International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Deriving Some Estimators of Panel Data Regression Models with Individual Effects

Megersa Tadesse Jirata[1], J. Cheruyot Chelule[2], R. O. Odhiambo[3]

[1]Pan African University Institute of Basic Sciences,
Technology and Innovation P.O. Box 62000 – 00200 Nairobi, Kenya

[2, 3]Department of Statistics and Actuarial Sciences
Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000 – 00200 Nairobi, Kenya

Abstract: ***The panel data models are becoming more common in relation to cross-section and time series models for innumerable present advantages, in addition to the computational advance that facilitated theirs utilization. The existing literature has focused on the application of the estimators. Many of panel data model estimators are severely inconsistent due to presence of endogeneity and heterogeneity problem. Hence, panel data offers various opportunities to derive estimators of those result consistent estimators. This paper considers estimation of linear panel data models with fixed effects and random effects when the equation of interest contains unobserved heterogeneity as well as endogenous explanatory variables. We offer a detailed analysis and derivation of the two-stage least squares (2SLS) and generalized least square (GLS) estimators in the context of panel data models.***

Keywords: Panel data, Fixed effects, Random effects, Two-stage least square and Generalized Least Square.

1. Introduction

Panel (or longitudinal) data is a kind of data in which observations are obtained on the same set of entities at several periods of time. It refers to the data with repeated time-series observations (T) for a large number (N) of cross-sectional units (e.g., states, regions, countries, firms, or randomly sampled individuals or households, etc). Since the panel data relate to these units over time, presence of heterogeneity in these units is a natural phenomenon. The techniques of panel data estimation can take such heterogeneity explicitly into account by allowing for individual specific variables. If individual heterogeneity is left completely unrestricted, then estimates of model parameters suffer from the incidental parameters problem, noted by creel (2014). This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which in turn, bias and inconsistent estimates of model parameters. An important advantage of using such data is that they allow researchers to control for unobservable heterogeneity, that is, systematic differences across cross-sectional units. Regressions using aggregated time-series and pure cross-section data are likely to be contaminated by these effects, and statistical inferences obtained by ignoring these effects could be seriously biased and inconsistent.

The two most widely applied panel data model estimation procedures are random effects (RE) and fixed effects (FE).It is well-known that the consistency of the RE and FE estimators requires the strict exogeneity of the regressors, but the strict exogeneity assumption generates many more moment conditions than these estimators use. Hence, problems that generally afflict fixed effect model (i.e. endogeneity) and random effect model (i.e. heteroscedasticity) need to be addressed while analyzing panel data. Because of many panel data models estimators becomes grossly inconsistent and inefficient [2], [7], [9] and [14] .

One of the critical assumptions of the classical linear regression model (CLRM) is that the error terms in the model are independent of all regressors. If this assumption is violated, then endogeneity is suspected $cov\,(\varepsilon_{it}\,, x_{it}\,) \neq 0\,, for\ ever\ i\ and\ t$ and hence within estimator is no longer consistent [1], [2]. Also, the error terms are expected to have the same variance. If this is not satisfied, there is heteroscedasticity (i.e. $ar\,(v_{it}) = var\,(\varepsilon_{it} + \alpha_i) = \sigma^2\Sigma$) see [6], [8] , [10], [13] and [17].

In the presence of heteroscedasticity, the usual OLS estimators are no longer having minimum variance among all linear unbiased estimators [3] and [8]. Thus, the OLS estimator is not efficient relative to GLS under such situations. The studies of [3], [4], [5], [12] and [15] focused on the existence of heteroscedasticity in panel data modelling.

A number of works on the methodologies and applications of panel data model estimation have appeared in the literature see [3], [5],[7],[9],[11],[12] and [14]. Situations where all the necessary assumptions underlying the use of classical linear regression methods are satisfied are rarely found in real life situations. Most of the studies that discussed panel data modelling considered the violation of each of the classical assumptions separately and the detailed derivation of the estimators has minimum attention in much literature.

The aim of this present paper is to elucidate the part of the earlier papers pertaining to panel data model estimators. The study contributes to the literature in several ways. First, we set out the assumptions behind the fixed and random effect approaches, highlight their strengths and weaknesses. Also, brief estimation method, procedures of estimation and detailed derivation the estimators are given. Results from this work would serve as useful guides to econometricians and students while estimating panel data that are characterized by the structure conjectured here.

2. Estimation Framework and Model Specification

Panel or longitudinal data provide more information than cross-sectional data, which increases estimation precision and also enables researchers to control for unobserved heterogeneity related to the omitted variable bias in cross-section models. Panel data can not only offer us the information across different individuals, but also the information for a given individual across the time.

2.1 Panel Data models

The basic panel data model takes the form;

$$y_{it} = X_{it}'\beta + \alpha_i + \varepsilon_{it}\ i = 1,2,\dots N\ ;\ t = 1,2,\dots .T \quad (1)$$

Where i is the individual dimension and t is the time dimension. Therefore, y_{it} is the response of individual i at time t, α_i are the unobserved individual-specific, time-invariant intercepts, X_{it} is the explanatory variable i at time t, β is a vector of regression coefficients, and ε_{it} is the error term of individual i at time t. They are also known as idiosyncratic errors because they change across i as well as across t (Hsiao, 2002). ε_{it} is iid over i and t. It has usual properties, i.e. mean 0, uncorrelated with itself , uncorrelated with α_i , and homoscedastic.

$$E(\varepsilon_{it}|\ \alpha_i, x_{i1}\ , \dots x_{iT}) = 0 \quad (2)$$

The various panel data models depend on the assumptions made about the in individual specific effects α_i. Mundlak (1978) and Chamberlain (1982) view individual effect α_i as random draws along with the observed variables. Then, one of the key issues is whether α_i is correlated with elements of X_{it} . The equation (3.4) is useful to emphasizing which factors change only across, which change only across t , and which change across i and t.

Wooldridge (2003) avoids referring to α_i as a random effect or a fixed effect. Instead, we will refer to α_i as unobserved effect, unobserved heterogeneity, and so on. Nevertheless, later we will label two different estimation methods random effects estimation and fixed effects estimation.

Fact that for Wooldridge (2003), these discussions about whether the α_i should be treated as random variables or as parameters to be estimated are wrongheaded for micro econometric panel data applications. With a large number of random draws from the cross section, it almost always makes sense to treat the unobserved effects, α_i , as random draws from the population, along with y_{it} and X_{it} . This approach is certainly appropriate from an omitted variables or neglected heterogeneity perspective. As suggested by Mundlak (1978), the key issue involving α_i is whether it is uncorrelated with the observed explanatory variables X_{it}, for $t = 1,..,T$.

In the traditional approach to panel data models, α_i is called a random effect, when it is treated as a random variable and a fixed effect, when it is treated as a parameter to be estimated for each cross section observation.

2.1.1 Fixed Effects Model

One variant of model (1) is called fixed effects (FE) model which treats the unobserved individual effects as random variables that are potentially correlated with the explanatory variables, $E\left(X_{it}\,\alpha_i\right) \neq 0$, (Wooldridge, 2002). Unlike the random effects estimators, the FE estimator assumes nothing regarding the correlation structure between α_i and the explanatory variables. As we don't know the statistical properties of α_i, it can be eliminated from the model. Among various ways to eliminate α_i, the within-group transformation or deviation from mean is easy to understand. The procedure of within transformation as follows;

Step 1: Average equation (3.4) over $t = 1,2,\dots\dots T$ to get the cross section equation:

$$\bar{y}_i = \bar{X}_i\beta + \alpha_i + \bar{\varepsilon}_i\ , i = 1,\dots\dots N \quad (3)$$

where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$; $\bar{X}_i = T^{-1}\sum_{t=1}^{T} X_{it}$; $\bar{\varepsilon}_i = T^{-1}\sum_{t=1}^{T}\varepsilon_{it}$ and $\bar{\alpha}_i = \alpha_i$ .These are called time means for each unit i. The OLS estimator for β obtained from (3) is called between estimator.

Step 2: To eliminate α_i subtract equation (3) from (1) for each t gives the fixed effects transformed equation,

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i\ )'\beta + \varepsilon_{it} - \bar{\varepsilon}_i$$

or equivalently

$$\ddot{y}_{it} = \ddot{X}_{it}'\beta + \ddot{\varepsilon}_{it}\ , i = 1,\dots N\ , t = 1,\dots T \quad (4)$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$; $\ddot{X}_{it} = X_{it} - \bar{X}_i$; $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ and $\alpha_i - \bar{\alpha}_i = 0$ and hence the effect is eliminated. Also, we define $\alpha = E(\alpha_i)$, so $E(\alpha_i - \alpha) = 0$, Since α_i is fixed or constant for every cross sectional unit. Like first differencing, time demeaning of the original equation has removed the individual effect α_i . With α_i out of the model, it is natural to estimate equation (4) by OLS if X_{it} is strictly exogenous .The OLS estimator obtained from (4) is often called the within estimator. Consistent estimation of this estimator requires X_{it} being strictly exogenous i.e. $E(\varepsilon_{it}|x_{i1}\ , \dots x_{iT}, \alpha_i) = 0$. However, when this assumption is violated, within estimator is no longer consistent. We suspect the correlation between α_i and X_{it} will leads to endogeneity problem. Hence, two-stage least square is treatment for this problem.

2.1.1.1 Two-Stage Least Square estimation

In regression model, we assume that variable y_{it} is determined by X_{it} but does not jointly determine y_{it}. However, many economic models involve endogeneity that in which response variable is determined by joint of X_{it}. When X_{it} is endogenous or jointly determined with y_{it} , then the estimation of the model will result inconsistent estimators and enlarge variance of estimators. This endogeneity problem is the consequence of omitted variable.

The treatment for this problem is to introduce instrumental variables Z_{it} which cut relationship between X_{it} and ε_{it} which depends on the following assumptions. (1) Z_{it} is uncorrelated with the error ε_{it}. (2) Z_{it} is correlated with the regressor X_{it}. To allow correlation between X_{it} and ε_{it} , we assume there exists a $1 \times L$ vector of instruments ($L \geq K$), Z_{it} which avoid correlation.

Now assume model with one endogenous explanatory variable X_K , $Y_{it} = X_{it}\beta + \varepsilon_{it}$

with assumption that $E(\varepsilon_{it}) = 0$, $Cov\,(x_k\,, \varepsilon_{it}) = 0\,, k = 1, 2\,\ldots, k-1$ and $Cov\,(x_k\,, \varepsilon_{it}) \neq 0$,for K, where $x_1\,, x_2\,, \ldots, x_{K-1}$ are exogenous and X_K is endogenous.

To fix the problem, consider z_1 as replacer of an endogenous explanatory X_K satisfies that $Cov\,(z_1\,, \varepsilon_{it}) = 0$ and $\theta_1 = \frac{\partial\, L\,(X_K|1, x_1\,, x_2\,, \ldots, x_{K-1}\,, z_1\,)}{\partial\, z_1} \neq 0$.Thus, we have $Z = (1,\, x_1\,, x_2\,, \ldots, x_{K-1}\,, z_1\,)$.Then, endogenous explanatory variable X_K can be written as

$$x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K\,, \theta_1 \neq 0 \qquad (5)$$

where, by definition $E(r_K) = 0$ and $Cov\,(r_K\,;\; x_1\,, x_2\,, \ldots, x_{K-1}\,, z_1\,) = 0$

By substituting estimated x_K in the regression model we can estimate the model by usual OLS.

For each i and t , define $\ddot{Z}_{it} = Z_{it} - \bar{Z}_i$, $\bar{Z}_i = T^{-1}\sum_{t=1}^{T} Z_{it}$ and similarly for $\ddot{y}_{it}\,, \ddot{X}_{it}\,, \ddot{\varepsilon}_{it}$.

Define also $\ddot{y} = (\ddot{y}_{i1}\,, \ddot{y}_{i2}\,, \ldots, \ddot{y}_{iT})\,, \ddot{X} = (\ddot{X}_{i1}\,, \ddot{X}_{i2}\,, \ldots, \ddot{X}_{iT})\,, \ddot{Z} = (\ddot{Z}_{i1}\,, \ddot{Z}_{i2}\,, \ldots, \ddot{Z}_{iT})$, and $\ddot{\varepsilon} = (\ddot{\varepsilon}_{i1}\,, \ddot{\varepsilon}_{i2}\,, \ldots, \ddot{\varepsilon}_{iT})$ .Then, the transformed model becomes $\ddot{y} = \ddot{X}\beta + \ddot{\varepsilon}$.

Suppose that $\ddot{Z}$ has the same number of variables as $\ddot{X}$, i.e. $L = K$. We assume that the rank of $\ddot{Z}'\ddot{X}$ is K, so now $\ddot{Z}'\ddot{X}$ is square matrix.

By premultiplying transformed model by $\ddot{Z}'$ and taking expectation we obtain instrumental variable estimator:

$$\hat{\beta}_{IV} = \left(\ddot{Z}'\ddot{X}\right)^{-1}\ddot{Z}'\ddot{y}$$
$$= \left(\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\ddot{Z}_{it}\ddot{X}_{it}'\right)\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\ddot{Z}_{\mathrm{it}}'\ddot{y}_{it}$$

However, the best way to get consistent estimate is to use all available instruments. If we have a single endogenous explanatory variable, but have more than one potential instrument and each of which would have a significant coefficient in (1).

Let $z_1\,, z_1\,, \ldots, z_M$ be instrumental variables such that $ov\,(Z_h\,, \varepsilon_{it}) = 0\,, h = 1, 2, \ldots, M$, so that each Z_h is exogenous in (1) and assume $E(\varepsilon_{it}) = 0$,
$Cov\,(x_k\,, \varepsilon_{it}) = 0\,, k = 1, \ldots, k-1$, $Cov\,(x_k\,, \varepsilon_{it}) \neq 0\,, for\ K$ and
$Cov\,(Z_h\,, \varepsilon_{it}) = 0\,, h = 1, 2, \ldots, M.$

Now, assume that Z_h has more number of variables than x_K , i.e. $L > K$. Define the vector of exogenous variables again by $= (1, x_1\,, x_2\,\ldots, x_{K-1}\,, z_1\,, z_1\,, \ldots, z_M\,)$, a $1 \times L$ vector($L = K + M$). The method 2SLS considers $z_1, z_2, \ldots, z_M$ as of replacer of an endogenous explanatory X_K satisfies that $Cov\,(\varepsilon_{it}\,;\; z_1, z_2, \ldots, z_M) = 0$ and

$$\theta_1 = \frac{\partial\, L\,(X_K|1,\, x_1\,, \ldots, x_{K-1}\,, z_1, z_2, \ldots, z_M\,)}{\partial\, z_1} \neq 0$$

The linear projection of x_K on Z can be written as

$$x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M + r_K \qquad (6)$$

where, $E(r_K) = 0$ and $Cov\,(r_K\,;\; x_1\,, x_2\,, \ldots, x_{K-1}\,, z_1\,, z_1\,, \ldots, z_M\,) = 0$

Fit (3.13) by OLS

$$\hat{x}_K = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \ldots + \hat{\delta}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \ldots + \hat{\theta}_M z_M$$

We denote $\hat{X} = (\,x_1\,, x_2\,, \ldots, x_{K-1}\,, \hat{x}_K\,)$. Two-stage estimation under instrumental variables to an endogenous explanatory x_K , referencing as

$$Y = X\beta + \varepsilon \qquad (7)$$

where $x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M + r_K$

Multiplying equation (7) by $\hat{X}'$

$$\hat{X}'Y = \left(\hat{X}'X\right)\beta + \hat{X}'\varepsilon$$

Again multiplying by $\left(\hat{X}'X\right)^{-1}$

$$\left(\hat{X}'X\right)^{-1}\left(\hat{X}'Y\right) = \left(\hat{X}'X\right)^{-1}\left(\hat{X}'X\right)\beta + \left(\hat{X}'X\right)^{-1}\hat{X}'\varepsilon$$

Taking expectation

$$E\left[\left(\hat{X}'X\right)^{-1}\left(\hat{X}'Y\right)\right] = E\left[\left(\hat{X}'X\right)^{-1}\left(\hat{X}'X\right)\right]\beta$$
$$+E\left[\left(\hat{X}'X\right)^{-1}\hat{X}'\varepsilon\right]$$

Estimation of β as in population

$$\beta = E\left[\left(\hat{X}'X\right)^{-1}\left(\hat{X}'Y\right)\right]$$

Estimation of β as in sample

$$\hat{\beta} = \left(\hat{X}'X\right)^{-1}\left(\hat{X}'Y\right)$$

Two -stage regression for estimation of β are given below.

First-stage regression - obtain fitted values of $\hat{x}_K$ from the regression x_K where

$x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M + r_K$

Then, we denote

$$\hat{x}_K = (x_K\,|1, x_1\,, x_2\,\ldots, x_{K-1}\,, z_1\,, z_1\,, \ldots, z_M\,)$$

Second -stage regression- Run the OLS regression y on $(1, x_1\,, x_2\,\ldots, x_{K-1}\,, \hat{x}_K)$

⋮⋮⋮

$$\theta_M = \frac{\partial\, L\,(X_K|1,\, x_1\,, \ldots, x_{K-1}\,, z_1, z_2, \ldots, z_M\,)}{\partial\, z_M} \neq 0$$

$$y = \beta_0 + \beta_1 X_1 + \ \ldots\ + \beta_{k-1} X_{k-1} + \ \beta_K \hat{x}_K \ + \varepsilon \quad (8)$$

It is x_K with $Cov\,(\varepsilon_{it}\,, x_K\,) \neq 0$ that leads the estimators of β to be inconsistent. But , we can use Z_h , $h = 1, 2, \ldots, M$ as a candidates for x_K .let $X_K = Z\pi_K + r_K$, and

$$X = Z\Pi + r_K\ , \Pi = (\pi_1\,, \pi_2, \ldots, \pi_K) \qquad (9)$$

Multiplying (9) by $Z^{'}$ and taking expectation

$$E\,(Z^{'}X) = E\,(Z^{'}Z)\Pi\ + E\,(Z^{'}r_K\,)$$

Then, we have $\Pi = \left(E\,(Z^{'}Z)\right)^{-1}\, E\,(Z^{'}X)$

Next , $X^{*} = E\,(x|z) = Z\Pi$,

Multiplying (9) by X^{*} and taking expectation gives

$$E\,(X^{*}y) = \ E\,(X^{*}X)\beta + E\,(\,X^{*}\varepsilon)$$

Solving for β gives

$$\beta = [E\,(X^{*}X)]^{-1}\, E\,(X^{*}y). \qquad (10)$$

But, $E\,(X^{*}X) = \ E\,(X^{'}Z)\left(E\,(Z^{'}Z)\right)^{-1}\, E\,(Z^{'}\,X)$ and $E\,(X^{*}y) = E\,(X^{'}Z)\left(E\,(Z^{'}Z)\right)^{-1}\, E\,(Z^{'}\,y).$

Therefore, substituting this results in (10) yields

$$\hat{\beta}_{2SLS} = [\,X^{'}Z\,(Z^{'}Z)^{-1}\,Z^{'}\,X]^{-1}\,X^{'}Z\,(Z^{'}Z)^{-1}\,Z^{'}\,y$$

$$= \left[\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^{'}Z_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^{'}Z_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^{'}X_{it}\right)\right]^{-1}$$

$$\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^{'}Z_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^{'}Z_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^{'}y_{it}\right)$$

Expressing $\hat{\beta}_{2SLS}$ in terms of transformed model in (4) :

$$\hat{\beta}_{2SLS} \ = \left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{y}$$

$$= \left[\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{X}_{it}^{'}\ddot{Z}_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{Z}_{it}^{'}\ddot{Z}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{Z}_{it}^{'}\ddot{X}_{it}\right)\right]^{-1}$$

$$\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{X}_{it}^{'}\ddot{Z}_{it}\right)\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{Z}_{it}\right)^{-1}\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{y}_{it}\right)$$

This is called fixed effect two-stage least square (2SLS) estimator.

2.1.1.2 Asymptotic variance of 2SLS estimator in fixed -effect model

Recall the definition of the 2SLS -estimator of transformed model

$$\hat{\beta}_{2SLS} \ = \left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{y}$$

$$= \left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{\varepsilon}$$

$$\hat{\beta}_{2SLS} - \beta \ = \left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{\varepsilon}$$

Therefore, the variance of 2SLS -estimator is defined by

$$\text{Avar}\left(\hat{\beta}_{2SLS}\right) \ = \ E\left\{\left(\hat{\beta}_{2SLS} - \beta\right)\left(\hat{\beta}_{2SLS} - \beta\right)^{'}\right\}$$

$$= E\left\{\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{\varepsilon}\,\ddot{\varepsilon}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\right\}$$

$$= \left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}E(\ddot{\varepsilon}\,\ddot{\varepsilon}^{'})\,\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}$$

Under Homoskedasticity (constant variance of error term), $E(\ddot{\varepsilon}\,\ddot{\varepsilon}^{'}) = \sigma^2$,then

$$\text{Avar}\left(\hat{\beta}_{2SLS}\right) =$$

$$\sigma^2\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}$$

$$\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}$$

$$= \sigma^2\left[\ddot{X}^{'}P_z\,\ddot{X}\right]^{-1}\ddot{X}^{'}P_z P_z\,\ddot{X}\left[\ddot{X}^{'}P_z\,\ddot{X}\right]^{-1}$$

where $P_z = \ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}$ is projection matrix

$$= \sigma^2\left[\ddot{X}^{'}P_z\,\ddot{X}\right]^{-1}, since\ P_z = P_z^{'}\ and\ P_z^2 = P_z$$

$$= \sigma^2\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}$$

When $E(\varepsilon\,\varepsilon^{'}) = \sigma^2$, then covariance matrix has the same form as OLS, but in terms of predicted values:

$$\text{Avar}\left(\hat{\beta}_{2SLS}\right) = \hat{\sigma}^{\,2}\left[\hat{\ddot{X}}^{'}\hat{\ddot{X}}\ \right]^{-1}$$

Recall $\hat{X} = Z\left(Z^{'}Z\right)^{-1}Z^{'}X$ implies $\hat{\ddot{X}} = \ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}$ (OLS formula applied to the first stage) , thus

$$\hat{\ddot{X}}^{'}\hat{\ddot{X}} = \ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}$$

Hence,

$$\text{Avar}\left(\hat{\beta}_{2SLS}\right) = \sigma^2\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}$$

$$= \sigma^2\left[\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{X}_{it}^{'}\ddot{Z}_{it}\right)\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{Z}_{it}\right)^{-1}\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{X}_{it}\right)\right]^{-1}$$

where σ^2 can be consistently estimated by

$$\hat{\sigma}^{\,2} = (NT - K)^{-1}\,\hat{\ddot{\varepsilon}}^{'}\hat{\ddot{\varepsilon}} = (NT - K)^{-1}(\ddot{y} - \ddot{\mathrm{X}}\hat{\beta}_{2SLS})^{'}(\ddot{y} - \ddot{\mathrm{X}}\hat{\beta}_{2SLS})$$

The $\hat{\ddot{\varepsilon}} = \ddot{Y} - \ddot{X}\hat{\beta}_{2SLS}$ which is the $NT \times 1$ column vector of estimated residuals. Notice that these residuals are not the residuals from the second stage OLS regression of dependent $\ddot{Y}$ on the predicted variables $\ddot{X}$.

Therefore, the estimated asymptotic variance of 2SLS estimator is

$$\hat{\text{Avar}}\left(\hat{\beta}_{2SLS}\right) = \hat{\sigma}^{\,2}\left[\ddot{X}^{'}\ddot{Z}\left(\ddot{Z}^{'}\ddot{Z}\right)^{-1}\ddot{Z}^{'}\ddot{X}\right]^{-1}$$

$$= \hat{\sigma}^{\,2}\left[\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{X}_{it}^{'}\ddot{Z}_{it}\right)\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{Z}_{it}\right)^{-1}\left(\tfrac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\ddot{Z}_{it}^{'}\ddot{X}_{it}\right)\right]^{-1}$$

However, a major limitation of the fixed effects estimator is that the coefficients of time-invariant explanatory variables are not identified. Thus it is not suited to estimate the effects of time constant variables, such as ethnic group, education before landing and immigration class on earnings.

2.1.2 Random effects model

It is commonly assumed in regression analysis that all factors that affect the dependent variable, but that have not been included as regressors, can be appropriately summarized by a random error term. In our case, this leads to the assumption that the α_i are random factors, independently and identically distributed over individuals and hence treated as error term.

This model is another variant of the model (1) which assumes that the unobserved individual effects α_i are random variables that are distributed independently of the explanatory variables i.e.

$$E\left(\alpha_i \middle| X_{it}\right) = 0 \qquad (11)$$

This model is called random effects model, which usually makes the additional assumptions that $\alpha_i \sim NIID(\alpha, \sigma_\alpha^2)$ and

$$\varepsilon_{it} \sim NIID(0,\ \sigma_\varepsilon^2) \qquad (12)$$

Thus, we write the random effects model as

$$y_{it} = X_{it}^{'}\beta + v_{it}\ i = 1,2,\dots N\ ;\ t = 1,2,\dots .T \qquad (13)$$

where $v_{it} = \alpha_i + \varepsilon_{it}$ which treated as an error term consisting of two components :

An individual specific component (α_i) , which does not vary over time, and a remainder component (ε_{it}), which is assumed to be uncorrelated over time. That is, all correlation of the error terms over time is attributed to the individual effects α_i.

The α_i are assumed independent of ε_{it} and X_{it} which are also independent of each other for all i and t. This assumption is not necessary in the fixed effect model. The components of $\mathrm{Cov}(v_{it}, v_{js}) = E(v_{it}, v_{js})$ are $\sigma_v^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ if $i = j$ and $t = s$, σ_α^2 if $i = j$ and $t \neq s$ and 0 if s, t and $i \neq j$. Thus, the Ω matrix or variance structure of errors looks like

$$\mathrm{Var}\,(v_{it}) = \sigma_\varepsilon^2 I_T + \sigma_\alpha^2\, i_T\, i_T^{'}$$

$$= \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \dots. & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix} = \Omega \qquad (14)$$

where i_T is a $T \times 1$ column vector of ones. when Ω has the above form, we say it has random effects structure.

A random effect model is estimated by generalized least squares (GLS) when the variance structure is known. Compared to fixed effect models, random effect models are relatively difficult to estimate. This document assumes panel data are balanced.

2.1.2.1 Generalized Least Squares (GLS)

It is well known that the omission of an explanatory variable(s) or uses of an incorrect functional form in a regression that otherwise satisfies the full ideal conditions, can lead to the erroneous conclusion that autocorrelation or heteroscedasticity is present among the disturbances. Thus, variance of error term is not constant. Heteroscedasticity is the case where $E(v_{it}\, v_{it}^{'}) = \Omega = \sigma^2\, \Sigma$ is a diagonal matrix, so that the errors are uncorrelated, but have different variances.

The common practice, however, is to use generalized least squares (GLS) and it achieves efficiency by transforming a heteroscedasticity variance covariance matrix into a homoscedastic one. When Ω is known (given), GLS based on the true variance components is BLUE and all the asymptotically efficient as either n or T approaches infinity (Baltagi 2001). When α_i is a random variables then OLS estimator is generally inefficient relative to GLS estimator. Because every y_{it} for $t = 1,2,\dots,T$ contains the same α_i , there will be covariance among the observation for each individual that GLS will exploit. The GLS estimator corresponding to this component structure has special structure. This need all of its reweighting within the time series y_i of an individual.

Therefore, to derive GLS we need to focus only on T-dimensional relationship,

$$y_i = X_i\beta + i_T\alpha_i + \varepsilon_i \qquad (15)$$

setting $v_i = i_T\alpha_i + \varepsilon_i$, model becomes $y_i = X_i\beta + v_i$.

Furthermore, the conditional variance of y_i given X_i depends on an orthogonal projector , α_i.

Define $i_T^{'} i_T = T$, we can write variance of random effect structure , Ω as

$$\Omega = \sigma_\alpha^2\, i_T\, i_T^{'} + \sigma_\varepsilon^2 I_T = T\sigma_\alpha^2\, i_T\, (i_T^{'} i_T)^{-1} i_T^{'} + \sigma_\varepsilon^2 I_T$$

Let $P_T = i_T\, (i_T^{'} i_T)^{-1} i_T^{'} = I_T - Q_T$ then,

$$\Omega = T\sigma_\alpha^2 P_{i_T} + \sigma_\varepsilon^2 I_T = (\,T\sigma_\alpha^2 + \sigma_\varepsilon^2\,)\, P_T + \sigma_\varepsilon^2\, (I_T - P_T)$$

$$= (\,T\sigma_\alpha^2 + \sigma_\varepsilon^2\,)\, P_T + \sigma_\varepsilon^2\, Q_T$$

$$= (\,T\sigma_\alpha^2 + \sigma_\varepsilon^2\,)\, (\,P_T + \theta\, Q_T), \text{ where } \theta = \frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

For application of GLS estimator, one needs to know the inverse of Ω^{-1} which van be written as

$$\Omega^{-1} = \sigma_\varepsilon^{-2}\left[I_T - \frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2} i_T i_T^{'}\right]$$

$$= \sigma_\varepsilon^{-2}\left[I_T - \frac{T\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}\frac{1}{T} i_T i_T^{'}\right]$$

which can also be written as

$$\Omega^{-1} = \sigma_\varepsilon^{-2}\left[\left(I_T - \frac{1}{T} i_T i_T^{'}\right) + \theta\, \frac{1}{T} i_T i_T^{'}\right]$$

$$= \sigma_\varepsilon^{-2}\left[P_T + \theta\, \frac{1}{T} i_T i_T^{'}\right]$$

Note that $P_T = I_T - Q_T = I_T - \frac{1}{T} i_T i_T'$ used to transform the data in deviation from individual means and $\frac{1}{T} i_T i_T'$ takes individual means.

Suppose that instead of $V(v_i) = \sigma^2 I_{NT}$, we may have $ar(v_i) = \Omega = \sigma^2 \Sigma$, where the matrix Σ contains terms for heterogeneity which is known, symmetric and positive definite but σ^2 is unknown.

Assume Ω has the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_T$, by Cholesky's decomposition, we can write as

$$\Omega = S \Lambda S'$$

where Λ is a diagonal matrix with the diagonal elements $(\lambda_1, \lambda_2, \dots, \lambda_T)$ and S is an orthogonal matrix. Columns of S are the characteristic vectors of Ω and the characteristic roots of Ω are arrayed in the diagonal matrix Λ .Thus,

$$\Omega^{-1} = S^{-1} \Lambda^{-1} S'^{-1} = S^{-1} \Lambda^{-1/2} \Lambda^{-1/2} S'^{-1} = PP'$$

where $P = S^{-1} \Lambda^{-1/2}$ and $\Lambda^{-1/2}$ is a diagonal matrix with the diagonal elements $\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}\right), \dots, \sqrt{\lambda_T}$. Then its straight forward to prove that $P' \Omega = I_T$, so $P'(P \Omega P') = P'$

Our interest is to make error terms to be iid which leads to have constant variance. Py, PX and Pv has typical element $(y_{it} - \lambda \bar{y}_i)$, $(X_{it} - \lambda \bar{X}_i)$ an $(\varepsilon_{it} - \lambda \bar{\varepsilon}_i)$ respectively , where $\lambda = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}}$. The term λ gives a measure of the relative sizes of the within and between unit variances.

The random effect estimator (GLS) uses both within-group (deviation from individual mean) and between-group (individual mean) variations, but weights them according to the relative sizes of $T\sigma_\alpha^2 + \sigma_\varepsilon^2$ and σ_ε^2. It is equivalent to the following two steps:

1) Transform the data: $y_{it}^* = y_{it} - \lambda \bar{y}_i, X_{it}^* = X_{it} - \lambda \bar{X}_i$ *and* $v_{it}^* = (1 - \lambda)\alpha_i + v_{it} - \lambda \bar{v}_i$.

2) Regress y_{it}^* on X_{it}^*. In GLS, we just need to compute λ using the matrix Ω. Then variance parameters σ_ε^2 and σ_α^2 can be estimated from the within-group and between-group regression residuals. Note that $\lambda = 0$ corresponds to pooled OLS, $\lambda = 1$ and $\sigma_\varepsilon^2 = 0$ corresponds to within estimation, and $\lambda \to 1$ as $T \to \infty$, this is a two-step estimator of β. λ=0 implies there is no covariance among observations. Then, $\hat{\beta}_{GLS} \xrightarrow{P} \hat{\beta}_{Pooled}$ where Parameter θ can take any value between one and zero, i.e. $0 \le \theta \le 1$.
Finally, to obtain GLS estimator run OLS on the transformed model:

$$y_{it}^* = X_{it}^* \beta + v_{it}^* \qquad (16)$$

Where $v_{it}^* = (1 - \lambda)\alpha_i + v_{it} - \lambda \bar{v}_i$ which is asymptotically iid. This transformed model satisfies the classical assumption. Because Ω is assumed to be known, y_{it}^* and X_{it}^* are observed data.

Therefore, the random effect estimator is given by $\hat{\beta}_{GLS} = \left(X^{*'} X^*\right)^{-1} X^{*'} y^*$
$= (X'PP'X)^{-1} X'PP'y$
$= (X'\Omega X)^{-1} X'\Omega y$
$= \left(\sum_{i=1}^N X_i' \Omega^{-1} X_i\right)^{-1} \sum_{i=1}^N X_i' \Omega^{-1} y_i$
$= \left(\sum_{i=1}^N \sum_{t=1}^T X_{it}' \Omega^{-1} X_{it}\right)^{-1} \sum_{i=1}^N \sum_{t=1}^T X_{it}' \Omega^{-1} y_{it}$
and
The variance of GLS estimator which is conditional on X_{it} can be calculated using
$\hat{\beta}_{GLS} = (X^{*'} X^*)^{-1} X^{*'} y^*$
$= (X^{*'} X^*)^{-1} X^{*'} (X_i^* \beta + v_i^*)$
$= \beta + (X^{*'} X^*)^{-1} X^{*'} v_i^*$
$\hat{\beta}_{GLS} - \beta = (X^{*'} X^*)^{-1} X^{*'} v_i^*$

Therefore,
$Var\left(\hat{\beta}_{GLS}\right) = E\left\{(\hat{\beta}_{GLS} - \beta)(\hat{\beta}_{GLS} - \beta)'\right\}$
$= E\{(X^{*'} X^*)^{-1} X^{*'} v^* \, v^{*'} X^* (X^{*'} X^*)^{-1}\}$
$= (X^{*'} X^*)^{-1} X^{*'} E(v^* \, v^{*'}) X^* (X^{*'} X^*)^{-1}$
$= (X^{*'} X^*)^{-1} X^{*'} X^* (X^{*'} X^*)^{-1}$
$= (X^{*'} X^*)^{-1}$
$= (X'PP'X)^{-1}$
$= \left(X'\Omega^{-1} X\right)^{-1}$
$= \left(\sum_{i=1}^N X_i' \Omega^{-1} X_i\right)^{-1}$
$= \left(\sum_{i=1}^N \sum_{t=1}^T X_{it}' \Omega^{-1} X_{it}\right)^{-1}$

Covariance matrix Ω is assumed to be known, since y_{it}^* and X_{it}^* are observed data. The gain to this approach is that it substantially reduces the number of parameters to be estimated. However, assumption (11) is unlikely to hold in many cases. In the present study, the unobserved individual invariant effects α_i could include personal characteristics such as ability, motivation and preferences which are very likely related to some explanatory variables for wages, like educational attainment, social network type and content and so on. In this case $E\left(\alpha_i \middle| X_{it}\right) \neq 0$ and the random effects estimator is biased and inconsistent.

3. Conclusion and Recommendation

In this paper, we have discussed brief estimation method and procedures for estimating panel data regression models. The assumptions behind the fixed and random effect approaches and their strengths and weaknesses are also presented. We have shown how to estimate fixed effect panel data models when the equation contains endogenous explanatory variables, where endogeneity is conditional on the unobserved effect and the estimation of random-effects is based on the assumption that the correlation between the regressors and the unobservable, individual-specific effects is zero. Two estimators are considered in estimating panel data models with endogeneity and heteroscedasticity. Detailed derivations of linear panel data models estimators are discussed. In particular, we derive two-stage least square(2SLS) estimator to estimate fixed effects and generalized least square (GLS) to estimate random effects. One of the most important uses of deriving these estimators is to increase understanding of estimators and reduce computational difficulty while estimating panel data models.

It is hereby recommended that for any econometric problems involving both cross-sectional and time series data, it is appropriate and adequate to use panel data model in analyzing such data. There are other methods of analyzing panel data in econometrics depending on the econometric problem to be addressed; such methods include the random intercept model, Pooled model, unrelated regression model, dynamic model, unbalance panel data model etc. It is recommended that further study / research work should focus on the use of these methods.

References

[1] Anatolyev (2011), Instrumental variables estimation and inference in the presence of many exogenous regressors. Working paper No.47, New Economic School, Nakhimovsky, Russia.
[2] Angris and Imben (1995), Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. Journal of the American Statistical Association June 1995, Vol. 90, No. 430, Applications and Case Studies.
[3] Baltagi, B. H. (2005), Econometrics analysis of panel data, 3rd edition, John Wiley and Sons Ltd, England.
[4] Baltagi and Griffin (1988), A generalized error component model with heteroscedastic disturbances, International Economic Review 29, 745-753.
[5] Bresson, et al (2006), Heteroskedasticity and random coefficient model on panel data, Working Papers ERMES 0601, ERMES, University Paris 2.
[6] Creel (2014), Econometrics. University Autonoma , Barcelona.
[7] Garba, et al (2013), Investigations of Certain Estimators for Modeling Panel Data under Violations of Some Basic Assumptions. Journal of Mathematical Theory and Modeling ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) Vol.3, No.10, 2013.
[8] Green. H (2012), Econometric Analysis, 6th edition, New York University
[9] Kruiniger (2001), On the estimation of panel regression model with fixed effects working paper, Department of Economics, Queen Mary, University of London, England.
[10] Maddala, G.S. (2008), Introduction to econometrics, 3rd edition, John Wiley & Sons, Ltd, Chichester, UK.
[11] Matyas ,et al (2012),The Formulation and Estimation of Random Effects Panel Data Models of Trade. Working paper No. 12/2, Central European University, Department of Economics, Hungary.
[12] Olofin, et al (2010), Testing for heteroscedasticity and serial correlation in a two way error component model. Ph.D dissertation submitted to the Department of Economics, University of Ibadan, Nigeria.
[13] Schmidt (2005), Econometrics, McGraw-Hill/Irwin, New York.
[14] Semykina and Wooldridge (2008), Estimating Panel Data Models in the Presence of Endogeneity and Selection. Working paper, Florida State University, USA.
[15] Wansbeek, T.J. (1989), An alternative heteroscedastic error component model, Econometric Theory 5, 326.
[16] Wooldridge, J. M. (2002), Econometric Analysis of Cross Section and Panel Data.Cambridge University , London, England.
[17] Wooldridge, J. M. (2012), Introductory Econometrics: A Modern Approach, 5th edition, South-Western College.

Author Profile

Megersa Tadesse Jirata received a BSc in Statistics from the Wollega University of Ethiopia in 2010. Currently he is studying towards a MSc Statistics degree at Pan African University Institute of Basic Sciences, Technology and Innovation (PAUISTI) hosted at Jomo Kenyatta University of Agriculture and Technology (JKUAT) in Kenya.

D r. Joel Cheruyot Chelule is the senior lecturer of the Department of Statistics and Actuarial Science at Jomo Kenyatta University of Agriculture and Technology. He has over 5 publications in statistics and he is a current Deputy Director of Academic Quality Assurance (Academic Affairs) at Jomo Kenyatta University of Agriculture and Technology.

Professor Romanus Otieno Odhiambo is a professor in statistics. He received Bachelor of Education, MSc (Statistics), PhD (Statistics) from Kenyatta University. He has published over 25 papers and he is the current Deputy Vice Chancellor (Academic Affairs) at Jomo Kenyatta University of Agriculture and Technology.