Panel Data Regression Modeling

Dr. Subrata Roy
Associate Professor
Department of Commerce, MGCUB

A. Today I am starting with the basic thing like research types. Research type can be divided into 3 forms namely (1) Based on research Purpose (2) Based on the research process (3) Based on research outcome and (4) Based on research logic.

1. Based on research purpose: *(i). Descriptive research* (It includes surveys and fact-finding enquiries of different kinds. The major purpose is generally to describe an event or a relationship) *(ii). Exploratory research* (The objective is the development of hypothesis rather than their testing) *(iii) Analytical explanatory research* (It is an extension of the descriptive research. Here, the researcher not only describes the issue, but also analyses and explains the reasons why or how the issue being studied happened) *(iv) Predictive research* (It aims to generalize findings from the analysis conducted on the basis of hypothesized, general relationships and provides explanations for interactions or interchanges in a specific and general situations)

2. Based on the Research Process: Some researcher adopts various approaches like qualitative approaches and quantitative approaches to address their research problems and design a study that involves collecting qualitative data or quantitative data. Qualitative and quantitative research can be used together in the same research such as;

a. collect data through qualitative techniques and quantify them by counting the frequency of occurrence of specific key words or themes in order to employ statistical methods.

b. collect qualitative data and analyze them by non-numerical methods or

c. collect numerical data and use statistical methods to analyze them

3. Based on the research outcome: *(i). Applied research* (It is a study that has been conducted in order to apply its findings to solve a specific or existing problem. It is the application of available knowledge to improve mgmt practices and policies. *(ii). Basic Research* (when the research problem is of less specific nature and the research is being conducted primarily to improve our understanding of some general issues without emphasis on its immediate application)

4. Based on the research logic: *(i). Deductive research* (Here, theoretical framework is developed and tested through empirical observations and particular instances are deducted from general inferences. So, this method of research moving from general to specific) *(ii). Inductive Research* (It represents type of study in which general inferences are induced from particular instances. Individual observations are used

to make general statements and it is referred to as moving from specific to general) *(iii). Conceptual research* (it is related to some abstracts or theory and generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones) *(iv). Empirical research* (It relies on experience or observation alone, often without seeking for system and theory. It is data based research, resulting in conclusions which are capable of being verified by observation or experiment, thus, it can be named as experimental research) (v). Historical research (It utilizes historical sources such as documents, remains etc. in order to study events or ides from the past, including the philosophers of persons and groups at any remote point of time)

B. Now come to various models. In a nutshell model means by which research problem can be solved. It may be economic models or econometrics models. *Economic models* help the researcher to rationally and logically support the hypothesis or answer predetermined research questions. Economic models can be divided in various forms like Visual model, Mathematical models, empirical models, simulation models, static models and dynamic models. On the other hand, *Econometrics* is a science and art of using economic theory and statistical techniques to analyze economic data. An econometric model should be constructed based on economic theory, experience or critical thinking. So econometrics aims to explore relationship between economic variables and interpret the results obtained through statistical techniques. Econometrics models can be classified in different forms like (i). The models developed to find out relationships between past and present (ii). The models that examine relationship between economic variables over time (iii). The models which investigate the relationship between different variables measured at a given point of time (iv). The models which consider the relationship between different variables for different units over time.

C. Now I shall discuss about data which is a piece of information or knowledge which are used for reasoning or calculation as measurement. Data may be various types:

i. Qualitative data

ii. Quantitative data

iii. Primary data

iv. Secondary data

v. Experimental data

vi. Observational data

vii. Time series data

viii. Cross sectional data

ix. Panel data or pooled data or longitudinal data or micro panel data

D. Definition

- The panel data have both cross-sectional and time series dimensions. Cross-sectional dimension of panel data implies that the data have been collected from a number of cross-sectional units. It may be country, state, district, city, village, firms, and individuals and so on. The time-series dimension of panel data is reflected by the fact that the data have been collected for chosen cross-sectional units for more than one time period. Here, the time point may refer to year, quarter, month, week and even day. Thus, to generate a panel data set, we are required to collect data (either from secondary sources or from primary sources or combination of both) for a number of time points for a set of cross-sectional units. Finally, the data so collected are presented in a systematic manner to generate the panel data set.

- Alternative names of panel data → pooled data, micro panel data (which refer specifically to individuals, firms and households), longitudinal data etc.

Types of Panel Data

- Balanced and unbalanced panels

  Balanced panel → each cross-sectional unit has same number of time series observations.

  If, I consider 5 countries like Japan, India, China, Brazil and Korea (N) and also consider 7 years (T) from 2001 to 2007. Then number of observation will be (N * T) = 5 * 7 = 35

  Unbalanced panel → the number of time series observations (T) differs between cross-sectional units (N).

Sometimes a distinction is also made between short panel and long panel:

Short panel → N > T in that case REM is more appropriate (Greene 2003, 283)

Long panel → T > N then FEM becomes more appropriate

Advantages usefulness of Panel Data:

1. It can be used to deal with heterogeneity among the cross-sectional units. Here, heterogeneity means that the cross-sectional units are all different from one another. In reality, a number of unmeasured (or unobserved) explanatory variables generate such heterogeneity among the cross-sectional units. If you don't consider such variables then causes biased estimation of unknown parameters of the model. Actually, the models that are estimated by using panel data are so specified that heterogeneity among the cross-sectional units is taken care of.
2. Combining time series with cross-sections it enhances of both quantity and quality of data. In this process, panel data create more variability in the data, less collinearity among the variables, and more degrees of freedom. As a consequence, more efficient estimation of unknown parameters of the model becomes possible.

3. Panel data makes possible examination of issues that cannot be studied using cross-sectional or time series data alone. In particular, panel data sets are better able to study complex issues of dynamic behavior. For, example with a cross-section data one can estimate the rate of unemployment at a given point of time. Repeated cross-sections can show how this proportion changes over time. But, panel data set alone can estimate what proportion of those who are unemployment in one period remained unemployed in another period.

4. Panel data allow us to estimate more complicated behavioural models which are not possible by using only cross-sectional or time series data sets. For example, Phenomena, such as economies of scale and technological change can be studied better by panel data rather by pure cross-sectional or time series data.

5. Panel data are able to measure effects that are not detectable in pure cross-sectional or time-series data

Example → to know whether union membership enhances wages of workers, we are required to analyze data on wages of a given set of workers at different points in time (e.g., before and after obtaining union membership). In a panel data regression model, holding individual workers' characteristics

Disadvantages of panel data → lengthy collection process, expensive, and prone to the problem of attrition/destroy due to 'sample decay' or 'non- response' from some cross-sectional units at some points in time.

Panel Data Models:

Suppose we want to examine inter-country variation in the Economic growth (GDP) due to foreign trade (EXPORT & IMPORT).

We collect Country-wise data for 7 years on GDP and some of its determinants such as EXPORT and IMPORT. [see Table 1].

Table 1: Incidence of Rural Poverty, Percentage of Rural Workers in Non-farm Employment and Rural Literacy Rate in 17 Major States of India

Country	*Year*	*GDP*	*IMP*	*EXP*	*Country*	*Year*	*GDP*	*IMP*	*EXP*
Japan	**2001**	27.3	19.9	29.5	**Japan**	**2006**	40.4	10.1	41.1
India	2001	41.9	20.1	41.8	India	2006	37.7	17.6	59.5
Brazil	2001	64.9	16.4	29.0	Brazil	2006	50.1	19.2	49.3
China	2001	27.9	14.7	45.6	China	2006	13.7	25.4	56.0
Korea	2001	21.8	22.7	39.9	Korea	2006	26.9	20.1	36.1
Japan	**2002**	17.8	12.9	51.6	**Japan**	**2007**	33.0	29.6	57.8
India	2002	25.2	25.3	28.8	India	2007	42.3	20.0	40.8
Brazil	2002	37.5	15.7	39.8	Brazil	2007	37.4	36.6	54.0
China	2002	38.5	37.1	82.2	China	2007	10.9	28.4	65.7
Korea	2002	48.2	9.1	30.7	Korea	2007	23.1	25.3	66.1
Japan	**2003**	45.0	14.2	47.8					
India	2003	67.5	20.8	39.4					
Brazil	2003	14.3	17.2	44.1					
China	2003	37.7	13.2	24.1					
Korea	2003	56.2	25.7	47.1	Here, N = 5 (Countries)				
Japan	**2004**	46.4	17.5	30.2	T = 7 (Years)				
India	2004	61.6	26.3	42.4	Total Obs. = 5 x 7 = 35				
Brazil	2004	16.6	20.6	40.5	It’s a balanced panel.				
China	2004	44.4	20.8	52.4					
Korea	2004	57.2	15.6	36.5					
Japan	**2005**	22.4	21.3	54.6					
India	2005	26.6	28.1	53.5					
Brazil	2005	29.3	19.8	65.2					
China	2005	19.7	24.0	39.1					
Korea	2005	30.2	18.7	50.8					

Now, start with the functional form like that, GDP is a function of foreign trade (or a function of import and export) corresponding to data in Table 1:

We can write it as $$GDP = f(IMP, EXP) \quad (1)$$

Or we can write the above function in functional form or mathematical form or economic form as below:

$$GDP_{it} = \alpha + \beta_1 IMP + \beta_2 EXP \quad (2)$$

Where, GDP is affected by foreign trade with-out any disturbances. But, in reality there must be some error or disturbances which are not consider in the above model so we have to consider it and then the econometric model which we are going to estimate as under:

$$GDP_{it} = \alpha_i + \beta_1 IMP + \beta_2 EXP + e_{it} \quad (3)$$

a. All the variables have *it* as subscript → *i* and *t* refer to cross- sectional and time series aspects of data, respectively. Here, i = 1, 2, 3, 4, 5 and t = 1, 2, 3, 4, 5, 6, 7. We have altogether 35 observations on each variable ($NT = 35$).

b. e_{it} is the disturbance term which is independently and identically distributed [$e_{it} \sim$ i.i.d. $(0, \sigma^2)$]

c. For the purpose of hypothesistesting, we assume that e_{it} is normally distributed with 0 mean and constant standard deviation.

d. Finally, all explanatory variables of the model are assumed to be non-stochastic and none of these is correlated with the disturbance term.

Estimation of model (3) ⟶ three different approaches:

(i) The Constant Coefficients Model (CCM),

(ii) The Fixed Effects Model (FEM), and

(iii) The Random Effects Model (REM).

The Constant Coefficients Model (CCM)

Assumption → homogeneity → all coefficients (intercept and slopes) remain unchanged across cross-sectional units, and over time

CCM ignores the space and time dimensions of panel data set.

Accepting this homogeneity assumption, the CCM applies the OLS method to estimate unknown parameters of themodel

CCM is nothing but straightforward application of OLS to a given panel data set to obtain estimates for unknown parameters of the model

Limitations of CCM:

- In reality, the homogeneity assumption may not be true.
- Different cross-sectional units may have different values for intercept and/or slope coefficients.

Why so?

Suppose there are two categories of countries → Developed and poor countries → Rich countries economic performance equation estimated by using data for developed economics would be different from the equation for poor countries.

Reason → apart from EXPORT and IMPORT, there are many other factors (e.g., stock market condition, crude oil prices, foreign exchange rate, Gold prices etc.) that cause low incidence of economic performance in some countries compared to others.

The economic performance function may also change over time following a major policy intervention by the government.

Implication → ignoring the differences between the countries, and estimating a single economic performance equation with pooled data is likely to provide distorted results (i.e., incorrect estimates for unknown parameters of the model).

Here one may argue that the disturbance term captures the effect of omitted variables, if any.

But some of those omitted variables are also likely to be correlated with the explanatory variables included in the model.
For example, an omitted variable like the foreign exchange and crude oil prices are likely to be correlated with import or export. In this situation, as the effects of these variables are captured by the disturbance term of the model, $Cov(Export_{it}, e_{it}) \neq 0$ and thus violation of one important OLS assumption and if you estimate the model 3 the estimation would be biased and inconsistent for the unknown parameters.

So, this problem can be avoided if we follow the FEM or the REM.

These models make a more rational specification of the model so that the heterogeneity among the cross-sectional units is explicitly recognized, although their methods of doing so are different.

The Fixed-Effects Model (FEM)

Consider another example → our objective is to examine the relationship between wages of workers and their education (years to schooling).

Assuming that the relationship between wages and education is same for all workers, we may formulate the following CCM:

$$WAGE_{it} = \alpha + \beta EDU_{it} + \varepsilon_{it} \qquad (4)$$

This model assumes that the values of α and β are same for all workers. This is an

unrealistic assumption for the following reason:

α represents the wages when EDU is zero.

Put differently, α is the benchmark from which a worker's wages develop. This will usually be different for different workers.

Example → holding education constant, the male worker probably obtains higher wages than the female worker → so value of intercept would vary across individuals.

Illustration: suppose that we gathered data on wages and education of two workers (one male and one female) for a period of 15 years. Such data are plotted in figure 1.

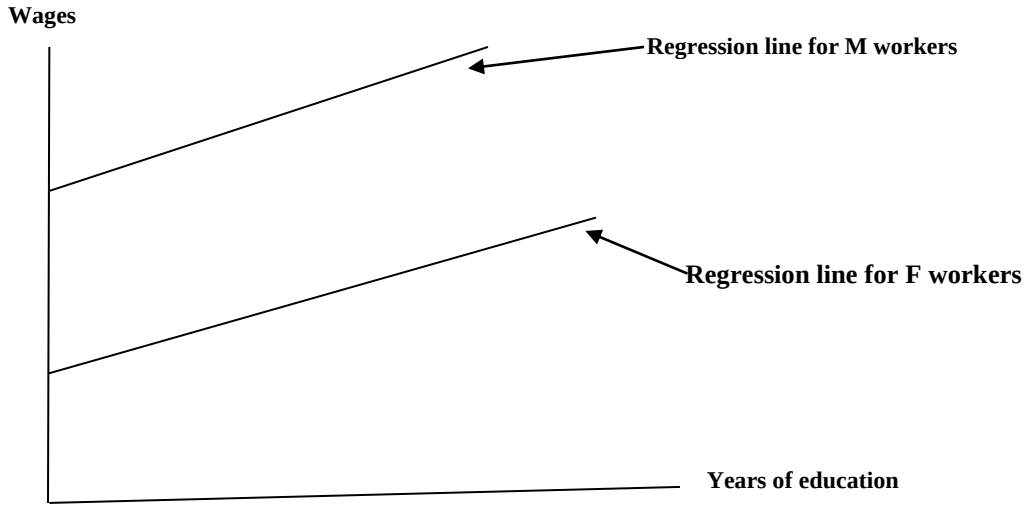

Fig.1 Relationship between wages & education of two workers

Corresponding to the scatter plots of data for two workers, we can visualise two separate regression lines → at every level of education, the regression line of the male worker lies above the regression line of the female worker → even with same level of education, the male worker gets higher wages than the female worker → the value of intercept is different for two workers (although the slopes are same).

If we ignore this reality and estimate a single wage function (equation 4) by using the pooled data set, we will have incorrect estimate of the intercept parameter of our model.

A more difficult situation arises if the data points for two workers are scattered in the manner shown in Figure 2.

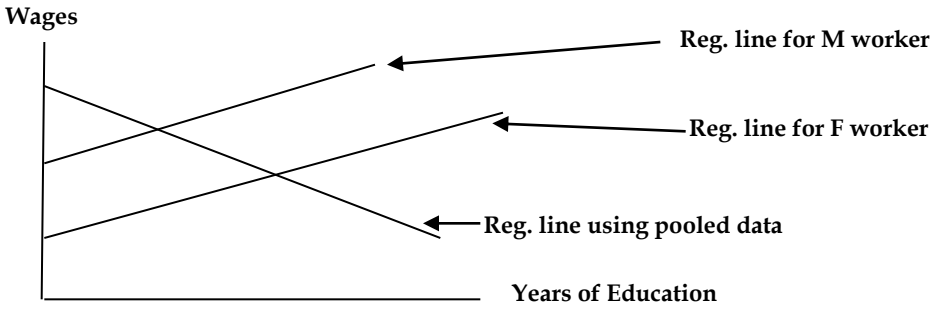

Fig. 2 Relationship between wages & education of two workers

Here estimating two separate regressions gives different intercept values for two workers but the value of slope coefficient (β) is same and positive, which leads to the conclusion that returns to education is positive.

However, if we pool all observations, the regression line corresponding to pooled data gives not only a different intercept but also turns the slope coefficient negative → we have a highly misleading conclusion that returns to education is negative.

Implication of above discussion → if different individuals have different regression lines (i.e., 'individual effects' are present) we must take this into consideration to avoid such a misleading conclusion.

In other words, the heterogeneity feature of the cross-sectional units (individuals) must be accommodated explicitly/clearly in the regression model while working with panel data set.

So, **FEM** does this by considering the intercept as a variable and uses ***dummy variables*** (D) to account for differences among the individuals with regard to the value of intercept.

The FEM model is written as

$$\mathbf{WAGE_{it} = \alpha_i + \beta EDU_{it} + e_{it}} \quad \textbf{(4)}$$

$$\mathbf{WAGE_{it} = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \ldots + \alpha_N D_{Nt} + \beta EDU_{it} + e_{it}} \quad \textbf{(5)}$$

The value of $D_{1t} = 1$ for first cross-sectional unit and 0 for all others, $D_{2t} = 1$ for second cross-sectional unit and 0 for all others and so on. In this situation, α_1 is the intercept of first cross-sectional unit, α_2 is the intercept of second cross-sectional unit, and so on. So, model 4 and 5 is called FEM because although the intercept term is allowed to vary across cross-sectional units, for each cross-sectional unit the intercept is assumed to remain constant over time i.e., the intercept of each cross-sectional unit is time invariant.

The model such as (3) or (4) is called FEM because although the intercept term is allowed to vary across cross-sectional units, for each cross-sectional unit the intercept is assumed to remain constant over time, i.e., the intercept of each cross-sectional unit is time-invariant.

The FEM is estimated by using the OLS method and thus FEM is also known as FELSDV or simple LSDV.

As FEM is simple a regression model we generally apply t test to examine whether an individual intercept is 0 or not.

Now we have to test the appropriateness of the FEM (equation 5) against CCM (equation 4) by applying restricted F-test.

The Restricted F-test

The null hypothesis is $H_N: \alpha_1 = \alpha_2 = ... = \alpha_N$

To test the validity of H_N, we compute

$$F^* = \frac{(R^2_{FEM} - R^2_{CCM})/(N-1)}{(1-R^2_{FEM})/(NT-N-k)} \sim F(N-1, NT-N-k)$$

Where,

R^2_{FEM} = computed R^2 value from the estimated FEM (called unrestricted regression);

R^2_{CCM} = computed R^2 value from the estimated CCM (restricted regression);

N = number of intercepts in FEM (equal to number of cross-sectional units);

NT = total number of observations; and

k = number of explanatory variables in the FEM.

Decision rule: If $F^* > F_\lambda(N-1, NT-N-k)$, reject H_N and conclude that compared with the CCM, the FEM is more appropriate in the context of our pooled or panel data set. This indicates that fixed effects are present and the intercepts of cross- sectional units are statistically significantly different from each other.

One-Way and Two-Way Fixed Effects Models:

Model (5) is known as One-Way FEM as we have allowed only the intercept to vary between cross-sectional units.

But we may also consider the period-effect if we believe that the intercept changes uniformly for all cross-sectional units over time (e.g., due to enforcement of minimum wages act by the government).

To account for such a period-effect, we add period dummies in model (5). The number of period dummies is determined by number of periods (T) for which data have been collected.

The model then is called Two-Way FEM that considers both individual (cross-section) and period (time) effects.

Further, we may consider variation in the slope coefficient among cross-sectional units as well as over time by introducing 'interaction or differential slope dummies' which are obtained by interacting N cross- section intercept-dummies and T time-dummies with the explanatory variable (EDU).

Limitations of FEM:

1. Using too many dummy variables → low degrees of freedom → imprecise estimates for unknown parameters of our model.
2. Many dummy variables → possibility of multicollinearity problem.
3. As $\varepsilon_{it} \sim N(0, \sigma^2)$ and the index i refers to cross-sectional units, and

   t to time series observations, ε_{it} is likely to suffer from problems of heteroskedasticty and autocorrelation.

So you have to eliminate those problems and thus Within-Group and First-Difference Estimators

First two of above-mentioned problems can be avoided by adopting a simple trick → either express all the variables (including dummy variables) as deviations from their respective means or consider first-difference of the variables in the FEM.

The former approach is known as Within-Group Estimator (WGE) while the latter is known as First-Difference Estimator (FDE).

Such transformations of the variables are performed separately for each cross-sectional unit.

Advantage of WGE/FDE → the terms involving dummy variables are removed from the FEM and this helps to attain maximum economy of degrees of freedom, and avoid the problem of multicollinearity.

Disadvantages of WGE/FDE

(i) While transforming the variables, apart from dummy variables, the time-invariant explanatory variables whose values remain unchanged for a given cross-sectional unit at different points in time (e.g., the workers' caste, gender, parental education etc.) get wiped out from the model. So, FEM is not applied if there are time-invariant variables in the data set.

(ii) Differencing leads to removal of long run effects in the variables and we are left with only the short run effects in the variables; and

(iii) The disturbance term of the Fixed Effects First-difference Model suffers from autocorrelation problem.

So, these problems are avoided if we apply theREM.

The Random Effects Model (REM)

REM doesn't use dummy variable to capture the presence of individual effect. It assumes that the individual effect (α_i) is a random variable with a mean value α_1. Then the i^{th} cross-sectional can be expressed as:

$$\alpha_i = \alpha_1 + \mu_i \qquad \textbf{(6)}$$

Here we assume that the cross-sectional units have been drawn from a much larger universe (population) and they have a common intercept α_1.

The individual differences in the intercept values of cross-sectional units is captured by the random error term u_i.

So the REM is specified as:

$$WAGE_{it} = (\alpha_1 + u_i) + \beta EDU_{it} + \varepsilon_{it} \qquad \textbf{(7)}$$

$$= \alpha_1 + \beta EDU_{it} + w_{it}$$

Here, w_{it} $(=\mu_i+\varepsilon_{it})$ **is the composite random error**

Term that has two components:

(i) u_i represents the cross-section or individual-specific random error component

(ii) ε_{it} is the combined time series and cross-section random error component, sometimes called the *idiosyncratic random term* as it varies over cross- sectional units as well as over time.

As this model considers individual effects (α_i) as a random variable, hence the name is REM.

The assumptions with regard to u_i and ε_{it} are:

$$u_i \sim N(0, \sigma_u^2)$$

$$\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$$

$$E(u_i \varepsilon_{it}) = 0$$

$$E(u_i u_j) = 0 \quad for\ i \neq j$$

$$E(\varepsilon_{it} \varepsilon_{is}) = E(\varepsilon_{ij} \varepsilon_{jt}) = E(\varepsilon_{it} \varepsilon_{js}) = 0 \quad for\ i \neq j\ and\ t \neq s$$

Here, $\sigma_u^2 = Var(u_i)$ **and** $\sigma_\varepsilon^2 = Var(\varepsilon_{it})$

These assumptions imply that individual error components are not correlated with each other and are not correlated across cross-section and time series units. Using these properties of μ_i and ε_{it}, we can find out the properties of w_{it}.

$$E(w_{it}) = 0$$

$$Var(w_{it}) = \sigma_u^2 + \sigma_\varepsilon^2$$

w_{it} has zero mean and constant variance (homoskedastic).

But w_{it} and w_{is} are correlated (i.e., the composite error term of a given cross sectional unit at two differed times are correlated).

The value of such a correlation coefficient (ρ) is given by:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \quad for \quad t \neq s$$

- If we ignore this correlation, and estimate the REM by OLS method, the resulting estimators will be inefficient.

- The most appropriate method to estimate the REM is the GLS. But for GLS to have good properties, we should have large N and relatively small T.

- Another important point that needs mention is that REM supposes the composit error term <u>not to be correlated</u> with any of the explanatory variables.

  However, since u_i is a component of w_{it}, sometimes w_{it} may be correlated with the explanatory variables included in the REM (This is because u_i captures the effect of omitted variables some of which may be correlated with the explanatory variable(s) of the REM.)

  In such a situation, REM will result in inconsistent estimation of regression coefficients, and we may have to depend on the FEM.

Whether or not w_{it} is correlated with the explanatory variables in a given application is checked through the Hausman test.

- Equation (7) represents One-way REM that considers only the cross- section effect. we may be extended equation 7 to allow for a random error component to take into account the period effect. The model then becomes Two-way REM.

How you will estimate REM

- The EViews package uses GLS method to estimate the REM. So, GLS is a technique for estimating the unknown parameters in a linear regression model when there is a certain degree of correlation between the residuals in a regression model.

  After estimation of the REM, testing of hypothesis involving regression coefficients are performed in the same manner as for any other regression model.

  The EViews conducts hypothesis tests (e.g. t- and F-tests) in standard ways using the appropriate estimate of the variance of the GLS estimator.

Choosing Between the FEM and REM: The Hausman Test

Rule of Thumb: REM is more suitable when the number of cross- sectional units (N) is large and number of time series observations (T) is small. Why so?

An perceptive/intuitive explanation of this is that (α_i) as the intercept in the REM is a random variable, it must be allowed to assume a wide spectrum of values over ($-\infty, \infty$). This is possible when N is sufficiently large. Thus, REM does not suit a data set satisfactorily with fewer cross-sectional units (i.e., long panels). In such a situation, the FEM that involves lesser number of dummy variables (as N is small) appears suitable. The FEM also enjoys computational convenience compared with the REM.

- However, selection between the FEM and REM is performed more rigorously by applying the Hausman test.

As pointed out earlier, REM is not preferred if the composite error term wit gets correlated with the explanatory variable (s) of the model, which at times becomes a possibility. Hausman adapted a test based on the idea that if there is no-correlation between Wit and explanatory variable (s) both OLS and GLS are consistent but OLS is insufficient. On the other hand, if such correlation exists, OLS is consistent but GLS is not. More specifically, Hausman assumed that there are two estimators $\hat{\beta}^{FE} and \hat{\beta}^{RE}$ of the parameter vector β and added two-hypothesis testing procedures. They are:

$$H_N: Both\ \hat{\beta}^{FE}\ and\ \hat{\beta}^{RE}\ are\ consistent\ ,\ but\ \hat{\beta}^{FE}\ is\ inefficient$$

$$H_A: \hat{\beta}^{FE}\ is\ consistent\ and\ efficient\ ,\ but\ \hat{\beta}^{RE}\ is\ inconsistent$$

- Here we actually test H_N (random effects are consistent and efficient) against H_A

(random effects are inconsistent, as the fixed effects will always be consistent).

- Hausman takes $\hat{q} = (\hat{\beta}^{FE} - \hat{\beta}^{RE})$ as the basis for the relevant test statistic. Then the Hausman test statistic is given by:

$$H = \hat{q}' [Var(\hat{\beta}^{FE}) - Var(\hat{\beta}^{RE})]^{-1} \hat{q} \sim \chi^2(k)$$

where k is the number of explanatory variables.

The decision rule is: If computed value of chi-square (χ^2) is greater than the theoretical/table value of chi-square (χ^2) at a chosen level of significance say 5% or 1% or 10% and k degrees of freedom, i.e., $\chi 2$ (computed)$>\chi^2\lambda(k)$, we reject H_N meaning that REM is consistent and accept the FEM estimator. In contrast, we don't reject HN if $\chi 2 \leq \chi^2\lambda(k)$ and prefer the Random effects estimator.

References

Asteriou, Dimitrios and Stephen G Hall (2007), *Applied Econometrics: A Modern Approach Using EViews and Microfit*, Palgrave Macmillan, New York.

Baltagi, Badi H (2008), *Econometric Analysis of Panel Data*, John Wiley & Sons, United Kingdom, 4th Edition.

Wooldridge, Jeffrey M (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, MA.