved taste for leisure' that does not change over time.

The approach to estimating the SEMs with panel data involves two steps: (i) eliminate the 'unobserved effects' from the equations by using the fixed effects transformation (or first differencing) and (ii) find instrumental variables for the endogenous variables in the transformed equation. This can be very challenging because, for a convincing analysis, we need to find instruments that change over time. Let us now consider an SEM for panel data like:

$$y_{it1} = \alpha_1 y_{it2} + z_{it1}\beta_1 + c_{i1} + u_{it1} \qquad (9.14)$$

$$y_{it2} = \alpha_1 y_{it2} + z_{it2}\beta_1 + c_{i2} + u_{it2} \qquad (9.15)$$

where i denotes cross section, t denotes time period, and, z_{it1} and z_{it2} represents linear functions of a set of exogenous explanatory variables. The most flexible analysis allows the 'unobserved effects', c_{i1} and c_{i2}, to be correlated with *all* explanatory variables, including the elements in z. However, we assume that the idiosyncratic structural error terms, u_{it1} and u_{it2}, are uncorrelated with z in both the equations and across all time periods. This is in the sense of being truly exogenous. Further, except under very special circumstance, y_{it2} is correlated with u_{it1}, and y_{it1} is correlated with u_{it2}.

Now, let us consider the Equation (9.14). This cannot be estimated by the OLS since the composite error term $c_{it1} + u_{it1}$ is potentially correlated with all explanatory variables. Suppose we take their difference over time to remove the unobserved effect $c_{it1.}$ That is:

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \Delta z_{it1} \beta_1 + \Delta u_{it1} \tag{9.16}$$

Now, with the differencing (or time-demeaning), we can only estimate the effects of variables that change over time for at least some cross-sectional units. The error term in the above equation is uncorrelated with Δz_{it1} by assumption but Δy_{it2} is not necessarily uncorrelated with Δu_{it1}. Therefore, we need an 'instrumental variable (IV)' for Δy_{it2}. It is possible that such IVs comes from other equations. For instance, elements of Δz_{it2} that are not also in Δz_{it1} are natural IVs for Δy_{it2}.

Let us now consider the example of labour supply of married working women. With the equilibrium condition imposed, the two structural equations are:

$$\begin{aligned} hours = \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 \\ + \beta_{14} nwifeinc + u_1 \end{aligned} \tag{9.17}$$

and

$$\begin{aligned} \log(wage) = \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} \exp er \\ + \beta_{23} \exp er^2 + u_2 \end{aligned} \tag{9.18}$$

The variable *age* is the woman's age in years, *kidslt6* is the number of children less than six years old, *nwifeinc* is the woman's non-wage income (like husband's earnings), and *educ* and *exper* are years of education and prior experience respectively. All variables except *hours* and log(*wage*) are assumed to be exogenous. Note that this is a tenuous assumption since *educ* might be correlated with the omitted variable 'ability' in either equation. But for illustrative purposes, we ignore the omitted variable problem here. Note the functional form in this system (where *hours* appear in level form but *wage* in logarithmic form) which is popularly used in labour economics. This is for the simple reason that hours of work would change less but changes in wages could be sharper (or minute) requiring to be captured by taking their logarithmic values.

The first equation is the 'supply function'. It satisfies the 'order condition' because the two exogenous variables, *exper* and *exper*^{2}, are omitted from the labour supply equation. These exclusions are crucial restrictions in the sense that it amounts to assuming that 'once wage, education, age, number of small children, and other income are controlled for, past experience has no effect on current labour supply'. The 'wage equation' is also identified as at least one of *age*, *kidslt6*, or *nwifeinc* has a non-zero coefficient in (9.17). After differencing, the labour supply function becomes:

$$\Delta hours = \beta_0 + \alpha_1 \Delta \log(wage_{it}) + \Delta(other\ \ factors_{it})\beta_1 + \Delta u_{it} \qquad (9.19)$$

We can use $\Delta expeer_{it}$ as an instrument for $\Delta \log(wage_{it})$. However, because we are looking at people who work in every time period, $\Delta expeer_{it}$ = 1 for all *i* and *t* (since each person gets another year of experience after a year passes). We cannot therefore use $\Delta expeer_{it}$ as an IV, as it takes same value for all *i* and *t*. In a panel data set up, often, participation in an experimental programme is used to obtain IVs. Let us now consider another example of 'job training' and 'worker productivity' where we want to estimate the effect of 'another hour of job training' on 'worker productivity'. For any two years (e.g. 1987 and 1988), consider the simple panel data model:

$$\log(scrap_{it}) = \delta_0 + \beta_1 \Delta hrsemp_{it} + \Delta u_{it} \qquad (9.20)$$

where $scrap_{it}$ (the 'scrap rate' i.e. number of items out of 100 that must be scrapped) is regressed on $hrsemp_{it}$. Normally, we would estimate this equation by OLS. But Δu_{it} could be correlated with $\Delta hrsemp_{it}$. For instance, a firm might hire more skilled workers, while at the same time reducing the level of job training. In this case, we need an instrumental variable for $\Delta hrsemp_{it}$. Generally, such an IV is hard to find. But we can use the fact that in $\Delta hrsemp_{it}$, some firms received job training grants in one of the two years (1988). If we assume that grant designation is uncorrelated with Δu_{it} (something that is reasonable because the grants are given at the beginning of the year), then $\Delta grant_{it}$ is a valid IV provided $\Delta hrsemp_{it}$ and $\Delta grant_{it}$ are correlated. Job training and worker productivity are jointly determined, but receiving a job training grant is exogenous in Equation (9.20).

Let us consider another example of SEM with panel data. In order to estimate the causal effect of increase in prison population on crime rates at the state level, let us consider the 'instances of prison overcrowding litigation as instruments for the growth in prison population'. For the equation estimated in first differences, we can write an underlying 'fixed effects model' as:

$$\log(crime_{it}) = \theta_t + \alpha_1 \log(prison_{it}) + z_{it1}\beta_1 + c_{i1} + u_{it1} \qquad (9.21)$$

where θ_t denotes different time intercepts, and *crime* and *prison* are measured per 100,000 people. The prison population variable is measured on the last day of the previous year. The vector zit_1 contains log of 'police per capita, log of income per capita, the unemployment rate, proportions of black and those living in metropolitan areas, and age distribution proportions'. Differencing (9.21) gives the equation to be estimated as:

$$\Delta \log(crime_{it}) = \xi_t + \alpha_1 \Delta \log(prison_{it}) + \Delta z_{it1}\beta_1 + \Delta u_{it1} \qquad (9.22)$$

Simultaneity between 'crime rates' and 'prison population' (or more precisely in their growth rates), makes the OLS estimation of (9.22) inconsistent. Hence, we can estimate (9.22) by the 'pooled 2SLS method'. After the pooled 2SLS estimation, we obtain the residuals $\hat{r}_{it1}$. Then, by

including one lag of these residuals in the original equation, we can estimate the equation by 2SLS with $\hat{r}_{it1}$ acting as its own instrument. The first year is lost because of the lagging. Hence, the usual 2SLS *t* statistic on the lagged residual is a valid test for serial correlation. Let us now consider another example of a panel data model to examine the effects of cigarette smoking on earnings as:

$$\log(wage_{it}) = z_{it}\gamma + \delta_1 cigs_{it} + c_i + u_{it} \qquad (9.23)$$

We want to know the causal effect of smoking on hourly wage. For concreteness, we assume $cigs_{it}$ is measured as 'average packs per day'. This equation has a causal interpretation i.e. holding the factors in z_{it} and c_i fixed, what is the effect of an exogenous change in cigarette smoking on wages? Thus, Equation (9.23) is a structural equation.

The presence of the individual heterogeneity, c_i, in Equation (9.23) recognises that cigarette smoking might be correlated with individual characteristics that also affect wage. An additional problem is that $cigs_{it}$ might also be correlated with u_{it}. In this example, the correlation could be from a variety of sources. But, simultaneity is one possibility because if cigarettes are treated as a normal good, then, as income increases, holding everything else fixed, cigarette consumption increases. Therefore, we might add another equation to Equation (9.23) that reflects $cigs_{it}$. It may depend on income which depends on wage. However, if Equation (9.23) is of interest, we do not need to add equations explicitly. But we need to find some instrumental variable. To get an estimable model, we must first deal with c_i since it can be correlated with z_{it} as well as $cigs_{it}$. We can use the FE method to eliminate c_i before addressing the correlation between $cigs_{it}$ and u_{it}. The approach that can be used to estimating the SEMs with panel data is to use the fixed effects transformation and then to apply an IV technique such as pooled 2SLS. A simple procedure is to estimate the time-demeaned equation by pooled 2SLS like:

$$\ddot{y}_{it1} = \alpha_1 \ddot{y}_{it2} + \ddot{z}_{it1}\beta_1 + \ddot{u}_{it1}, \; t=1,2,....,T \qquad (9.24)$$

where $\ddot{z}_{it1}$ and $\ddot{z}_{it2}$ are IVs.

Check Your Progress 4 [answer within the space given in about 50-100 words]

1) What are the two factors to be considered in the approach to estimate the SEMs in panel data regressions?

……

……

……

……

2) How is the 'unobserved effect' removed from a composite error term like $c_{it1} + u_{it1}$ in Equation (9.14)? What is a consequent disadvantage of this approach and how is this dealt with?

……

……

……

……

……

9.6 LET US SUM UP

The unit introduces the methods of estimating the panel data models. Two cases of 'random effects' and 'fixed effects' are separately discussed. The set of assumptions governing the estimation process are listed. The method of converting or transforming the data into 'quasi de-meaned data' is indicated. For making a choice between the 'fixed effects' model and the 'random effects' model, the procedure for applying the Haussman's test is outlined. Since situations of simultaneity affects the estimation process in panel data contexts, methods of dealing with such cases is explained with many illustrations.

9.7 KEY WORDS

Linear Static Panel Data Model : It is a model like in Equation (9.1). In this, a separate term for 'unobserved effects', c_i, assumed to be random variables, is introduced. Further, in order to be able to assume the effect of c_i to be 'zero' on an average, we introduce a separate intercept term as in Equation (9.2).

Quasi-Demeaned Data : Refers to the procedure adopted for transforming the variables to remove the effect of their means. It is a partial (quasi) cleaning process since the 'random effects transformation subtracts a fraction of the time average effect'.

Fixed Effects Estimator : Refers to a 'pooled OLS estimator' based on the time-demeaned variables' used in the panel regression models.

Hausman Test : This is a test procedure proposed by Hausman which helps us to make a choice between a 'random effects' model and a 'fixed effects' model. The test statistic [as in Equation (9.12)] is distributed as $\chi 2$ with the d.f. equal to 'the rank of the difference in the two covariance matrices' i.e. the covariance matrix of the 'consistent estimator' and the 'efficient

estimator'. It is based on the difference between the random effects and fixed effects estimates.

9.8 SUGGESTED BOOKS FOR FURTHER READING

1) Wooldridge J M (2006). Introduction to Econometrics: A Modern Approach. *Michigan State University. USA*.

2) Greene W H (2016). *Econometrics*. Prentice Hall.

9.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A linear static panel data model is like $Y_{it} = \beta_0 + X_{it}\,\beta_i + c_i + u_{it}$. Here X_i's are fixed or pre-determined (and hence the model called as the 'static' model) corresponding to the value of Y in period t. The c_i's stand for the 'unobserved effect' and u_{it}'s are the 'idiosyncratic error terms'. The term 'idiosyncratic' refers to the random character of the observation-specific 'error term'. The zero mean characteristic of c_i's is achieved by the inclusion of the explicit intercept term β_0.

2) The OLS method cannot be applied because of the composite error term $v_{it} = c_i + u_{it}$ which implies that the error terms are serially correlated across time. In such cases, we have to apply the GLS method of estimation with 'auto regressive serial correlation'.

3) The assumptions required for RE model are: (i) the model is defined as in equation (9.3) with a term for 'unobserved effect' (c_i) and 'random error term' (u_{it}), (ii) explanatory variables change over time but not with perfect collinearity, (iii) the expected value of error terms (u_{it}) given X_i and c_i is 'zero', (iv) *var* (u_{it}) = σ_u^2 (v) the covariance of uncorrelated error terms conditional to x_i and c_i is 'zero', (vi) $E(c_i) = 0$ and (vii) $\text{var}(c_i \mid x_i) = \sigma_c^2$ i.e. the assumption of homoscedasticity hold for c_i's.

4) The term 'quasi' means 'partial'. The term 'demeaned' can be split as 'de-meaned' which means removing the effect of mean. Hence, the term 'quasi demeaned data' refers to the partial adjustment made in each variable to remove the effects of their means. In effect, it is a partial cleaning process attained by the estimation process of GLS.

Check Your Progress 2

1) It means the two are correlated with each other. In mathematical terminology, it means that they are not parallel (when they will never meet or cross) but perpendicular and hence intersecting with each other.

2) Firstly, in terms of the effect on the estimated parameters. They are in the first stage consistent and in the later stage 'best' i.e. BLUE. With the additional assumption on the normality of u_{it}'s (conditional on x_i and c_i) i.e. making the nature of error terms i.i.d., the t and F statistics have the exact distributions without the need for approximation of large N and small T.

3) By a process of 'time demeaning' transformation for removing the effect of 'average' in Equation (9.6). The transformation eliminates the 'unobserved effect' c_i as in the transformed Equations (9.8) or (9.9). The estimation of (9.9) becomes the 'pooled OLS' method. A major difference between the FE and the RE methods is that, FE is consistent when c_i and x_{it} are correlated, whereas the RE estimate is not.

4) $N(T-1)-k$.

5) The pooled OLS's standard errors and test statistics are generally invalid, because, they ignore the serial correlation in the composite errors: $v_{it} = c_i + u_{it}$.

Check Your Progress 3

1) It is 'whether c_i and x_{it} are correlated'. It is achieved by the Hausman test. It is based on the difference between the random effects and fixed effects estimates.

2) A statistically significant difference 'between the random effects and fixed effects estimates' is interpreted as 'evidence against the random effects'.

3) Hausman's specification test compares an estimator $\hat{\beta}_1$ to be consistent with an estimator $\hat{\beta}_2$ that is efficient. The null hypothesis it tests is that 'the estimator $\hat{\beta}_2$ is efficient (and consistent) of the true parameters'. Hausman statistic is distributed as χ^2 and is computed as: $H = (\beta_c - \beta_e)'(V_c - V_e)^{-1}(\beta_c - \beta_e)$ where β_c is the coefficient vector from the consistent estimator, β_e is the coefficient vector from the efficient estimator, V_c is the covariance matrix of the consistent estimator and V_e is the covariance matrix of the efficient estimator.

Check Your Progress 4

1) Eliminating the 'unobserved effects' from the equations by using the 'fixed effects transformation' and (ii) finding instrumental variables for the endogenous variables in the transformed equation.

2) By considering the difference equation over time as in Equation (9.16). The error term in (9.16) is uncorrelated with Δz_{it1} (by assumption) but Δy_{it2} is not necessarily uncorrelated with Δu_{it1}. As a result, we need an 'instrumental variable (IV)' for $\Delta y_{it2.}$