COMPARISON OF SOME PANEL DATA REGRESSION MODEL ESTIMATORS USING SIMULATED DATA

J. T. Megesa[1], J. C. Chelule[2] and R. O. Odhiambo[3]
[1]Pan African University Institute of Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2,3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: magetade2003@gmail.com

Abstract
This paper presents estimation of panel data regression models with individual effects. We discuss estimation techniques for both fixed and random effects panel data regression models. We derive two-stage least squares and generalized least squares estimators, and discuss their limitations. Under specified conditions, we investigate the asymptotic properties of the derived estimators, in particular, the consistency and asymptotic normality, and the Hausman test for panel data regression models with large number of cross-section and fixed time-series observations. We show that both estimators are consistent and asymptotically normally distributed and have different convergence rates dependent on the assumptions of the regressors and the remainder disturbances. We also perform simulation studies to see the performance of our estimates for large cross sections. Our simulation results show that the estimator based on the bigger sample is more consistent than the one based on the smaller sample size. We find that the two-stage least squares estimator performs better in the presence of endogeneity, while the generalized least squares estimator performs better under strict exogeneity conditions. We also note that the generalized least squares estimator performs better than the ordinary least squares estimator in the absence of correlation between individual effects and the regressors.

Key words: panel data, fixed effects, random effects, two-stage least square, generalized least square, consistency, asymptotic normality, endogeneity and heterogeneity

1.0 Introduction

Panel data are repeated observations on the same cross section, typically of individuals or firms observed for several time periods. This could be generated by pooling time-series observations across a variety of cross-sectional units including countries, states, regions, firms, or randomly sampled individuals or households. The panel data models are being widely used, amongst other reasons, due to the computational advance. An important advantage of using these data is that they allow researchers to control for unobservable individual time-invariant heterogeneity, that is, systematic differences across cross-sectional units (e.g., individuals, households, firms, countries). It also presents advantages in relation to cross-section and time series models for, in addition to increasing the degrees of freedom, they manage to remove the influence of the individual of the independent variables, thus making the estimates of model coefficients more realistic. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates. Panel data sets are also better able to identify and estimate effects that are simply not detectable in pure cross-sections or pure time-series data. If the intercept or individual effects are incorrectly estimated, then estimates of model parameters suffer from the incidental parameters problem, noted by Alan and Franco (2007). It is possible to estimate the model considering the individual effects as being fixed or random effects.

This work aimed at presenting a computational procedure for estimation of panel data models with fixed effects and random effects estimators.

In the past, researchers have regarded estimated fixed effects in panel data models as nuisance or ancillary parameters. Despite the varied uses of estimated fixed effects, little is known about the performance of commonly-used panel data estimators with respect to fixed effects. It has been argued that the least squares dummy variable (LSDV) and within estimators produce estimated fixed effects which are unbiased but inconsistent in short panel when regressors are not strictly exogenous. However, there are few practical guides as to the definition of a short panel Cameron and Trivedi (2005). Even less is known a priori about the properties of fixed effects in panel data Judson and Owen (1999). The within estimator could be inconsistent for models in which regressors are only weakly exogenous. In response to these problems, a number of studies have developed alternative two stage least square (2SLS) estimation methods Cornwell et al (1989) and Bulkley et al (2004).The asymptotic properties of this method received minimal attention in most of the literature Semykina and Wooldridge (2008).

A number of studies have addressed the problems of heterogeneity under the assumption of strictly exogenous explanatory variables Verbeek and Nijman (1992), Wooldridge (1995), Kyriazidou (1997), Rochina-Barrachina (1999), Dustmann et el (2007) and *Akossou, et al , 2013.*

When panel data are available,a random effect model can be used to control for these individual differences. Such a model typically assumes that the stochastic error term has two components: a time-invariant individual effect which captures the unobservable individual heterogeneity and the usual random noise term. A serious defect of the within estimator is its inability to estimate the impact of time-invariant regressors. The generalized least square (GLS) estimator is often used in the literature as a treatment of this problem when variance structures are known and Feasible GLS when variances are not known Baltagi and Khanti-Akom (1990) and Green (2012).

Enormous works on the methodologies and applications of panel data model estimation have appeared in the literature see Mundlak (1978), Hausman *et al* (1981), Breusch (1989), Baltagi (2005), Bresson *et al* (2006), Garba *et al* (2013), Kruiniger (2001), Matyas *et al* (2012), Olofin, *et al* (2010) and Semykina and Wooldridge (2008). 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 purpose of this paper is to elucidate the part of the earlier work pertaining to these panel data model estimators. The paper also contributes to the existing literature in several ways. First , we set out the assumptions behind the fixed and random effect approaches, highlight their strengths and weaknesses. Also, we give brief estimation methods and procedures for the models and derive the estimators. We study asymptotic properties of Two stage least square and generalized least square estimators and examine the finite sample properties of estimators with simulation study.

2.0 Estimation Framework and Model Specification

Panel data can not only offer us the information across different individuals, but also the information for a given individual across the time. The panel data model in which the intercept coefficients vary over the individual and slopes remain the same is a kind of classical regression model to present the special feature of panel data. The paper describes two general approaches to the estimation of panel data: fixed and random effect models. These are presented in the paper as separate and rather distinct estimation frameworks for didactic purposes.

2.1 Panel Data Models

As described by Hsiao (1986), there are many benefits in using panel data, the principal ones being: control of the individual heterogeneity; panel data models have a greater variability, less collinearity between variables, more degrees of freedom and more efficiency; they are more capable in identifying and measuring

effects that aren't detected in cross-section or time series data. However, these models also have some limitations, the requirement of more computational resources being the most apparent of them. The general formularization of the panel data model is ;

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

where i denotes individuals and t denotes time. X_{it} is a vector of observations on k explanatory variables, β is a k vector of unknown coefficients, random variable α_i is an unobserved individual specific effects, and ε_{it} is iid over i and t. It has a zero mean random disturbance with variance σ_ε^2.

In panel data models, the individual intercept $\alpha_i$is meant to control for the effect of unobservable regressors that are specific to individual i. The various panel data models depend on the assumptions made about the individual specific effects α_i. 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

This model assumes that differences across units of observation can be captured in the constant term. Each α_i is treated as an unknown parameter to be estimated. It also assumes that there is unit-specific heterogeneity in the model which might be correlated with the regressors and needs to be removed from the regression before estimation. In this model, we estimate parameters for fixed effects between units and thereby remove variance from the error term. Hence, the fixed effects estimation method eliminates the time invariant unobserved effect. However, if the number of units is large, the estimation of the parameters may be inefficient. If the individual effects are randomly distributed in each cross sectional unit, the fixed effects model gives inconsistent estimates and hence, we use random effect instead.

2.1.2 Random Effect Model

In regression analysis it is commonly assumed 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. Thus, this leads to the assumption that the α_i are random factors, independently and identically distributed over individuals and treated as an error term. In this model, it´s necessary to assume that the explanatory variables are uncorrelated to the specific term for each cross sectional unit. The gain to this approach is that it substantially reduces the number of parameters to be estimated.

3.0 Estimation

3.1 Fixed Effects Model

One variant of model (1) is called the fixed effects 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 is given by Megersa et al (2014) and the transformed equation is given by

$$\ddot{y}_{it} = \ddot{X}'_{it}\beta + \ddot{\varepsilon}_{it}, \quad i = 1, \ldots \ldots N, t = 1, \ldots \ldots T \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

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. The OLS estimator obtained from (2) is often called the within estimator. Consistent estimation of this estimator requires X_{it} to be strictly exogenous i.e. $E(\varepsilon_{it}|x_{i1}, \ldots x_{iT}, \alpha_i) = 0$. However, when this strong assumption is not satisfied, the within estimator is no longer consistent. We suspect the correlation between α_i and X_{it} will result in an endogeneity problem. Hence, two-stage least square is the treatment for this problem.

3.1.1 Two-Stage Least Square estimation (2SLS)

In a 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 the response variable is determined jointly with X_{it}. When X_{it} is endogenous or jointly determined with y_{it}, then the estimation of the model will result in inconsistent estimators and enlarge variance of estimators. This endogeneity problem is the consequence of omitted variables. 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. Z_{it} is uncorrelated with the error ε_{it}. Z_{it} is correlated with the regressor X_{it}. Now, 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} that cut correlation. Now assume a model with one endogenous explanatory variable X_K, $Y_{it} = X_{it}\beta + \varepsilon_{it}$ assume $E(\varepsilon_{it}) = 0$, $Cov\,(X_K, \varepsilon_{it}) = 0$, $i = 1, 2 \ldots, K-1$ 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. 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 $\ddot{\varepsilon} = (\ddot{\varepsilon}_{i1}, \ddot{\varepsilon}_{i2}, \ldots, \ddot{\varepsilon}_{iT})$.The detailed derivation of two stage least square estimation is given in Megersa et al (2014). Suppose that Z has the same number of variables as X, i.e. $L = K$. The instrumental variable estimator is given as;

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

But, the better way in order to get a consistent estimate is to use all available instruments. Now, Z_h has a greater number of variables than X_K, i.e. $L > K$.The two stage least square estimation is given as derived in Megersa et al (2014).

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

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

And the asymptotic variance of $\hat{\beta}_{2SLS}$

$$\mathrm{Avar}\left(\hat{\beta}_{2SLS}\right)$$
$$= \sigma^2\left[\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}'_{it}\ddot{Z}_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{Z}'_{it}\ddot{Z}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{Z}'_{it}\ddot{X}_{it}\right)\right]^{-1}$$

where σ^2 can be consistently estimated by $\hat{\sigma}^2 = (NT-K)^{-1}\hat{\ddot{\varepsilon}}'\hat{\ddot{\varepsilon}}$.

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.

3.2 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 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.

$$E\left(\alpha_i \middle| X_{it}\right) = 0 \quad \text{.....(3)}$$

This model is called the random effects model, which usually makes the additional assumptions that $\alpha_i \sim NIID(\alpha, \sigma_\alpha^2)$, $\varepsilon_{it} \sim NIID(0, \sigma_\varepsilon^2)$. So that both the random effects and the error term in (1) are assumed to be i.i.d. (Cameron and Trivedi,2005). Thus, we write the random effects model as

$$y_{it} = X_{it}'\beta + v_{it} \quad i = 1,2,\dots\dots N\ ;\ t = 1,2,\dots\dots T \quad \text{.....(4)}$$

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

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 $\text{Cov}(v_{it}, v_{js}) = E(v_{it}, v_{js})$ 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

$$var\,(v_{it}) = \sigma_\varepsilon^2 I_T + \sigma_\alpha^2\, i_T\, i_T' = \Omega \text{ .....(5)}$$

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 with the fixed effect models, random effects models are relatively difficult to estimate. This document assumes panel data are balanced.

3.2.1 Generalized Least Squares (GLS)

It is well known that the omission of an explanatory variable(s) or use 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, the variance of the 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.
Therefore, to derive GLS we need to focus only on T-dimensional relationship,

$$y_i = X_i\beta + i_T\alpha_i + \varepsilon_i \text{ .....(6)}$$

setting $v_i = i_T\alpha_i + \varepsilon_i$, the 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.

Therefore, as noted by Megersa et al (2014) the random effect estimator is given by
$\hat{\beta}_{GLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (\,X'\Omega X\,)^{-1}X'\Omega\, y\ =$
$\left(\sum_{i=1}^N \sum_{t=1}^T X_{it}'\Omega^{-1}X_{it}\right)^{-1} \sum_{i=1}^N \sum_{t=1}^T X_{it}'\Omega^{-1}y_{it}$ .....(7)

However, assumption (3) 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.

The variance of the GLS estimator which is conditional on X_{it} can be obtained using

$$Var\left(\hat{\beta}_{GLS}\right) = (X'PP'X)^{-1} = (X'\Omega^{-1}X)^{-1} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} \quad (8)$$

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 (3) 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 see Megersa et al (2015b).

4.0 Simulation Study

This section presents simulation of the Panel Data Regression Model. We simulate panel data under some specified conditions likely to be encountered in a real life situation. We then use the data to illustrate estimation of the model and the asymptotic properties (consistency and normality) of the estimators as described in the previous section.

4.1 Simulation Set up

We assume a study on crime where the response variable y_{it} is the crime rate. Crime rate is a function of many variables. In this study, we consider one time-variant endogenous regressor; probability of arrest (x_3), and three time-variant exogenous variables; probability of conviction given arrest(x_1), probability of prison sentencing given conviction(x_2) and average duration of a prison sentence(x_4). Further, we generate an instrument, Z, for endogenous regressor which follow standard normal distribution Cornwell and Trumbull (1994) and Cornwell and Rupert (1988).

As described above, the Panel Data Regression Model is;

$$y_{it} = X_{it}'\beta + \alpha_i + \varepsilon_{it}$$

where $i = 1, \ldots . N$ represent individual offenders and $t = 1,2, \ldots , T$ represent time periods. The error-term $\varepsilon_{it} \sim iid\ (0, \sigma_\varepsilon^2)$ and the individual effects $\alpha_i \sim iid\ (0, \sigma_\alpha^2)$. The independent variable X_{it} contains both exogenous and endogenous variables.

We investigate the finite sample asymptotic properties of the 2SLS and GLS, then compare it with pooled OLS and within estimators. Our comparison is based on consistency and standard errors of the estimator pooled OLS, Within, 2SLS and GLS. The disturbances are considered as independent normally distributed random variables independent of the x_{it} values, for pooled OLS, Within and GLS estimators and correlated for 2SLS estimator. The values of *N* were chosen to be 30,50, 100, 200 and T=10 to represent large samples for the number of individuals and fixed time dimension, respectively. We are interested in the performance of the 2SLS and GLS estimators in estimating α and β.

For benchmark design, we consider three exogenous regressor and one endogenous with coefficient $\beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = 1.5,\ \beta_4 = 2$ which enter the equation. Those parameters are set at several different values to allow study of the estimators under conditions where the panel data model was properly specified. The values of parameter are assumed as conducted in Clark and Linzer (2012) , Liu (2010) , Hielke et al (2008) and Peter E. (2004), For each combination of parameters we vary the size of our panel N, the cross-sectional dimension, takes on values of 30 , 50 ,100, 200 and T, the time dimension, is assigned value of 10. Choices of these values are random to show large cross sections.
The four explanatory variables are denoted x_1, x_2, x_3 and x_4 . Here, x_3 is an endogenous variable, where z is an instrumental variable for x_3 and u is an unobserved error term. Thus, x_3 is the function of z and u. In the context of omitted variables, this means that z should have no partial effect on *y* and z should not be correlated with other factors that affect y. This means that z must be related, either positively or negatively, to the endogenous explanatory variable x_3. These variables are described in table 1. The variables are all normally distributed with different means and standard deviations. All variables vary freely in time. All variables and parameters of the model necessary to calculate the dependent variable y were simulated as well: the coefficients of the variables, β_1, β_2, β_3, and β_4 were sampled from a normal distribution according to Clark and Linzer (2012). The mean of the constants, is assumed to be 10 and the variance of its normal distribution is 2. Having determined these variables, the response variable, y, is calculated. The simulations are similar to those conducted by Plumper and Troeger (2007). The settings of the model variables of the simulation study are given in the following table.

Table 1: Description of variables

VARIABLE	PARAMETER
$\boldsymbol{x_1}$	$x_1 \sim N(\mathrm{N} * \mathrm{T}, 0.4, 0.3)$
$\boldsymbol{x_2}$	$x_2 \sim N(\mathrm{N} * \mathrm{T}, 0.67, 0.05)$
Z	$Z \sim N(\mathrm{N} * \mathrm{T}, 0.35, 0.2)$
$\boldsymbol{x_3}$	z + u
$\boldsymbol{x_4}$	$x_4 \sim N(\mathrm{N} * \mathrm{T}, 12, 3)$

To simulate a data-generating process in which observations are clustered by units, we first generate a series of N within-unit means $\bar{X}_i$ and corresponding unit effects α_i. The following table shows the descriptions of assumed true values of parameters used for our simulation studies.

Table 2: Parameters Manipulated in Simulation and Their Assumed Values

Parameter values	**Description**	**Assumed**
N	Number of cross sectional units	30, 50 , 100, 200
T	Time periods	10
$\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$	Parameter	$(0.5, 1, 1.5, 2)$
σ_α	Standard deviation of unit effects	2
σ_ε	Standard deviation of error terms	1

We then draw N observations of X_i within each unit $i = 1,2,\ldots.N$. The total sample size is $N \times T$. Finally, we apply (1) to produce y_i as a linear function of X_i, with slope β and unit-level constant terms α_i. Our simulations considered only balanced panel data. In order to highlight the differences between the usual fixed effects and random effects approaches, we generate our data as typical empirical crime problem.

4.3 Simulation Results

The specific purpose of these simulations is to analyze the finite sample asymptotic properties of the previous panel data model estimators for different values of cross sectional units. For each simulated dataset, we estimate the fixed effects and the random effects model estimators and record the estimates of betas produced by each methods.

In particular, we focus on investigating the asymptotic properties of within, 2SLS, and GLS estimators which depend crucially on fixed T and large N for static panel data model. We examine how the fixed-effect and random effect consistency associated with each of these estimators varies across cross sectional dimension for fixed time dimension. This section reports the results of simulation designed to investigate the finite sample relative consistency of OLS, Within, 2SLS and GLS estimators. In assessing the performance for the these estimators, an examination of the means and standard deviations of the estimates of parameters was made. The simulation results of each estimator were reported in table 3 consisting of different values of cross sectional units 30, 50, 100,200 and fixed time dimension T=10.

Table 3: Simulated Results for Panel Data model estimators

N=30 , T=10								
Estimates	**Pooled OLS**		**Fixed effects**		**2SLS**		**Random effects**	
	β	Se	β	Se	β	Se	β	Se
β_1	0.4118	0.19887	0.47863	0.10912	0.4865	0.10412	0.684976	0.13029
β_2	0.5007	0.59488	0.42513	0.59975	0.49723	0.57975	0.582934	0.652202
β_3	1.3624	0.1601	1.3024	0.15311	1.26831	0.13493	1.026521	0.148384
β_4	1.496	0.02998	1.48601	0.01057	1.45601	0.01046	1.496028	0.011769
σ_ε^2							0.3912	
σ_α^2							0.0544	
θ							0.353	
N=50,T=10								
Estimates	**Pooled OLS**		**Fixed effects**		**2SLS**		**Random effects**	
	β	Se	β	Se	β	Se	β	Se
β_1	0.44968	0.09059	0.47761	0.07298	0.47761	0.09298	0.45794	0.08991
β_2	0.654722	0.54692	0.6768	0.5316	0.7768	0.5216	0.44322	0.5128
β_3	1.29895	0.12807	1.40549	0.112	1.2134	0.11375	1.22881	0.12727
β_4	1.495309	0.00903	1.49823	0.0092	1.4952	0.00912	1.4968	0.00896
σ_ε^2							0.3489	
σ_α^2							0.0229	
θ							0.2235	
N=100 , T=10								
Estimates	**Pooled OLS**		**Fixed effects**		**2SLS**		**Random effects**	
	β	Se	β	Se	β	Se	β	Se
β_1	0.69986	0.0902	0.5881	0.0903	0.5241	0.0903	0.6142	0.08856
β_2	0.95136	0.5251	1.1297	0.4036	1.0197	0.5129	0.9993	0.45188
β_3	0.89355	0.13264	0.9663	0.1256	1.2147	0.1145	0.92643	0.1511
β_4	1.43594	0.0087	1.5008	0.009	1.5008	0.008356	1.49767	0.0286
σ_ε^2							0.7037	
σ_α^2							0.0323	
θ							0.172	
N=200 , T=10								
Estimates	**Pooled OLS**		**Fixed effects**		**2SLS**		**Random effects**	
	β	Se	β	Se	β	Se	β	Se
β_1	0.52655	0.08486	0.52807	0.0687	0.5381	0.0763	0.5266	0.0648
β_2	0.76084	0.49742	0.8775	0.3959	0.8975	0.3959	0.7687	0.4373
β_3	1.3127	0.0897	1.30401	0.09368	1.372	0.08357	1.31206	0.0812
β_4	1.4959	0.0083	1.4963	0.00664	1.4983	0.00664	1.4959	0.02629

σ_{ε}^2							0.6617
σ_{α}^2							0.01866
θ							0.117

The results in table 3 present the method of estimation, mean and standard error of estimate of β. Choice of either technique could be justified on the basis of our results, given the size of the standard deviation of the panel data model estimates. In our simulation, we look at mean and standard error of pooled OLS, Fixed effects, 2SLS and GLS based on results given in table 3. As N increases, the mean of Pooled OLS, Fixed effects, 2SLS and Random effects estimators increases with fixed T. Standard errors are generally decreasing as N increases with fixed T for all techniques except for 2SLS.
Results in table 3 indicate that in the presence of endogeneity, the 2SLS estimator has a lower standard error than the fixed effect estimation except for N=50 and T=10. This is an indication of the theoretical result that the variance of the 2SLS estimator is lower than the variance of the fixed effects or within estimator. This also implies 2SLS is consistent when there is an endogenous variable. The results show that the 2SLS performs well for estimating parameters of the model. The random effects or GLS estimator performs well relative to pooled OLS throughout cross sectional units as it has a small standard error. In general, based on our simulation results, the pooled OLS has a high standard error and the 2SLS has smaller standard error compared with all other estimators.

For instance, as N= 100,T=10, the 2SLS estimator of $\beta(=5)$ converges to 4.2593 and the GLS estimator converges to 4.0376.Thus , the 2SLS is a more consistent estimator than the within and the GLS in the presence of an endogeneity problem. The averages for the pooled OLS and the within estimator are 3.98071 and 4.1849 respectively, while their true coefficients value for mean is 5 for N=100,T=10. As we can see the within fixed estimator is outperformed relative to that for the pooled OLS. Increasing the number of individuals data will make the estimators better with fixed time periods.

If theta in table 3 is close to unity, the random effects and fixed effects estimates tend to be close to each other. This is especially the case if T gets large, or the variance of the estimated unit effects gets large as compared to the error variance. However, from our simulation results theta is close to zero rather than one. This indicates that the estimate obtained from fixed effects and random effects are quite different as cross sectional units increases.

As the random effects estimator relies on the strict exogeneity assumption it will produce biased estimation results whenever the unit specific effects are correlated with any of the RHS variables. However, in this case the unit effects do not covary with the explanatory variables and the random effects estimator generates more

efficient results and therefore more reliable point estimates. This finding is important - although the standard error reported by a fixed-effects model is smaller than that reported by random- effects model and the fixed-effects estimate is actually likely to be closer to the parameter of interest beta.

Figure 4.7 shows asymptotic normality of panel data regression model estimators using standard deviation of estimators for values of N=30, 50,100,200 and T=10 given below.

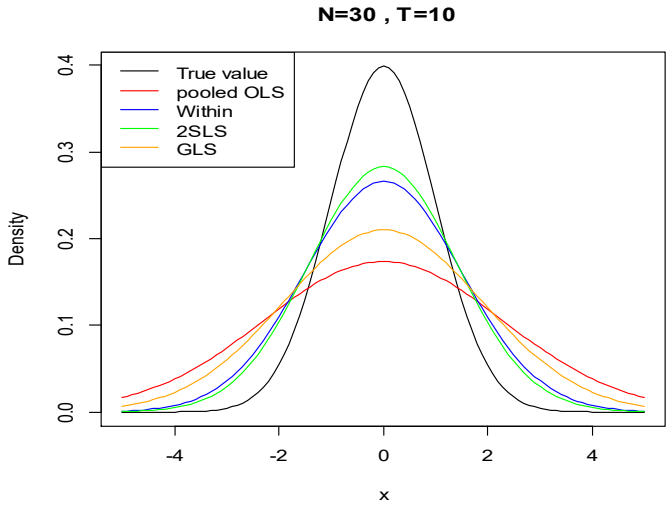

Figure 4.7 a: Distribution of estimators using standard deviation from simulated data for N=30, T=10.

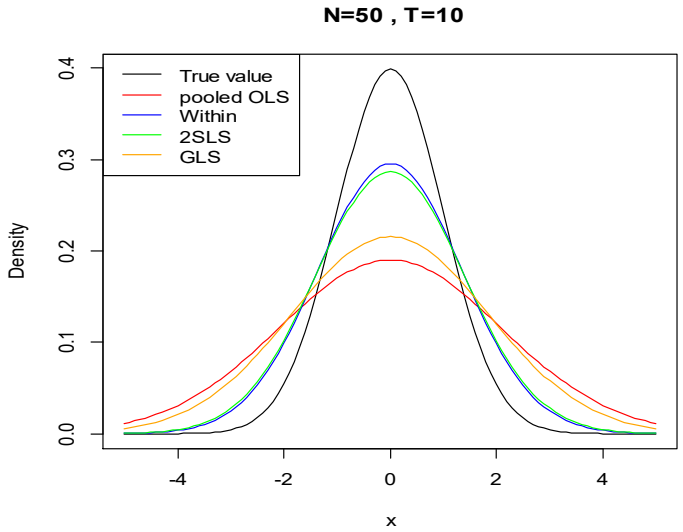

Figure 4.7 b: Distribution of estimators using standard deviation from simulated data for N=50 ,T=10.

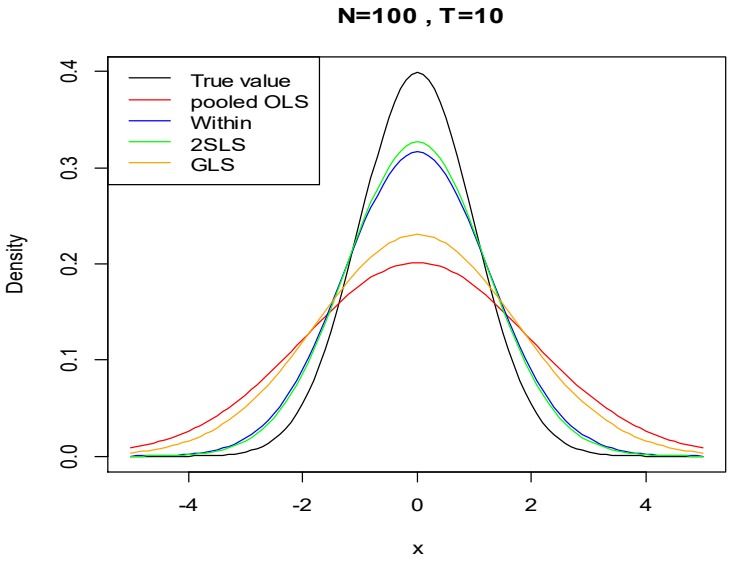

Figure 4.7c: Distribution of estimators using standard deviation from simulated data for N=100 ,T=10.

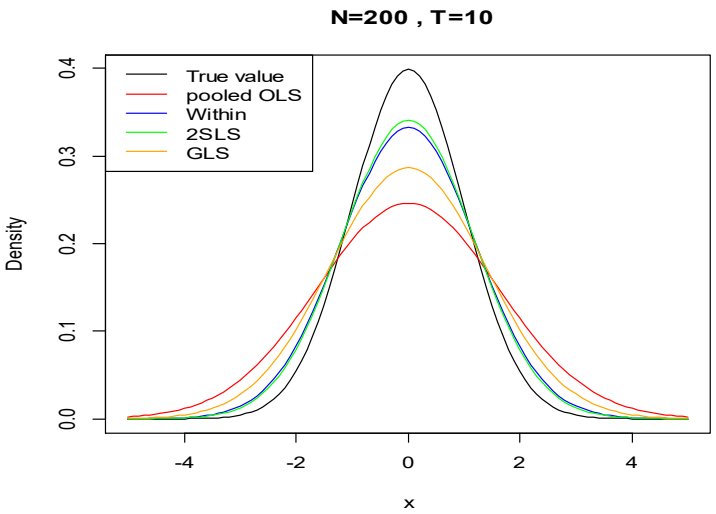

fig. 4.7 d: distribution of estimators using standard deviation from simulated data for N=200, t=10.

Figures 4.7 a, b, c, and d display estimated probability density functions for the panel data model estimators for varied values of individuals. For N=30, T=10, as it can be seen, and as expected, the 2SLS estimator is outperformed by the Within, GLS and Pooled OLS estimators in terms of mean of estimates. For N=50, T=10,the Within estimator is better than the other estimators. In the figure 4.7 c and d 2SLS estimator is more close to true value as N increases. In general, as the number of

cross sections units increase and time dimension is fixed, the panel data estimators become closer and closer to true value.

Figures 4.8 shows the distributions of the Pooled OLS, Within, 2SLS and random effects coefficients estimated from a data generating process with four RHS variables and their standard deviations.

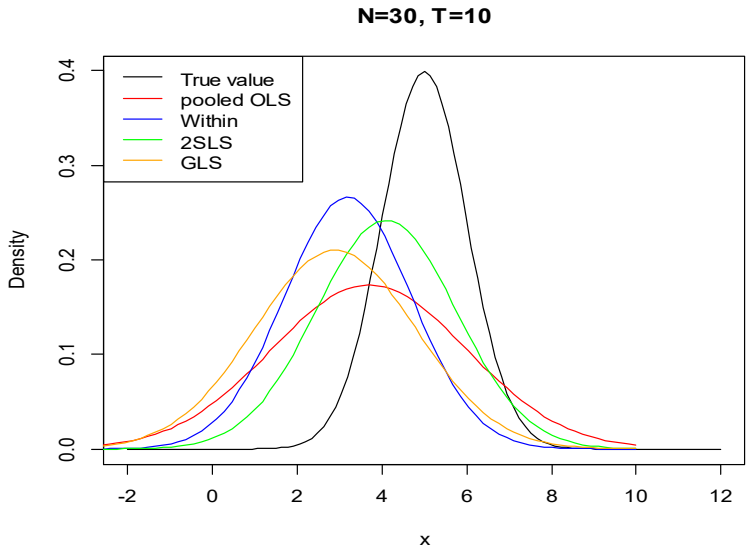

Figure 4.8 a: Distribution of estimators using mean and standard deviation from simulated data for N=30, T=10

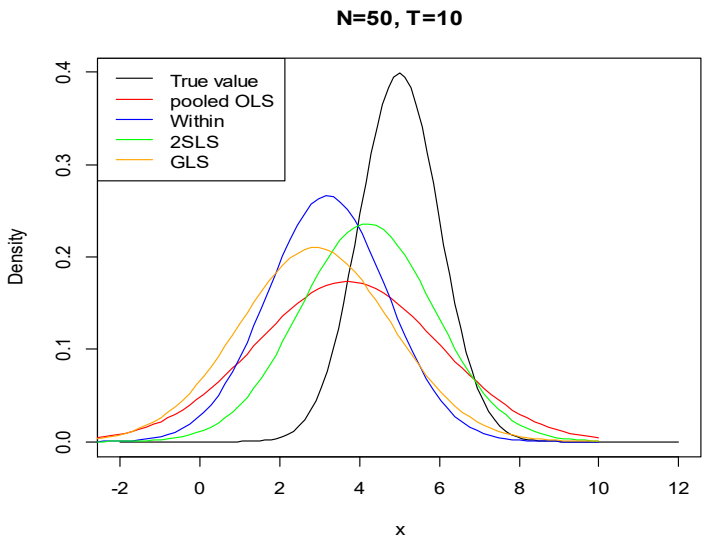

Figure 4.8b: distribution of estimators using mean and standard deviation from simulated data for N=50,T=10

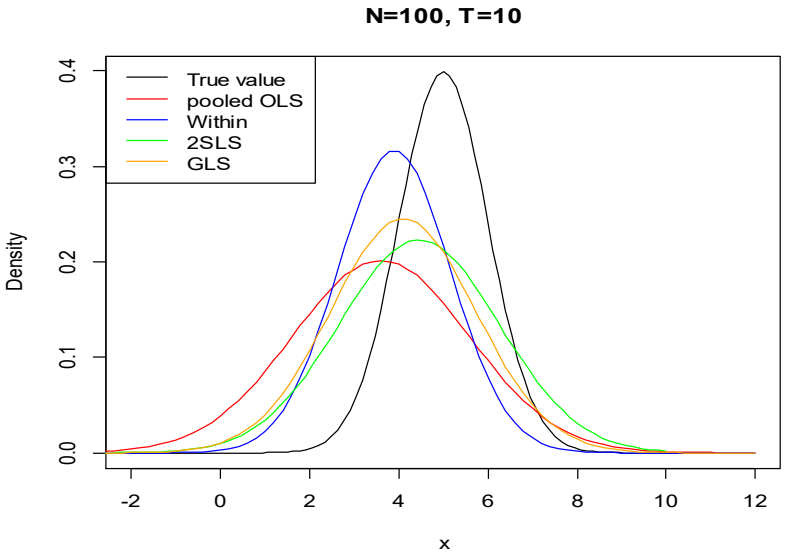

Figure 4.8c: Distribution of estimators using mean and standard deviation from simulated data for N=100,T=10

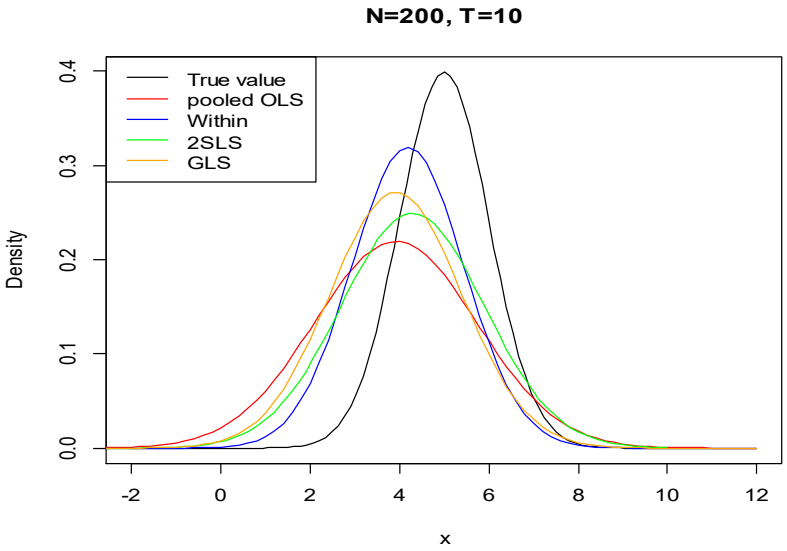

Figure 4.8d: distribution of estimators using mean and standard deviation from simulated data for N=200, T=10

As we can see from figure 4.8 variation across cross sectional units gets smaller as the number of individuals gets larger with fixed time periods. As we expected, the 2SLS estimator produces consistent estimates for true beta of 5 but the distribution is somewhat wider than other estimators. However, the within and GLS estimators are not consistent when there is an endogenous variable and individual effects are

correlated with regressors. The pooled OLS estimator is far away from the true relationship and its distribution is wider than all other estimators. Overall, our simulation results clearly show that as N increases and T is fixed, then the standard error of Pooled OLS, within, 2SLS and GLS estimators decreases.

5.0 Conclusion

Panel data, by blending the inter-individual differences and intra-individual dynamics have advantages over cross-sectional or time series data. The method of panel data has greater capacity for capturing the complexity of human behavior and more accurate inference of model parameters can be obtained through panel data. This work aimed at estimation of panel data regression models with fixed effects and random effects when the equation of interest contains unobserved heterogeneity as well as endogenous explanatory variables, where endogeneity is conditional on the unobserved effect. The assumptions behind the fixed and random effect approaches and their strengths, weaknesses and complications which arise in implementing estimation are also presented. We have departed from the existing literature by deriving and investigating asymptotic properties of panel data model estimators including the 2SLS and GLS estimators for large cross-sections and fixed time periods. In particular, we provided consistency and asymptotic normality of model estimators under specified conditions.

We showed that both estimators are consistent and have asymptotically normal distributions and have different convergence rates dependent on the assumptions of the regressors and the remainder disturbances.

We performed simulations studies to analyze the finite sample asymptotic properties of the model estimators for $N = 30\,, N = 50, N = 100, N = 200$ and $T = 10$. The simulation were made to compute the mean and standard errors of the estimator for the within, 2SLS, OLS and GLS estimators. The summary of results presented in table 3 suggest that all estimators perform well across a range of different panel dimensions. In the presence of endogeneity, the 2SLS estimator performs better relative to the within estimator with large cross-sections. The standard errors of the GLS are smaller than OLS, which is consistent with the theoretical result under exogeneity between individual effects and regressors. The pooled OLS estimator is far away from the true relationship and its distribution is wider than all other estimators in a large sample size. Overall, our simulation results show that the estimated standard error of estimators decreases in large cross-sections with fixed time periods. One of the most important uses of model estimations is to increase understanding of estimators and reduce computational complication 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 the panel data model in analyzing such data. Therefore, there are many important issues such as

modeling of the random intercept model, varying parameter models (Hsiao 1992, 2003; Hsiao and Pesaran 2006), nonparametric or semiparametric approach, bootstrap approach, repeated cross-section data, unrelated regression model, dynamic model, two-way random components, etc, that are not discussed here but are of no less importance. An important avenue of research is to find estimators which are efficient, or nearly so, and yet have better finite sample properties than the existing estimators.

References

Akossou A. Y. and Fonton N. H. (2013). Empirical Comparison of Three Estimators of Collinearity International Journal of Mathematics and Computation , Volume 20, Issue Number 3.

Alan N. and Franco A. (2007). Computational Algorithm For Estimation Of Panel Data Models With Two Fixed Effects.

Baltagi, B. and Khanti-Akom S. (1990). On efficient estimation with panel data: An empirical comparison of instrumental variables estimators, Journal of Applied Econometrics 5, 401-406.

Baltagi, B. H.(2005). Econometrics analysis of panel data, 3rd edition, John Wiley and Sons Ltd, England.

Bresson, G. et al , (2006). Heteroskedasticity and random coefficient model on panel data, Working Papers ERMES 0601, ERMES, University Paris 2.

Breusch, T., Mizon G. and Schmidt P. (1989). Efficient estimation using panel data, Econometrica 57, 695-700.

Bulkley, G.,Harrif R. and Herrerias, R. (2004). Why does book-to-market value of equity forecast cross-section stock returns? *International Review of Financial Analysis* 13,153-160.

Cameron A. and Trivedi P. (2005). *Microeconometrics. Methods and Applications.* Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo.

Clark T. S. and Linzer D. A. (2012). Should I Use Fixed or Random Effects? Political Science Research and Methods Vol 3, No. 2, 399–408.

Cornwell C. and Rupert P. (1988), Efficient estimation with panel data: An empirical comparison of instrumental variables estimators, Journal of Applied Econometrics 3, 149-155.

Cornwell C. and Trumbull W. (1994), Estimating the Economic Model of Crime with Panel Data. The Review of Economics and Statistics, Vol. 76, No. 2. (May, 1994), pp. 360-366.

Dustmann, C. and Rochina-Barrachina M. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. Econometrics Journal 10, 263-293.

Garba, M. K. , Oyejola B. A. and Yahya, W. B (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.

Green. H .(2012). Econometric Analysis, 6th edition, New York University

Hausman J. and Taylor W.(1981). Panel data and unobservable individual effects, Econometrica 49, 1377-1398.

Hielke B., Paul H., Umut O. And Elsabet W.(2008), Fixed Effects Bias In Panel Data Estimators. Discussion Papers, Forschungsinstitut zur Zukunft der Arbeit Institute for the Study of Labor

Hsiao, C. (2003). Analysis of Panel Data, Vol. 34 of Econometric Society monographs. Cambridge University Press, Cambridge, 2nd ed.

Hsiao, C. (1986), Analysis of Panel Data, New York: Cambridge University Press.

Judson, R. and Owen A. (1999). Estimating dynamic panel data models: a guide for macroeconomists. Economics Letters 65, 9-15.

Hsiao, C. (1992). Random coefficient models. In L. M´aty´as and P. Sevestre, eds., The Econometrics of Panel Data, pp. 223–241. Kluwer Academic Publishers, Dordrecht, 1st ed. Reprinted in 2nd ed. (1996) pp. 410–428.

Hsiao, C. and Pesaran, M. (2006). Random coefficients models. In L. M´aty´as and P. Sevestre, eds., The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory, Chap. 5. Kluwer Academic Publishers, Dordrecht, 3rd ed.

Kruiniger J. (2001). On the estimation of panel regression model with fixed effects working paper, Department of Economics , Queen Mary College , University of London, England.

Kyriazidou, E.(1997), Estimation of a panel data sample selection model. Econometrica 65, 1335-1364.

Matyas L., Cecilia H. and Daria P. (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.

Megersa T. J, Chelule J. C. and Odhiambo R. O. (2014). Deriving Some Estimators of Panel Data Regression Models with Individual Effects, International Journal of Science and Research 3, 53-59.

Megersa T.J., Chelule J.C. and Odhiambo R.O. (2015a). On Estimation of Panel Data Regression Model in the Presence of Endogeneity and Heterogeneity. International Journal Mathematics and Computation, CESER PUBLICATIONS Vol. 26, Issue No. 1; 2015.

Megersa T.J, Chelule J.C, and Odhiambo R.O. (2015b). Investigating the Asymptotic Properties of Some Estimators for Panel Data Regression Model with Individual Effects. International Journal of Statistics and Economics, CESER PUBLICATION, Volume 16, Issue Number: 2, 2015.

Mengque Liu (2010), The Hausman test in dynamic panel model. Master thesis.

Mundlak, Y. (1978). On the pooling of time series and cross section data, Econometrica 46, 69-85.

Olofin, S. O., Kouassi, E.W. and Salisu, A. A. (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.

Peter E. (2004), Regressor and random-effects dependencies in multilevel models. Statistica Neerlandica (2004) Vol. 58, nr. 2, pp. 161–178.

Plümper, T. and Troeger T. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis,* 15 (2): 124-39.

Rochina-Barrachina, M.E. (1999), A new estimator for panel data sample selection models. Annales d'Economie et de Statistique 55/56, 153-181.

Semykina A. and Wooldridge M. (2008). Estimating Panel Data Models in the Presence of Endogeneity and Selection. Working paper , Florida State University, USA.

Vella, F. and Verbeek M. (1999). Two-step estimation of panel data models with censored endogenous variables and selection bias, Journal of Econometrics 90, 239-263.

Verbeek, M. and Nijman T. (1992). Testing for selectivity bias in panel data models, International Economic Review 33, 681-703.

Wooldridge J.M. (2002), Econometric Analysis of Cross Section and Panel Data. MIT: Cambridge, MA.

Wooldridge, J.M. (1995). Selection corrections for panel data models under conditional mean independence assumptions, Journal of Econometrics 68, 115-132.