Estimation of Panel Data Regression Models with Individual Effects

Megersa Tadesse Jirata

MS300-0010/12

A thesis submitted to Pan African University, Institute of Basic Sciences, Technology and Innovation in partial fulfillment of the requirement for the degree of Master of Science in Mathematics (Statistics Option)

September 2014

DECLARATION

This thesis is my original work and has not been submitted to any other University for examination.

Signature ……………………………………. Date ………………………….

Megersa Tadesse Jirata

This thesis report has been submitted for examination with our approval as University Supervisors:

Signature ………………………………………………… Date …………………………

Dr. Joel Cheruiyot Chelule

JKUAT, Kenya

Signature ………………………………………………… Date …………………………

Prof. Romanus O. Odhiambo

JKUAT, Kenya

ACKNOWLEDGEMENTS

First of all, I give Glory to God for bringing this far and giving me good health and knowledge while writing this research. My heartfelt gratitude goes to my respectable Supervisor Dr. Joel Cheruiyot Chelule for his unlimited constructive advice, suggestions, ideas and comments in all phases of the thesis. I cannot forget to appreciate my Second Supervisor, Prof. Romanus Odhiambo for his instruction and guidance throughout this research. Furthermore, I appreciate my dear classmates and friends for encouraging me to work hard. Finally, I would like to extend my sincere appreciation and thanks to Pan Africa University and Jomo Kenyatta University for kind assistance in many ways.

DEDICATION

To my mother Dagitu Nagaya, my father Tadesse Jirata, my sisters Mafte Tadesse and Gadise Tadesse my brother Mezgebu Tadesse, my friends Assebe Ragasa, Kasahun Misgana, Garamu Fufa, Yosef Hamba, Tola Bayisa, Chalchisa Desale, Dawit Hapte, Dagaga Bikila, Biranu Worku and Chala Boru.

ABSTRACT

This thesis 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 estimators 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 generalized least squares estimator performs better than ordinary least squares estimator in the absence of correlation between individual effects and the regressors.

TABLE OF CONTENTS

LISTS OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONOMY

2SLS	Two- stage least square
BLUE	Best linear unbiased estimator
BLUP	Best Linear unbiased predictor
CLRM	Classical linear regression model
CLT	Central Limit Theorem
DGP	Data-generating process
DOLS	Dynamic ordinary least square
FE	Fixed effects
FGLS	Feasible generalized least square
FMOLS	Fully modified ordinary least square
GLS-	Generalized least square
GMM	Generalized method of moments
HT	Hausman and Taylor
I.I.D	Independently and identically distributed
IV	Instrumental variable
LHS	Left hand side
LLN	Law of large numbers
LSDV	Least square dummy variable
MLE	Maximum likelihood estimation
N	Cross-section entities and T-Time periods
NLS	National Longitudinal Survey
OLS	Ordinary least square
plm	Panel linear model
PSID	Panel Study of Income Dynamics
RE	Random effects
RMSE	Root Mean Square error
RHS	Right hand side
R^2	Coefficient of determination
$\otimes$	Kronecker product

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

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). Two well known examples in the U.S. are the PSID and NLS. Since the panel data relate to individuals, firms, regions, states, countries , etc 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.

Due to the increased availability of longitudinal data and recent theoretical advances, use of panel data regression methods have become widely used in applied economics research because they allow researchers to control for unobserved individual time-invariant heterogeneity which is not easily done with pure cross-sectional data. If individual heterogeneity is left completely unrestricted, then estimates of model parameters suffer from the incidental parameters problem, first noted by Neyman and Scott (1948). 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.

Given the immense interest in testing and estimation of cross sectional and time series data, not much attention has been paid to estimation of panel data models. In this paper we consider the estimation of panel data models containing unobserved individual effects. The two most widely applied panel data model estimation procedures are RE and FE. It is well-known that the consistency of the RE and FE estimators (as the cross section dimension tends to infinity with the time dimension fixed) requires the strict exogeneity of the regressors, but that 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.

A number of works on the methodologies and applications of panel data modelling have appeared in the literature Li and Stengos (1994), Roy (2002), Baltagi et al.,(2005, 2008), Bresson et al.,(2006), Olofin et al.,(2010). 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 and statistical properties of the estimators has minimum attention in many literature.

One of the critical assumptions of the 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 every i and t. Also, the error terms are expected to have the same variance. If this is not satisfied, there is heteroscedasticity (i.e. $ar\ (v_{it}) = var\ (\varepsilon_{it} + \alpha_i) = \sigma^2\Omega$) See Schmidt (2005), Greene (2008), Maddala (2008), Creel (2011) and Wooldridge (2012).

The estimators of Wooldridge (1995), Kyriazidou (1997) and Rochina-Barrachina (1999) help to resolve the endogeneity issues that arise because of non-zero correlation between individual unobserved effects and explanatory variables. However, other endogeneity problem may arise due to a different factor – a nonzero correlation between explanatory variables and idiosyncratic errors. Such type of endogeneity can become an issue due to omission of relevant time-varying factors. The resulting biases cannot be removed via differencing or within transformation, and hence, require special consideration. The approach removes individual effects via within transformation. The estimator is a two stage least squares on the transformed data can achieve unbiased and consistent estimator.

Random effect model treat the individual effects as part of error term, hence variance become non constant. In the presence of heteroscedasticity, the usual OLS estimators, are no longer having minimum variance among all linear unbiased estimators. See Greene (2008), Baltagi et al. (2008), Olofin et al. (2010). Thus, the OLS estimator is not efficient relative to GLS under such situations. The studies of Mazodier and Trognon (1978), Rao et al. (1981), Magnus (1982), Baltagi and Griffin (1988) and Wansbeek (1989) focused on the existence of heteroscedasticity in panel data modelling.

The most popular estimation methods for panel data models are the within and the GLS estimators. For the panel data with large N and small T, the appropriate choice of estimators depends on whether or not regressors are correlated with the unobservable individual effect. An important advantage of using the within estimator (least squares on data transformed into deviations from individual means) is that it is consistent even if regressors are correlated with the individual effect. Some explanatory variables (e.g., years of schooling in the earnings equation) are likely to be correlated with the individual effects (e.g., unobservable talent or IQ). A simple treatment to this problem is the within estimator which is equivalent to least squares after transformation of the data to deviations from means.

However, the within method has two serious defects. First, the within transformation of a model wipes out time invariant regressors as well as the individual effect, so that it is not possible to estimate the effects of time-invariant regressors on the dependent variable. The GLS estimator is often used in the literature as a treatment of this problem. The consistency of the GLS crucially depends on a strong assumption that no regressor is correlated with the effect (random effects assumption). Second, consistency of the within estimator requires that all the regressors in a given model be strictly exogenous with respect to the random noise. The within estimator could be inconsistent for models in which regressors are only weakly exogenous, such as endogenous regressors. In response to these problems, a number of studies have developed alternative estimation methods called 2SLS.

The question of whether to use random or fixed effects naturally arises with panel data. The choice between FE and RE estimators continues to generate a hot debate among econometricians. Use of the A Hausman statistic (1978) is commonly used for this purpose (e.g., Hausman and Taylor (1981), Cornwell and Rupert (1988) and Baltagi and Khanti-Akom (1990). estimator thus requires a statistical test that can empirically validate the above strong assumption. The latter statistic is based upon a contrast between the FE and RE estimators, see Hausman (1978) or Baltagi (2001). If this standard Hausman test rejects the null hypothesis that the conditional mean of the disturbances given the regressors is zero, the applied researcher reports the FE estimator. Otherwise, the researcher reports the RE estimator, see Owusu-Gyapong (1986) and Cardellichio (1990) for two such applications.

This thesis extends the literature by studying the asymptotic properties of 2SLS, GLS and the Hausman test for panel data models with large numbers of cross-section (N) and small time-series (T) observations.

The aim of this present thesis is to elucidate the part of the earlier work pertaining to these panel data model estimators. The thesis also contribute 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 method and procedures for the models and derive the estimators. We study asymptotic properties of 2SLS , GLS estimators and Hausman test statistic and examines the finite sample properties of estimators with simulation study.

This thesis is organized as follows. Chapter 1 introduces the panel data model of interest and chapter 2 provides existing literature done by others on panel data models. In chapter 3 , we describes the general model of interest, some basic assumptions, some notation and defines the 2SLS, GLS estimators and the Hausman test. For several simple illustrative models, we derive the asymptotic consistency and asymptotic distributions of the within estimators, the GLS estimators, and the Hausman test statistic. Chapter 4 reports some simulation evidence about the finite sample properties of our estimators. Finally, chapter 5 concludes by reviewing our contribution and making some suggestions for future work.

1.2 Statement of the Problem

Much work on estimation of regression models has been done using time series and cross-sectional data, separately. Time series and cross-sectional data are special cases of panel data which involve observations made on the same individual over time, and observations made on several individuals at the same time point, respectively. When time series and cross-sectional data are combined, we have panel data. That is, Panel data are repeated observations on the same unit over time. A regression model of panel data is called panel data regression model. Not much work has been done on estimation of panel data, especially investigation of asymptotic properties of the estimation. This thesis seeks to estimate a panel data regression model with individual effects, and investigate the statistical properties of the estimators.

1.3 Objectives of the Study

1.3.1 General Objective

To estimate a panel data regression model with individual effects

1.3.2 Specific Objectives

i. To derive estimators of the parameters of the panel data regression model
ii. To investigate some asymptotic properties of the estimators developed in (i)
iii. To investigate the empirical properties of estimators using simulated data

1.4 Significance of the Study

As more and more panel data are available, many scholars, practitioners, and students have been interested in panel data modeling because the longitudinal data have more variability and allows for exploration of more issues than cross- sectional or time-series data alone (Kennedy, 2008). It is apparent that most researchers, and in particular most econometricians and statictians lack basic understanding to properly interpret economic statements. There is, therefore, need to develop estimation techniques of reporting economic statements in an easy way to understand way.

The analysis of panel data allows the model builder to learn about economic processes while accounting for both heterogeneity across individuals, firms, countries, and so on, and for dynamic effects that are not visible in cross sections. More importantly, longitudinal data allow a researcher to analyze a number of important economic questions that cannot be addressed using cross-sectional or time-series data sets and provides a means of resolving the magnitude of econometric problems that often arise in empirical studies.

Therefore, panel data model is known to improve the accuracy and efficiency of model parameter estimates and one can test more sophisticated behavioral models with less restrictive assumptions. At the end of the day, the main question for an applied researcher, as in any panel data setup, is whether to use a fixed effects or a random effects specification to get the answer.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows; Chapter two presents a review of literature relating to our research objectives. In chapter three, we discuss the methodology in which we give a detailed procedure of estimation of the panel data regression models. We consider both the fixedeffects and the random-effects approaches. We derive some estimators of the panel data regression model, and investigate some asymptotic properties of estimators. In particular, we study consistency and asymptotic normality of the estimators. An empirical study to test the suitability of the derived estimators is carried out in chapter four while the last chapter offers conclusions and suggestions for further study, based on this research.

1.6 Conclusion

Chapter one has given an background of the study, the statement of the problem, research objectives, Significance of the study and organization of the thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, we give literature review in which we review some recent studies related to our study as guided by our objectives. This will enable us to gain an insight of our research while avoiding repetition of work already done by others.

2.2 Literature Review

Panel data econometrics is a continuously developing field. The increasing availability of data observed on cross-sections of units (like households, firms, countries etc.) and over time has given rise to a number of estimation approaches exploiting this double dimensionality to cope with some of the typical problems associated with economic data, first of all that of unobserved heterogeneity. Time wise observation of data from different observational units has long been common in other fields of statistics (where they are often termed as longitudinal data). In the panel data field as well as in others, the econometric approach is nevertheless peculiar with respect to experimental contexts, as it is emphasizing model specification and testing and tackling a number of issues arising from the particular statistical problems associated with economic data Croissant and Millo (2011).

The literature on statistical models for panel data has experienced enormous growth over the last twenty years with several recent textbooks focusing on that subfield of econometrics; see Arellano (2005), Hsiao (2003), Halaby (2004), Baltagi (2005), Cameron and Trivedi (2005), Wooldridge (2010) and Green(2012).The growth is mostly focused on developing asymptotically justifiable estimation techniques by making probabilistic assumptions for error terms that allow for certain forms of heterogeneity and dependence. Despite the impressive development of statistical techniques in analyzing panel data Wooldridge (2010), the statistical foundations of panel data modeling are rather weak in so far as the current textbook perspective is inadequate for securing the reliability and precision of inference based on panel data models.

In this section, we will provide a brief overview of the panel data, fixed and random effects models for pooled cross-sectional and time-series data that are commonly used. More detailed accounts are given in above textbooks. Our overview begin with a general model that can represent either the random or fixed effects models depending on the restrictions placed on it. Then, we discuss the random effects model, the fixed effects model, and the Hausman test that distinguishes between them.

2.3 Panel Data

In recent years, panel data has become widely utilized in econometric analysis in many social sciences. Panel data combines cross-sectional and time-series data and, therefore, provides a more appealing structure of data analysis than either cross sectional or time-series data, alone. Although it is more costly to gather, the advantages of this data type include better and more precise parameter estimation due to a larger sample size as well as simplification of data modeling Hsiao (2005).

Panel data analysis refers to data containing time-series for a cross-section or group of people who are surveyed periodically over a given period of time Yaffee (2003). The observations in panel data involve at least two dimensions i.e. a cross-sectional dimension indicated by subscript i and a time-series dimension indicated by subscript t. Panel data analysis have become very popular in the social sciences, having been used in economics to study behavior of firms and wages of people over time as well as in marketing to study market share changes across different market structures Hsiao (2005) and Yaffee (2003).

Panel data analysis has many advantages over analysis using time-series and cross-sectional data alone. For example, the increased sample size due to the utilization of cross-sectional and time-series data improves the accuracy of model parameters' estimates due to a greater number of degrees of freedom and less multicollinearity compared to either cross-section or time-series data alone. Additionally, since panel data contains information on both the inter-temporal dynamics and the individuality of entities, it controls for the effect of missing variables on the estimation results. Finally, panel data allows for identification of previously not identified model specification Hsiao (2005).

2.3.1 Panel Data Arrangement

A panel data set contains N entities or subjects (e.g., firms and states), each of which includes T observations measured at i through t time period. Thus, the total number of observations is NT. Ideally, panel data are measured at regular time intervals (e.g., year, quarter, and month). Otherwise, panel data should be analyzed with caution. A short panel data set has many entities but few time periods (small T), while a long panel has many time periods (large T) but few entities Cameron and Trivedi (2009).

Typical panels involve annual data covering a short span of time for each individual. This means that asymptotic arguments rely crucially on the number of individuals in the panel tending to infinity. Increasing the time span of the panel is not without cost either. In fact, this increases the chances of attrition with every new wave and increases the degree of computational difficulty in the estimation of qualitative limited dependent variable panel data models, Baltagi (1995b).

2 .4 Overview of Panel Data Models

Analysis of panel data requires to take account of the panel specific structure of several observations for each individual. If OLS regression is used, the standard assumptions must be fulfilled Greene (2012). But it is unlikely, that the error terms are uncorrelated between individuals and over time. The two most popular approaches to take account of the special time structure are fixed and random effects models. The fixed effects model assumes that the differences across units can be captured in differences in the constant term which needs to be estimated as parameters. The model can be reformulated by taking the deviation of the mean of all explaining variables instead of including individual specific dummy variables by applying the Frisch-Waugh Theorem Frisch and Waugh (1989). This reformulation has no effect on the results of the estimated parameters but since the number of variables is reduced this formulation has computational advantages.

The most appealing aspect of the fixed effect model is that it is robust to the omission of any relevant time-invariant regressors. On the other hand, time-invariant regressors cannot be estimated because their influence is captured in the individual specific dummy or, in the case of the simplified formulation, because the variables are zero. The second most popular approach is the random effects model. It is assumed that the individual specific effects are uncorrelated with the explaining variables and this specific effects are treated as part of error term Ogunwale, et el (2011).

Data therefore do not carry useful information about the error term. A variance-covariance matrix can be used to describe how much certain observation depend on each other. In a frequentist framework, this is identical to a GLS estimation where the variance covariance matrix for the FGLS can be taken either from fixed effects or OLS regression. For a detailed discussion of random effects model Baltagi (2001) and Lancaster (2004).

The major difference of random effects models from the fixed effects model is that in the former the omitted time-invariant variables are assumed to be uncorrelated with the included time-varying covariates while in the latter they are allowed to correlate Mundlak (1978). The random effects model has the advantage of greater efficiency relative to the fixed effects model leading to smaller standard errors and higher statistical power to detect effects Hsiao (2003). A Hausman test enables researchers to distinguish between the random and fixed effects model Hausman, (1978).

Despite the many desirable features of the random and fixed effects models for longitudinal data there are a number of limitations of the standard implementations that are not fully appreciated by users. First, these models have implicit restrictions that are rarely tested but that if wrong, could bias the estimated effects, Bollen and Brand (2008).

Panel data models are widely used in empirical economics because they allow researchers to control for unobserved individual time-invariant heterogeneity. However, these models pose important technical challenges in panel settings. In particular, if individual heterogeneity is left completely unrestricted, then estimates of model parameters suffer from the incidental parameters problem, first noted by Neyman and Scott (1948). This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which, in turn, bias estimates of model parameters, Greene (2002), Katz (2001) and Hahn and Newey, (2004).

Applied researchers face a wide choice of theoretically acceptable estimators when confronted with a panel data models Ahn and Schmidt (1995), and Arellano and Bover (1995). Typically, the theoretical literature mitigates against some standard estimators on the basis of inconsistency or inefficiency of the estimates. For example, in the popular one-way error component model, the standard OLS estimator is often not recommended in a static panel data model because it yields biased and inconsistent estimators. Nevertheless, the standard GLS estimator as well as the within estimator are frequently applied.

Maddala and Mount (1973) compared OLS, FE, RE and MLE methods using Monte Carlo experiments. They found little to choose among the various FGLS estimators in small samples and argued in favor of methods that were easier to compute.

Taylor (1980) derived exact finite sample results for the one-way error component model ignoring the time-effects. He found the following important results. (1) FGLS is more efficient than the FE estimator for all but the fewest degrees of freedom. (2) The variance of FGLS is never more than 17% above the Cramer-Rao lower bound. (3) More efficient estimators of the variance components do not necessarily yield more efficient FGLS estimators. These finite sample results are confirmed by the Monte Carlo experiments carried out by Baltagi (1981a).

Baillie and Baltagi (1995) derived the asymptotic mean square prediction error for the fixed effects and random effects predictors as well as two other misspecified predictors and compared their performance using Monte Carlo experiments.

Wallace and Hussain (1996) compared the RE and FE estimators of β in the case of nonstochastic (repetitive) x_{it}'s and find that both are (i) asymptotically normal (ii) consistent and unbiased and that (iii) β_{RE} has a smaller generalized variance (i.e., more efficient) in finite samples. In the case of nonstochastic (nonrepetitive) x_{it}'s they find that both β_{RE} and β_{FE} are consistent, asymptotically unbiased and have equivalent asymptotic variance-covariance matrices, as both N and T are large. Under the random effects model, GLS based on the true variance components is BLUE, and all the FGLS estimators considered are asymptotically efficient as N and T tend to infinity.

Some previous studies have examined the large-N and large-T properties of the within and GLS estimators for error-component models. For example, Phillips and Moon (1999) and Kao (1999) establish the asymptotic normality of the within estimator for the cases in which regressors follow unit root processes. Extending these studies, Choi (1998) considers a general random effects model and derives the asymptotic distributions of both the within and GLS estimators which contains both unit-root and covariance-stationary regressors. However, they did not consider the asymptotic properties of the Hausman test.

The estimators of Wooldridge (1995), Kyriazidou (1997) and Rochina-Barrachina (1999) help to resolve the endogeneity issues that arise because of non-zero correlation between individual unobserved effects and explanatory variables. However, other endogeneity biases may arise due to a different factor – a nonzero correlation between explanatory variables and idiosyncratic errors. Such type of endogeneity can become an issue due to omission of relevant time-varying factors, simultaneous responses to idiosyncratic shocks, or measurement error. The resulting biases cannot be removed via differencing or fixed effects estimation, and hence, require special consideration.

Hsiao (2003) and Halaby (2004) provide good summaries of the complications that emerge with the usual random and fixed effects estimators in models with lagged endogenous variables and the corrections needed might discourage researchers from exploring these possibilities. A closely related alternative is that there are lagged effects of a covariate on the dependent variable, yet these lagged variables are rarely considered.

Kao and Chiang (1999) studied the asymptotic distributions for OLS, FMOLS, and DOLS estimators in cointegrated regression models in panel data. They shows that the OLS, FMOLS, and DOLS estimators are all asymptotically normally distributed.

One test for the usefulness of panel data models is their ability to predict. For the RE model, the BLUP was derived by Wansbeek and Kapteyn (1978) and Taub (1979). The derivation was generalized by Baltagi and Li (1992) to the RE model with serially correlated remainder disturbances.

Ahn and Moon (2001) examined the asymptotic properties of the popular within, GLS estimators and the Hausman test for panel data models with both large numbers of cross-section and time-series observations. They found that find that the within estimator is as efficient as the GLS estimator.

More recently, Ogunwale, et el (2011) studied on the use of simulated panel data to compare the performance of two methods of analyzing such data. Features of the ordinary least squares model that uses pooled data and fixed effects of the LSDV model were discussed. Several statistics (R^2, Standard error, t and F distributions) were used in comparing the result of the estimates obtained from the two methods applied to two sample sizes. The comparison of the results showed that the analysis based on the bigger sample is more consistent and efficient than the one based on the smaller sample size and this made the least square dummy variables to be more superior than the pooled data model. The conclusion is that the fixed effects model of LSDV is superior and better in the analysis of panel data.

2.5 Hausman Specification Test

Much of the testing literature pertaining to panel data models builds on the work of Hausman (1978). In general, the Hausman test can be applied anytime an econometric model can be consistently estimated under the alternative hypothesis as well as under the null. The test is based on comparing the two estimates. Since, under the null hypothesis both estimation procedures are consistent, therefore, observing a statistical difference between the two provides evidence against the null.

Fixed versus random effects has generated a continues lively debate in the biometrics literature. In econometrics, see Mundlak (1978) and Ahn and Moon (2001). The random and fixed effects models yield different estimation results, especially if T is small and N is large. A specification test based on the difference between these estimates is given by Hausman (1978). The null hypothesis is that the individual and time-effects are not correlated with the $x_{it}'s$. The basic idea behind this test is that the fixed effects estimator is consistent whether the effects are or are not correlated with the $x_{it}'s$.

Mundalk (1978) argued that the RE model assumes exogeneity of all the regressors and the random individual effects. In contrast, the FE model allows for endogeneity of all the regressors and the individual effects. This all or nothing choice of correlation between the individual effects and the regressors prompted Hausman and Taylor (1981) to propose a model where some of the regressors are correlated with the individual effects. The resulting estimator is called the HT estimator and it is based upon an instrumental variable estimator which uses both the between and within variation of the strictly exogenous variables as instruments.

More specifically, the individual means of the strictly exogenous regressors are used as instruments for the time invariant regressors that are correlated with the individual effects, see Baltagi (2001). The choice of the strictly exogenous regressors is a testable hypothesis. In fact, this is a Hausman test based upon the contrast between the FE and the HT estimators.

Most applications in economics since the 1980s have made the choice between the RE and FE estimators based upon the standard Hausman test. The latter statistic is based upon a contrast between the FE and RE estimators, see Hausman (1978). If this standard Hausman test rejects the applied researcher reports the FE estimator. Otherwise, the researcher reports the RE estimator, see Hausman and Taylor (1981), Owusu-Gyapong (1986), Cornwell and Rupert (1988) and Cardellichio (1990) and Baltagi and Khanti-Akom (1990) for two such applications.

When T is finite and N is large, whether to treat the effects as fixed or random is not an easy question to answer. It can make surprising amount of differences in the estimates of parameters. In fact, when only a few observations are available for different individuals over time, it is exceptionally important to make the best use of the lesser amount of information over time for the efficient estimation of the common behavioral relationship Hsiao(2005).

Choi (2002) made simple generalization of the Hausman test, which is previously considered by Arellano (1993). The Hausman statistic incorporates and tests a specific set of moment restrictions implying that individual means of the time-varying regressors are exogenous.

Moulton (1986) developed Hausman statistic based on linear panel data estimators. More specifically, He performed Monte Carlo experiments to compare the performance of the standard panel data estimators under various designs. The estimators considered are: OLS, FE, RE and the Hausman–Taylor estimators. Their result shows that when there is endogeneity among the regressors, there is substantial bias in OLS and the RE estimators and both yield misleading inference, then FE estimator is preferred. Even when there is no correlation between the individual effects and the regressors, i.e. in a RE , inference based on OLS can be seriously misleading.

Hausman and Taylor (1981) derive testing methods in the context of linear panel data models containing correlated fixed effects where interest lies in the parameters associated with observed time invariant explanatory variables. More specifically they develop an estimation procedure using instruments to estimate the parameters of the observed time-invariant variables. Where the instruments are the within transformations of the time-varying explanatory variables that are assumed to have no relationship with the unobserved component. Therefore, Hausman and Taylor do not rely on instrument from outside the model.

Metcalf (1996) extends procedures of Hausman and Taylor to models containing endogenous variables in addition to the correlated fixed effects. That is, Metcalf requires instruments outside of the model. The test statistic developed is then pertaining to possible correlation between the instruments and the unobserved component.

Ahn and Low (1996) further extend the testing literature regarding panel data models through reformulating the Hausman test in the context of GMM estimation. They note that the Hausman test statistic for testing correlation between the unobserved component and the regressors implies that the individual means or time averages of the regressors are exogenous. Their alternative GMM statistic incorporates a much broader set of moment conditions signifying that each of the time-varying explanatory variables is exogenous. They described that It's not clear how to apply the Hausman test to panel data models containing both correlated unobserved components and endogenous explanatory variables.

The analysis of panel data is a field of econometrics that is experiencing increased methodological progress. Recent contributions include, among others; Elhorst (2010), Elhorst, Piras, and Arbia (2010), Lee and Yu (2010a), Lee and Yu (2010c), Lee and Yu (2010d), Lee and Yu (2010b), Mutl (2006), Mutl and Pfaffermayr (2011), Pesaran and Tosetti (2011). Empirical applications are hindered by the lack of readily available software.

Panel data econometrics is obviously one of the main fields in the profession, but most of the models used are difficult to estimate with R. plm is a package for R which intends to make the estimation of linear panel models straightforward. plm is an R package for the estimation and testing of various panel data specifications. plm provides functions to estimate a wide variety of models and to make (robust) inference Croissant and Millo (2011). They slightly modified version of Croissant and Millo (2008) on plm package.

A very comprehensive software framework for (among many other features) maximum likelihood estimation of linear regression models for longitudinal data, packages nlme (Pinheiro, Bates, DebRoy, and the R Core team 2007) and lme4 (Bates 2007), is available in the R (R Development Core Team (2008)) environment and can be used, e.g., for estimation of random effects panel models, its use is not intuitive for a practicing econometrician, and maximum likelihood estimation is only one of the possible approaches to panel data econometrics.

The aim of this paper is to derive 2SLS and GLS estimators, study their asymptotic property and apply Hausman test to make choice between FE and RE. We consider the implementation of two-stage least square and generalized least square estimators in the context of fixed as well as random effects linear panel data models. We perform comparisons of our estimators using simulation studies.

2.6 Conclusion

In this chapter we have given literature review. The panel data regression model with fixed and random effects and simulation studies have also been reviewed. In the next chapter we give panel data models, estimation procedure, panel model estimators and consistency and asymptotic normality of estimators.

CHAPTER 3

MODEL ESTIMATION

3.1 Introduction

In this chapter, we define the two-dimensional panel data regression model, discuss model estimation procedure, estimation techniques and investigate some asymptotic properties of the estimators.

3.2 Brief Overview of Some Concepts

3.2.1 Multiple Regression

Since a multi-dimensional panel data regression model entails relationship between many independent variables and a response variable, there is need to briefly review multiple regression concepts. In general, a multiple regression model takes the form;

$$Y = f\left(\xi_1, \xi_2, ..., \xi_k\right) + \varepsilon \tag{3.1}$$

where Y is the response variable, $\xi_1, \xi_2, ..., \xi_k$ are the independent variables and ε is an error term representing other sources of variability not accounted for in the function f. This ε may include effects such as measurement errors on the response, background noises and even effects of other variables. It is treated as a statistical error that is normally distributed with zero mean and variance σ^2 i.e. $\varepsilon \sim N\left(0, \sigma^2\right)$.

Consequently;

$$E\left[Y | \xi_1, \xi_2, ..., \xi_k\right] = E\left[f\left(\xi_1, \xi_2, ..., \xi_k\right)\right] + E\left[\varepsilon \,|\, \xi_1, \xi_2, ..., \xi_k\right] = E\left[f\left(\xi_1, \xi_2, ..., \xi_k\right)\right] \tag{3.2}$$

The variables $\xi_1, \xi_2, ..., \xi_k$ in equation (3.1) are called natural variables because they are expressed in the natural units in which the measurements being studied were made. It is convenient to transform these natural variables into coded variables, say $X_1, X_2, ..., X_k$, which are dimensionless with mean zero and same standard deviation. Accordingly, in terms of the coded variables, the response function (3.1) can be written as;

$$Y = f\left(x_1, x_2, ..., x_k\right) + \varepsilon \tag{3.3}$$

where ε are random variables called error terms which are assumed to be identically and independently distributed, independent of X and normally distributed with zero mathematical expectation i.e. $E(\varepsilon)=0$, and constant and finite variance i.e. $Var(\varepsilon)=\sigma^2<\infty$. The explanatory variables X are assumed to be non-random.

Since the true response function f is unknown, it is estimated. The efficiency of the estimation procedure depends on the ability to develop a suitable approximation for this function. This tenability of an efficient approximation is usually the focus in model estimation.

3.2.2 Panel Data

Panel data, also known as longitudinal data, refers to multi-dimensional data normally involving measurements over time. It contains observations on several phenomena obtained over time where time is sub-divided into equal time periods e.g. days, months, years, e.tc. The observations are obtained from the same set of entities or units which may be individuals, households, firms, regions or countries.

Time series and cross-sectional data are special cases of panel data that are in one dimension only. A time series dataset has one panel member or individual whose characteristic(s) of interest are observed over several time periods. On the other hand, cross-sectional data involves one time point at which many panel members or individuals are observed for certain characteristic(s) of interest. In most cases, when $T \gg N$, the pane data set is likely to be a time series data, and when $N \gg T$, the panel data is likely to be a cross-section data. Panel data set, therefore, possess a combination of the characteristics of both time series and cross-sectional data Hasio (2005) .

Another important distinction between time series and cross-sectional data sets is that, given a time dimension t a panel member behaves like a time series; depicting natural ordering, and systematic dependence, but a cross-section data set has no natural ordering.

Given that time series and cross-sectional data sets are special types of panel data, it may be necessary to address the problems that generally afflict time series data i.e. autocorrelation, and cross-sectional data i.e. heteroscedasticity, while analyzing panel data.

Panel members are normally denoted by X_{it} where $i = 1, 2, \dots, N$ represent the individual dimension and $t = 1, 2, \dots, T$ represent the time dimension.

We have balanced and unbalanced panel data sets. A balanced panel data set involves observation of a characteristic of interest on each panel member for each time period in a given time duration while an unbalanced panel data set may miss out some observation(s) on certain time periods.

Panel data possess some advantages over the time series and cross-sectional data. For instance, they are more informative than time series and cross-sectional data because they allow tracking individual histories, reflect dynamics and Granger causality across variables. They are also useful in situations in which one suspects that the outcome variable may depend on some unobservable explanatory variables that are possibly correlated with the observed explanatory variables. If such omitted variables are constant over time, then, panel data estimators allow for consistent estimation of the effect of the unobserved explanatory variables on the response.

Hsiao (1986) also enumerates more benefits of panel data. These include control of the individual heterogeneity; panel data models have greater variability, less collinearity between variables, more degrees of freedom and more efficiency; they are more capable to identify and measure effects that aren't detected in cross-section or time series data.

3.2.3 Panel Data Models

This sub-section presents a brief overview of panel data models. We have static linear, non-linear and dynamic panel data models.

3.2.3.1 Static Linear Panel Data Models

This type of panel data models does not allow inclusion of the lagged , current and future value of dependent variable as one of the regressors for time periods of the same individual. This is a strong assumption which e.g. rules out lagged dependent variables from the model. The linear is the part of the designation relates to the appearance of the regression coefficients. Therefore, the model is linear in parameters in β , individual effect α_i and error term ε_{it} .

3.2.3.2 Static Non-Linear Panel data models

The nonlinearity is defined in terms of the techniques needed to estimate the parameters, not the shape of the regression function . A nonlinear panel data regression model is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.

3.2.3.3 Dynamic linear panel data models

This approach to panel data models involves the use of a dynamic effect, in this case adding a lagged dependent variable to the explanatory variables. these models take into account the dynamic processes by allowing the lagged value of the dependent variable as one of the explanatory variables as well as containing observed and unobserved permanent (heterogeneous) or transitory (serially-correlated) individual differences. The main theoretical reason for the dynamic panel is that it is modelling a partial adjustment based approach. If it is a partial adjustment process, the coefficient on the lagged dependent variable measures the speed of adjustment (i.e. 1 – coefficient is speed of adjustment). In addition the lagged dependent variable can remove any autocorrelation.

When explanatory variables contain lagged dependent variables , because a typical panel contains a large number of cross-sectional units followed over a short period of time, it turns out that how the initial value of the dependent variable is modelled plays a crucial role with regard to the consistency and efficiency of an estimator Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Blundell and Bond (1998).

Each of these models can be either fixed-effect or random-effect i.e. has both fixed and random effects depending some specified assumptions. A fixed effect model examines if intercepts vary across group or time period, whereas a random effect model explores differences in error variance components across individual or time period.

Panel data models are becoming increasingly common as compared to cross-sectional and time-series models due to the advantages of panel data explained in sub-section (3.3.2) above. In addition, panel data models examine group (individual-specific) effects, time effects, or both in order to deal with heterogeneity or individual effect that may or may not be observed.

3.2.4 Definitions of Important Terminologies

Definition 3.1 (Heterogeneity)

Heterogeneity in a panel data model setting implies that panel data model parameters (constant and slope coefficients) vary across individuals.

Definition 3.2 (Homogeneity)

Homogeneity in a panel data model setting implies that panel data model parameters (constant and slope coefficients) are constant across individuals.

Definition 3.3 (Endogeneity)

A variable is said to be endogenous when there is a correlation between the independent variable and the error term in regression model. Endogeneity can arise as a result of measurement error, autoregression with autocorrelated errors, simultaneity and omitted variables.

Definition 3.4 (Exogeneity)

A variable is said to be exogenous when there is no correlation between the independent variable and the error term in the regression model. It is also a factor in a causal model whose value is independent from the states of other variables in the model.

Definition 3.5 (Time series data)

A time series data is a sequence of numerical data points in successive order, measured typically at successive points in time spaced at uniform time intervals. It is a collection of observations of well-defined data items obtained through repeated measurements over time. Quantities that represent the values taken by a variable over a period such as a month, quarter, or year.

Definition 3.6 (Cross-sectional data)

It refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

Definition 3.7 (Instrumental variable)

An instrumental variable is a variable which uncorrelated with the disturbance but is correlated with independent variable in the regression model. It used to estimate causal relationships when controlled variables are not feasible in regression model.

3.3 Notations

The following are the notations used throughout this thesis;

y_{it} − the value of the dependent (continuous) variable for cross-section individual i at time t where $i = 1, \ldots, N$ and $t = 1, \ldots, T$

X_{it}^{j} − the value of the jth explanatory variable for cross-section individual i at time t .There are K explanatory variables indexed by $j = 1, \ldots, K$.

ε_{it} − are called the idiosyncratic errors or idiosyncratic disturbances because these change accross t as well as accross i .

3.4 The Model

In sub-section 3.2.3, we had an overview of three types of panel data models; Static Linear, Non-Linear, and Dynamic Panel data models. Our research, however, focuses on the Static Linear Panel data model. This sub-section, therefore, defines a Linear Panel Data Regression Model. The General Panel Data Regression Model can be written as;

$$Y_{it} = \alpha_i + \beta X_{it}^{'} + \varepsilon_{it}, i = 1, 2, \ldots, N; t = 1, 2, \ldots, T \qquad (3.4)$$

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.

We will assume throughout this thesis that each individual i is observed in all time periods t. This is a so-called balanced panel. The total number of observations thus is NT. The treatment of unbalanced panels is straightforward but tedious. For analysis (and computation), it is useful to organize the observations in vectors in which all the observations for n = 1 are stacked on top of all the observations for n = 2, etc. For panel data models the usual convention is to stack observations in the opposite order of subscripts, that is, first collecting the observations across time for each individual as vector form. The T observations for individual i can be summarized as

$$y_i = X_i\beta + \alpha_i i_T + \varepsilon_i \qquad (3.5)$$

for $= 1,2,\dots N$, where y_i and ε_i are T-vectors and X_i is a $T \times K$ matrix ,

$$\underset{(T\times 1)}{y_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \underset{(T\times K)}{X_i} = \begin{bmatrix} x'_{i1} \\ x'_{it} \\ \vdots \\ x'_{iT} \end{bmatrix} = \begin{bmatrix} x_{1i1} & x_{2i1} & \cdots & x_{ki1} \\ x_{1i2} & x_{2i2} & \cdots & x_{ki2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1iT} & x_{2iT} & \dots & x_{iT} \end{bmatrix}, \underset{(T\times 1)}{i} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \underset{(T\times 1)}{\varepsilon_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}$$

and $\alpha_i = \alpha_i i_T$. Then, stacking the entire data set by individuals,

$$\underset{(NT\times 1)}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{iT} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix}, \underset{(\mathrm{NT}\times \mathrm{k})}{\mathrm{X}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \underset{(NT\times 1)}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}, \underset{(N\times 1)}{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix},$$

$$\underset{(K\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}.$$

Then we can write this as ;

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} i \\ 0 \\ \vdots \\ 0 \end{bmatrix}\alpha_1 + \begin{bmatrix} 0 \\ i \\ \vdots \\ 0 \end{bmatrix}\alpha_2 + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ i \end{bmatrix}\alpha_N + \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

Then the data can be represented by the single (relatively simple) equation by pilling over all observations on top as

$$\mathrm{Y} = \mathrm{X}\beta + \alpha + \varepsilon \qquad (3.6)$$

Most of the paper are concerned with an unobserved effects model defined for a large population. Therefore, we assume random sampling in the cross section dimension. Unless stated otherwise, the asymptotic results are for a fixed number of time periods, T, with the number of cross section observations, N, getting large.

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.

The individual effect is a random variable in both fixed and random effects models and assume that $E(y_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, \alpha_i) = E(y_{it}|X_{it}, \alpha_i) = \alpha_i + X_{it}^{'}\beta$. Individual effect α_i is unkown and can not be consistently estimated in short panels, so we cannot estimate $E(y_{it}|\alpha_i, x_{it})$. Instead, we can eliminate α_i again by taking the expectation with respect to α_i,leading to $E(y_{it} \mid x_{it}) = E(\alpha_i \mid x_{it}) + X_{it}\beta$ .For the random effect model it is assumed that $E(\alpha_i \mid x_{it}) = 0$ i.e. α_i is treated as error term. So, $E(R_{it}|\alpha_i, x_{it}) = X_{it}\beta$ and hence it is possible to idenstify $E(y_{it} \mid x_{it})$. In fixed effect model, however , $E(\alpha_i \mid x_{it})$ varies with x_{it} and it is not known how it varies ,we cannot identify $E(y_{it} \mid x_{it})$.

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 in 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 observations. The major distinction between fixed- and random-effects models rests in the conceptualization and estimation of the individual effect.

3.4.1 Fixed-Effects Model

This approach 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 the fixed effects approach, we estimate parameters for fixed effects between units and thereby remove variance from the error term. Hence, 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 fixed effects approach give inconsistent estimate and hence, we use random effect instead.

If we treat α_i as an unobserved random variable that is correlated with the observed regressors , then we consider these effect as parameters $\alpha_1 , \alpha_2, \ldots \ldots \ldots \alpha_N$ to be estimated. In such cases many estimators such as OLS are inconsistent. Instead, alternative estimation methods that eliminate the α_i are needed to ensure consistent estimation of parameters.

In micro econometric applications, the term fixed effect, does not usually mean that α_i is being treated as nonrandom; rather, it means that one is allowing for arbitrary correlation between the unobserved effect α_i and the observed explanatory variables X_{it} .

3.4.2 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 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.

3.5 Model Assumptions

The following assumptions are made on model(3.4);

PL1: Linearity

The model in (3.4) is linear in parameters in β , individual effect α_i and error ε_{it}. ε_{it} are normally distributed with mean 0 and variance σ_ε^2 i.e. $\varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$ distributed, where $0 < \sigma_\varepsilon^2 < \infty$.

PL2: Independence

The $\{x_{i11}, \ldots x_{iTK}\}_{i=1}^{N}$ is independently and identically distributed. The observations are independent across individuals but not necessarily across time. The ε_{it} are independent and identically distributed ($i.i.d$) random variables.

PL3: Strict Exogeneity

The ε_{it} are independent of the explanatory variables X_{it} as well as the individual-specific time-invariant intercepts α_i i.e. $E(\varepsilon_{it}|x_{i11}, \ldots x_{iTK}, \alpha_i) = 0$ (**mean independence**). The idiosyncratic error term ε_{it} is assumed uncorrelated with the explanatory variables of all past, current and future time periods of the same individual. It also assumes that the idiosyncratic error is uncorrelated with the individual specific effect.

PL4: Error Variance

The covariance between the error terms in any two different observations equals to zero i.e. $\operatorname{cov}\left(\varepsilon_{it}, \varepsilon_{is} \mid X_{it}, \alpha_i\right) = 0$, $\varepsilon_{it} \neq \varepsilon_{is}$ for all i and $s \neq t$.

a) $Var\ (\varepsilon_{it}|x_{i11}, \ldots x_{iTK}, \alpha_i) = \sigma_\varepsilon^2 > 0$ and $< \infty$ for all i, t .Corr $(\varepsilon_{it}, \varepsilon_{is}|x_{i11}, \ldots x_{iTK}, \alpha_i) = 0$ for all i and $s \neq t$ (**homoscedastic and no serial correlation**).

b) $Var\ (\varepsilon_{it}|x_{i11}, \ldots x_{iTK}, \alpha_i) = \sigma_{\varepsilon,it}^2 > 0$ and $< \infty$ for all i, t . $Corr(\varepsilon_{it}, \varepsilon_{is}|x_{i11}, \ldots x_{iTK}, \alpha_i) = 0$ for all i and $s \neq t$ (**no serial correlation**).

c) $Var\ (\varepsilon_{it}|x_{i11}, \ldots x_{iTK}, \alpha_i) = \sigma_{\varepsilon,it}^2 > 0$ and $< \infty$ for all i, t. $Corr(\varepsilon_{it}, \varepsilon_{is}|x_{i11}, \ldots x_{iTK}, \alpha_i) < 1$ and > -1 for all $s \neq t$.

The remaining assumptions are divided into two sets of assumptions: the random effects model and the fixed effects model.

3.5.1 Fixed Effects Assumptions

In the fixed effects model , the α_i is a random variable that is allowed to be with the explanatory variables.

FE1: We have a random sample from the cross sectional dimension.

FE2: There are no time-constant variables (for at least some), and no perfect collinearity in the model variables, that all regressors have non-zero within-variance (i.e. variation over time for a given individual) and not too many extreme values. Hence, x_{it} can not include a constant or any other time-invariant variables.

FE.3: $E(\varepsilon_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i) = 0$, $\forall t = 1,2, \ldots T$, which implies that the following (unfeasible) regression $E(y_{it}|X_{it}, \alpha_i) = X_{it}'\beta + \alpha_i$ (**Strict exogeneity**).

FE.4: $Var(\varepsilon_{it}|X_i, \alpha_i) = E(\varepsilon_i'\varepsilon_i|X_i, \alpha_i) = \sigma_\varepsilon^2 I_T$, $\forall t = 1,2, \ldots T$, where I_T is $T \times T$ identity matrix **(Homoskedasticity).**

FE.5: $Cov(\varepsilon_{it}\, \varepsilon_{is}|X_i, \alpha_i) = 0$ **(no serial correlation).** This implies for all $\neq s$, the idiosyncratic errors ε_{it} are serially uncorrelated across time (conditional on all explanatory variables and α_i).

FE.6: $(X_{i1}, X_{i2}, \ldots, X_{iT}, \varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})$, $i = 1,2, \ldots N$ are i.i.d drawn from their joint distribution. This also indicate $(\ddot{x}_{it1}, \ldots, \ddot{x}_{itK})$ are linearly independent and $Var(\ddot{x}_{itK}) > 0$ and $< \infty$ for all k , where $\ddot{x}_{itk} = x_{itk} - \bar{x}_{ik}$ and $\bar{x}_{ik} = T^{-1} \sum_{t=1}^{T} x_{itk}$.

FE.7: Rank $\left(\sum_{t=1}^{T} E\left(\ddot{X}_{it}'\ddot{X}_{it}\right)\right) = rank\left[E\left(\ddot{X}_i'\ddot{X}_i\right) = K\right]$

With all these assumptions, fixed effect is BLUE, which means it has smaller variance than random effect. However, if we believe that assumption FE 6 is violated (there is Serial Correlation in the error ε_{it}), fixed effect is not BLUE anymore. If there is strong correlation in errors over time (the error term follows random walk process) first Difference will be a better estimator. If Serial correlation is relatively mild, it is not straightforward which one of these estimators to use. As T gets large (relative to N), Serial correlation might be more of a problem, and using fixed effect with long panel data might not be the best strategy.

3.5.2 Random Effect Assumptions

In the random effects model , the individual-specific effect is a random variable that is unrelated with the explanatory variables.

RE.1: Unrelated effects

a) $E(\alpha_i|x_{i1},...x_{iT}) = E(\alpha_i) = 0$ and $Var\ (\alpha_i|x_{i1},...x_{iT}) = \sigma_\alpha^2 < \infty$

b) $E(\alpha_i|x_{i1},...x_{iT}) = 0$ and $Var\ (\alpha_i|x_{i1},...x_{iT}) = \sigma_{\alpha,i}^2(x_{i1},...x_{iT}) < \infty$

It assumes that α_i is uncorrelated with the explanatory variables of all past, current and future time periods of the same individual. Version RE1.a assumes constant variance and (orthogonality of α_i and x_{it}), where $X_i \equiv (X_{1t}, X_{i2}, ..., X_{iT})$. RE1.a always implied by the assumption that the X_{it} are fixed and $E(\alpha_i) = 0$, or by the assumption that α_i is independent of X_i. The important part is $(\alpha_i|X_i) = E(\alpha_i) = 0$; the assumption $E(\alpha_i) = 0$ is without loss of generality, provided an intercept is included in X_{it} .

RE.2: a) $E(\varepsilon_{it}|X_i, \alpha_i) = 0, \quad t = 1,2,...,T$ (**Strict exogeneity**).

b) $E(\varepsilon_i\ \varepsilon_i'|X_i, \alpha_i) = \sigma_\varepsilon^2 I_T$

Under RE.1a , $E(\varepsilon_{it}^2|X_i, \alpha_i) = \sigma_\varepsilon^2\ , t = 1,2,...T$ and $E(\varepsilon_{it}\varepsilon_{is}|X_i, \alpha_i) = 0\ , s \neq t,$ s, t =1,2,...,T (both by iterated expectation argument).

RE3: Identifiability

The $(1, x_{i1},...x_{iT})$ are not linearly dependent and $Var\ (\ddot{x}_{it}) > 0$ and $< \infty$ for all k **.** It assumes that the regressors including a constant are not perfectly collinear, that all regressors (but the constant) have non-zero variance and not too many extreme values.

Assuming PL.2 , PL.4 and RE.1 in the special versions PL.4.a and RE.1.a leads to

$$Var\ (V_{it}|x_{i1},...x_{iT}) = \sigma_v^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2 \text{ for all } i, t$$

$$Cov\ (V_{it}, V_{is}|x_{i1},...x_{iT}) = \sigma_\alpha^2 \text{ for all } i \text{ and } s \neq t$$

$$Cov\left(V_{it}, V_{jt}\middle|x_{i1},...x_{iT}, x_{j1},...x_{jT}\right) = 0 \text{ for all } s, t \text{ and } i \neq j$$

This special case under the a) versions of PL.4 and RE.1 is therefore called the equicorrelated random effects model.

RE.4: Same as FE.1.

RE.5: Same as FE.6

RE.6: Rank $E(X_i'\Omega^{-1}X_i) = K$

3.6 Estimation Techniques

This section provides estimation procedure of our model. That is, we derive estimators of the parameters β and α_i. Since a panel data regression model comprise of both fixed and random effects, we discuss the estimation of each of these panel data models as follows;

3.6.1 Fixed Effect Estimation

In this model, the individual effects α_i ; $i = 1,2,\ldots,N,$ are estimated as a time-invariant set of constants. We treat them as unobserved random variables that are correlated with the observed regressors. If α_i are observed for all individuals, then, the entire model can be treated as an ordinary linear regression model and hence we can estimate it by ordinary least squares method. This is relatively simple if the predictor variables are exogenous and the error terms are homoscedastic and serially uncorrelated.

If α_i are unobserved, but are correlated with the regressors, then, OLS estimator for β will be biased and inconsistent. The particular advantage of the fixed effects model is the removal of the individual - specific heterogeneity from the model and wipe out predictor variables that vary slowly within units. Those effects that do not vary at all over time can't be estimated in fixed effect model.

Consider the panel data model (3.4), we have N equations written as;

$$y_i = \alpha_i i_T + X_i'\beta + \varepsilon_i \tag{3.7}$$

where i_T is a $T \times 1$ vector of ones. The above equation represents a single random draw from a cross section. Fixed effect analysis allows $E(\alpha_i|X_{it})$ to be any function of X_i and therefore, it is more robust than the random-effects analysis. In addition, if α_i can be arbitrarily correlated with each element of X_{it}, there is no way to distinguish the effects of time-constant observables from the time-constant unobservable α_i. For instance, when analyzing individuals, factors such as gender or race cannot be included in X_{it}; when analyzing firms, industry cannot be included in X_{it} unless industry designation changes over time for at least some firms; for cities, variables describing fixed city attributes, such as whether or not the city is near a river, cannot be included in X_{it}. The fact that X_{it} cannot include time-constant explanatory variables is a drawback in certain applications, but when the interest is only on time-varying explanatory variables, it is convenient not to have to worry about modeling time-constant factors that are not of direct interest.

In panel data analysis the term time-varying explanatory variables means that each element of X_{it} varies over time for some cross section units. Often there are elements of X_{it} that are constant across time for a subset of the cross section.

The coefficients on the time-invariant variables cannot be estimated. This lack of identification is the price of the robustness of the specification to unmeasured correlation between the individual effect and the exogenous variables. Thus, this time-invariant unobserved individual or group effect needs to be eliminated or removed before estimation. The particular advantage of the fixed effects model is the removal of the individual -specific heterogeneity from the model which can be shown by deviations-from-means approach. The idea for estimating β is to transform the equations to eliminate the unobserved effect α_i. When at least two time periods are available, there are several transformations that accomplish this purpose. In this section we present only the within transformation.

3.6.1.1 Within Transformation

We eliminate the individual effects α_i by transforming the model. Transformation of the model requires the following steps;

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

$$\bar{y}_i = \bar{X}_i\beta + \alpha_i + \bar{\varepsilon}_i , \quad i = 1, \dots . . N \tag{3.8}$$

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 (3.8) is called between estimator.

Step 2: To eliminate α_i subtract equation (3.8) from (3.4) 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

$$\ddot{y}_{it} = \ddot{X}_{it}'\beta + \ddot{\varepsilon}_{it} , \quad i = 1, \dots \dots N , t = 1, \dots \dots . T \tag{3.9}$$

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 (3.9) by OLS if X_{it} is strictly exogenous .

Equation (3.9) is a regression model in deviations from individual means and does not include the individual effects α_i . The transformation that produces observations in deviation from individual means, as in (3.9), is called the within transformation. The OLS estimator for β obtained from this transformed model is often called the within estimator. Note that time-invariant regressors (e.g. the constant) where $X_{it} = \bar{X}_i$ cancel as $\ddot{X}_{it} = X_{it} - \bar{X}_i = 0$ and their effect cannot be estimated by the within estimator.

Essentially, the fixed effects model concentrates on differences within individuals. That is, it is explaining to what extent y_{it} differs from $\bar{y}_i$ and does not explain why • $\bar{y}_i$ is different from $\bar{y}_j$. The parametric assumptions about β on the other hand, impose that a change in X_{it} has the same (ceteris paribus) effect, whether it is a change from one period to the other or a change from one individual to the other. When interpreting the results, however, from a fixed effects regression, it may be important to realize that the parameters are identified only through the within dimension of the data.

By exploiting the within dimension of the data (differences within individuals), determined as the OLS estimator in a regression in deviations from individual means. It is consistent for β for $T \longrightarrow \infty$ or $N \longrightarrow \infty$ and $TN \longrightarrow \infty$ provided that $E\left(\ddot{X}_{it}\, \varepsilon_{it}\right) = 0$. Again this requires the X-variables to be strictly exogenous, but it does not impose any restrictions upon the relationship between α_i and X_{it}. When individual specific effects are correlated with relevant covariates , it is appealing to prefer the fixed effect over the random effect estimator.

If at least one of regressors are endogenous implies $E(X_{it}\varepsilon_{it}) \neq 0$, OLS produce inconsistent estimate. However, there are also perils relying on the fixed effect only. First , as pointed out , time-invariant variables cannot be used. Furthermore, measurement error in X and endogenous changes in X might lead to biased and inconsistent results also within estimator. When at least one of regressors are endogenous implies $E(X_{it}\varepsilon_{it}) \neq 0$, fixed effect estimator is no longer consistent. For this thesis when this is a case we apply 2SLS to estimate parameters consistently.

3.6.1.2 Two Stage Least Square Estimation

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

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.

Assumptions

(1) Z_{it} is uncorrelated with the error ε_{it}

(2) Z_{it} is correlated with the regressor X_{it}

Consider in (3.4) and (3.7), We allow for arbitrary correlation between the α_i and X_{it}. In addition, we allow some elements of X_{it} to be correlated with the ε_{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 , $Y_{it} = X_{it}\beta + \varepsilon_{it}$. Assume $E(\varepsilon_{it}) = 0,$ $Cov\,(x_k\,,\varepsilon_{it}) = 0\,, k = 1, 2\, \dots, k-1$ and $Cov\,(x_k\,,\varepsilon_{it}) \neq 0$,for K. where $x_1, x_2, \dots, x_{K-1}$ are exogenous and X_K is endogenous since $Cov\,(x_k\,,\varepsilon_{it}) \neq 0$. To fix the problem , consider z_1 as replacer of an endogenous explanatory X_K satisfies that

$$Cov\,(z_1\,,\varepsilon_{it}) = 0 \qquad \text{and} \qquad \theta_1 = \frac{\partial\, L\,(X_K|1, x_1, x_2, \dots, x_{K-1}\,, z_1\,)}{\partial\, z_1} \neq 0.$$

Thus, $Z = (1,\, x_1, x_2, \dots, x_{K-1}\,, z_1)$. Then, endogenous variable X_K can be written as

$$x_K = \delta_0 + \delta_1 x_1 + \dots + \delta_{K-1}x_{K-1} + \theta_1 z_1 + r_K\,, \theta_1 \neq 0 \qquad (3.10)$$

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

By substituting estimated x_K in the regression model we can estimate the model by usual OLS. Then, the model with IV for an explanatory X_K becomes

$$y = \alpha_0 + \alpha_1 x_1 + \ ... + \ \alpha_{K-1} x_{K-1} + \ \gamma_1 z_1 + v \qquad (3.11)$$

where $v = \varepsilon + \beta_K r_K$, $\alpha_K = \beta_K + \beta_K \delta_K$ and $\gamma_1 = \beta_K \theta_1$

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

Suppose that Z has the same number of variables as, i.e. $L = K$. We assume that the rank of $Z'X$ is K, so now $Z'X$ is square matrix. To obtain instrumental variable estimator multiply the transformed model by $\ddot{Z}'$

$$\ddot{Z}'\ddot{y} = \ddot{Z}'\ddot{X}\beta + \ddot{Z}'\ddot{\varepsilon}$$

Taking expectation

$$E\left(\ddot{Z}'\ddot{y}\right) = E\left(\ddot{Z}'\ddot{X}\right)\beta + E\left(\ddot{Z}'\ddot{\varepsilon}\right)$$

Then , $\beta = E\left(\ddot{Z}'\ddot{X}\right)^{-1} E\left(\ddot{Z}'\ddot{y}\right)$ in population

The expectations $E\left(\ddot{Z}'\ddot{X}\right)$ and $E\left(\ddot{Z}'\ddot{y}\right)$ can be consistently estimated using a random sample on (x_{it}, y_{it}, Z_{it}), and so we can identify the vector β. Given random sample $\{(x_i, y_i, Z_{i1}) : i = 1,2, ..., N\}$ from the population, then the instrumental variable estimator of β is given by;

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

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 (3.9). Let $z_1, z_1, ..., z_M$ be instrumental variables such that $Cov\,(Z_h, \varepsilon_{it}) = 0, h = 1, 2, ..., M$, so that each Z_h is exogenous in (3.4), and assume $(\varepsilon_{it}) = 0$, $Cov\,(x_k, \varepsilon_{it}) = 0, k = 1, ..., k-1$,$Cov\,(x_k, \varepsilon_{it}) \neq 0, for\ K$ and $Cov\,(Z_h, \varepsilon_{it}) = 0, h = 1,2, ..., M$.

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

$$\theta_1 = \frac{\partial\, L\,(X_K | 1,\ x_1, x_2, \dots, x_{K-1}\ , z_1, z_2, \dots, z_M\,)}{\partial\, z_1} \neq 0$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\theta_M = \frac{\partial\, L\,(X_K | 1,\ x_1, x_2, \dots, x_{K-1}\ , z_1, z_2, \dots, z_M\,)}{\partial\, z_M} \neq 0$$

The linear projection of x_K on Z can be written as

$$x_K = \delta_0 + \delta_1 x_1 + \dots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \dots + \theta_M z_M + r_K \tag{3.12}$$

where , by definition $E(r_K) = 0$ and $Cov\,(r_K\,;\ x_1, x_2, \dots, x_{K-1}\ , z_1, z_1, \dots, z_M\,) = 0$
Then , it can be simply fitted by OLS

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

So, we denote $\hat{X} = (\,x_1, x_2, \dots, x_{K-1}, \hat{x}_K\,)$. The two-stage estimation under instrumental variables to an endogenous variable x_K referencing as

$$Y = X\beta + \varepsilon \tag{3.13}$$

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

Multiplying equation (3.11) 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

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

There are two -stage regression for estimation of β

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

$x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M + r_K$. 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 $(1, x_1, x_2 \ldots, x_{K-1}, \hat{x}_K)$

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \beta_K \hat{x}_K + \varepsilon \quad (3.14)$$

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 (only one endogenous explanatory variable).

Let $X_K = Z\pi_K + r_K$, and $X = Z\Pi + r_K$, where $\Pi = (\pi_1, \pi_2, \ldots, \pi_K)$

Multiplying by Z' and taking expectation

$$E\ (Z'X) = E\ (Z'Z)\Pi + E\ (Z' r_K)$$

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

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

Multiplying (3.11) by X^* and taking expectation

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

solving for β gives

$$\beta = [E\ (X^* X)]^{-1}\ E\ (X^* y) \quad (3.15)$$

But, $E\ (X^* X) = E\ ((Z\Pi)' X)$, since $X^* = Z\Pi$

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

$$E\ (X^* y) = E\ ((Z\Pi)' y)$$

$$= E\ (X'Z) \left(E\ (Z'Z)\right)^{-1} E\ (Z' y)$$

Therefore , substituting in (3.15) yields

$$\beta_{2SLS} = \left[E\ (X'Z) \left(E\ (Z'Z)\right)^{-1} E\ (Z' X)\right]^{-1}\ E\ (X'Z) \left(E\ (Z'Z)\right)^{-1} E\ (Z' y)$$

This is estimation of β in population

Estimation of β as in sample is

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

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

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

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

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

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

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

This is called fixed 2SLS estimator.

3.6.1.3 Asymptotic Variance of Fixed Effect -2SLS Estimator

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

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

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

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

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

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

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

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 $\hat{X} = Z(Z'Z)^{-1}Z'X$ 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 ,

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

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

where σ^2 can be consistently estimated by

$$\hat{\sigma}^2 = (NT-K)^{-1}\,\hat{\ddot{\varepsilon}}'\hat{\ddot{\varepsilon}}$$

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

Proof

We have $\hat{\sigma}^{2} = (NT-K)^{-1}\left(\ddot{y} - \ddot{X}\hat{\beta}_{2SLS}\right)'\left(\ddot{y} - \ddot{X}\hat{\beta}_{2SLS}\right)$

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

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

$$+\left[\beta - \hat{\beta}_{2SLS}\right]'(NT-K)^{-1}\,\ddot{X}'\ddot{X}\left[\beta - \hat{\beta}_{2SLS}\right]$$

Therefore,

$$Plim\ \hat{\sigma}^{2} = Plim\ (NT-K)^{-1}\,\varepsilon'\varepsilon + (2)(0)E\left((NT-K)^{-1}\,\ddot{\varepsilon}_i\ddot{X}_i\right) + 0\,.E\left(\ddot{X}_i'\ddot{X}_i\right).0$$

$$= E(\ddot{\varepsilon}_i^{2}) = E\left(\ddot{\varepsilon}_i^{2}\,|\ddot{Z}_i\right) = \sigma^{2}$$

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

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

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

3.6.2 Random Effect Estimation

The random effects model specifies α_i is a group-specific random element and treats as part of the error term , similar to ε_i except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic .Thus, the method assumes $E(\alpha_i|X_{it}) = E(\alpha_i) = 0$. As long as $E(X_{it}v_{it}) = 0$, that is X_{it} are uncorrelated with α_i and ε_{it} , the estimates are consistent.

If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. Thus we write the random effects model as

$$y_{it} = X'_{it}\beta + v_{it} \qquad i = 1,2, \dots \dots N \, ; \, t = 1,2, \dots \dots . T \tag{3.16}$$

where $v_{it} = \alpha_i + \varepsilon_{it}$ is 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. It is assumed that α_i and ε_{it} are mutually independent and independent of X_{js} (for all j and s).

In modern econometric parlance, random effect is synonymous with zero correlation between the observed explanatory variables and the unobserved effect: $cov(X_{it}, \alpha_i) = 0$, $t = 1, \dots, T$. Actually, a stronger conditional mean independence assumption, $E(\alpha_i | X_{i1}, \dots, X_{iT}) = 0$ will be needed to fully justify statistical inference. In applied papers, when α_i is referred to as, say, an individual random effect, then α_i is probably being assumed to be uncorrelated with X_{it} .

The random specification of unobserved effects corresponds to a particular case of variance-component or error-component model, in which the residual is assumed to consist of two components : $v_{it} = \alpha_i + \varepsilon_{it}$.

As suggested by Wooldridge (2001), the fixed effect specification can be viewed as a case in which α_i is a random parameter with $cov(\alpha_i, X'_{it}) \neq 0$, whereas the random effect model correspond to the situation in which $cov(\alpha_i, X'_{it}) = 0$. The variance of y_{it} conditional on X_{it} is the sum of two components:

$$var\,(y_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$$

Under assumptions of random effect model , the variance-covariance matrix of ε_i is equal to:

$$\begin{aligned} var\,(v_{it}) &= E(v_{it}\, v'_{it}) \\ &= E[(\alpha_i + \varepsilon_{it})(\alpha_i + \varepsilon_{it})'] \\ &= \sigma_\varepsilon^2 I_T + \sigma_\alpha^2\, i_T\, i'_T \end{aligned}$$

$$= \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

$$= \Omega \tag{3.17}$$

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.

Let $\Omega = (v_i\, v_i' \,|X)$ be the $T \times T$ matrix given in (3.17) , the disturbance covariance matrix for the full NT observations then is

$$\mathrm{V} = I_T \otimes \Omega = E[v_i\, v_i'] = \begin{bmatrix} \Omega & 0 & \dots & 0 \\ 0 & \Omega & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega \end{bmatrix}$$

where $\otimes$ is Kronecker product.

A random effect model is estimated by GLS when the variance structure is known. Compared to fixed effect models, random effect models are relatively difficult to estimate.

3.6.2.1 Generalized Least Square Estimation

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, 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 GLS and it achieves efficiency by transforming a heteroscedasticity variance covariance matrix into a homoscedastic one. In the systems of equation case this is done by premultiplying the system of equations by Ω^{-1}.The basic idea behind GLS is to transform the observation matrix $[y\, X]$ so that the variance in the transformed model is I_n (or $\sigma^2 I_n$).

If $\mathrm{E}(\alpha_i|X_{it}) = \mathrm{E}(\alpha_i) = 0$, random effect model can be estimated via GLS which transforms the data such that the error terms in the transformed model are uncorrelated across all N

individuals and all time periods T. This GLS achieves efficiency relative OLS. 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,

$$y_i = X_i\beta + i_T\alpha_i + \varepsilon_i \tag{3.18}$$

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

$$\Omega = Var\ (\alpha_i + \varepsilon_i)$$

$$= Var\ (i_T\alpha_i + \varepsilon_i)$$

$$= \sigma_\alpha^2\ i_T\ i'_T\ +\ \sigma_\varepsilon^2 I_T$$

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

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

$$\Omega = T\sigma_\alpha^2 P_{i_T}\ +\ \sigma_\varepsilon^2 I_T$$

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

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

$$= T\sigma_\alpha^2 P_T + \sigma_\varepsilon^2 P_T\ +\ \sigma_\varepsilon^2 I_T\ - \sigma_\varepsilon^2 P_T$$

$$= (\ T\sigma_\alpha^2 + \sigma_\varepsilon^2\ )\ P_T\ +\ \sigma_\varepsilon^2\ (I_T\ - P_T)$$

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

Dividing all by $T\sigma_\alpha^2 + \sigma_\varepsilon^2$ gives,

$$\Omega = (T\sigma_\alpha^2 + \sigma_\varepsilon^2) (P_T + \theta \; Q_T)$$

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

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

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

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 $Var\,(v_i) = \Omega = \sigma^2\Sigma$, 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 $\lambda_1, \lambda_2, \ldots, \lambda_T$, by Cholesky's decomposition , we can write as

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

where Λ is a diagonal matrix with the diagonal elements $(\lambda_1, \lambda_2, \ldots, \lambda_T)$ 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 $PP'\Omega = I_T$, so $P'(P\Omega P') = P'$. Our interest is to make error terms to be i.i.d 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 $\lambda = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}} = 1 - \sqrt{\theta}$. The term λ gives a measure of the relative sizes of the within and between unit variances.

If λ=1 when $\theta = 0$ (T is large; important variation in α_i), we have FE = RE. If λ=0 when $\theta = 1$ $(\sigma_\alpha^2 = 0)$ (small T, very little if any heterogeneity, $Var\ (\alpha_i)$ we have pooled OLS. This implies there is no covariance among observations. Then, $\hat{\beta}_{GLS} \xrightarrow{P} \hat{\beta}_{Pooled}$. Parameter θ can take any value between one and zero, i.e. $0 \le \theta \le 1$.Thus, we can transform (3.18) as

$$(y_{it} - \lambda\bar{y}_i) = (X_{it} - \lambda\bar{X}_i)\beta + (v_{it} - \lambda\bar{v}_i) \tag{3.19}$$

This equation involves quasi-demeaned data. To simplify the model,

we let $y_i^* = P'y_i = (y_{it} - \lambda\bar{y}_i)$, $X_i^* = P'X_i = (X_{it} - \lambda\bar{X}_i)$, $v_i^* = P'v_i = (v_{it} - \lambda\bar{v}_i)$

Then , the transformed model can be written as

$$y_i^* = X_i^*\beta + v_i^* \tag{3.20}$$

$$v_i^* = v_{it}^* = (v_{it} - \lambda\bar{v}_i) = (1-\lambda)\alpha_i + (\varepsilon_{it} - \lambda\bar{\varepsilon}_i)$$

$$E\ (v_{it}^*) = 0\ , \qquad Var\ (v_{it}^*) = \sigma_\varepsilon^2 = I_T\ , \qquad E\ (X_{it}^* v_{it}^*) = 0$$

$$Cov\ (v_{it}^*, v_{is}^*) = Cov\left(v_{it}^*, v_{jt}^*\right) = 0$$

The transformed model is homoscedastic, i.e. conditional variance, $Var\ (v_i^*) = Var\ (P'v_i) = I_T$.

Now, we can write the transformed model as

$$y_i^* = X_i^*\beta + v_i^*$$

$$E\ (v_i^*) = 0\ , \qquad Var\ (v_i^*) = I_T\ , \qquad E\ (X_i^* v_i^*) = 0$$

This transformed model satisfies the classical assumption. Because Ω is assumed to be known, y_i^* and X_i^* are observed data. To obtain GLS estimator we apply usual OLS to transformed model . Therefore, the GLS estimator is given by

$$\hat{\beta}_{GLS} = (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*}$$

$$= (\ X'PP'X\ )^{-1}X'PP'y$$

$$= (\ X'\Omega X\ )^{-1}X'\Omega\, y$$

$$= \left(\sum_{i=1}^{N} X_i^{'}\Omega^{-1}X_i\right)^{-1}\sum_{i=1}^{N} X_i^{'}\,\Omega^{-1}y_i$$

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

is the efficient estimator of β. This estimator is the GLS or Aitken (1935) estimator of β. The variance of GLS estimator which is conditional on X can be calculated using

$$\hat{\beta}_{GLS} = (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*}$$

$$= (X^{*\prime}X^{*})^{-1}X^{*\prime}(\ X_i^{*}\beta + v_i^{*}\ )$$

$$= \beta + (X^{*\prime}X^{*})^{-1}X^{*\prime}v_i^{*}$$

$$\hat{\beta}_{GLS} - \beta = (X^{*\prime}X^{*})^{-1}X^{*\prime}v_i^{*}$$

Therefore ,

$$Var\left(\hat{\beta}_{GLS}\right) = E\left\{\left(\hat{\beta}_{GLS}-\beta\right)\left(\hat{\beta}_{GLS}-\beta\right)'\right\}$$

$$= E\ \{\ (X^{*\prime}X^{*})^{-1}X^{*\prime}v^{*}\ v^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}\ \}$$

$$= (X^{*\prime}X^{*})^{-1}X^{*\prime}E\ (v^{*}\ v^{*\prime})X^{*}(X^{*\prime}X^{*})^{-1}$$

$$= (X^{*\prime}X^{*})^{-1}X^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}$$

$$= (X^{*\prime}X^{*})^{-1}$$

$$= (X'PP'X)^{-1}$$

$$= \left(X'\Omega^{-1}X\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} X_i^{'}\,\Omega^{-1}X_i\right)^{-1}$$

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

All the above result is regarding on the desirable properties of OLS estimator hold when dealing with the transformed model. The GLS estimator is more efficient than OLS estimator. This is a consequence of the Gauss-Markov theorem, since the GLS estimator is based on a model that satisfies the classical assumptions but the OLS estimator is not.

3.7 Asymptotic Properties of the Estimators

We investigate consistency and asymptotic normality of panel data model estimators.

3.7.1 Consistency

Consistency means that when sample size is sufficiently large the estimator will be likely to be very close to the actual parameter value. When an estimator converges in probability to the true value as the sample size increases, we say that the estimator is asymptotically consistent. That is,

$$P\left\{\lim_{n\to\infty} (\hat{\beta})\right\} = \beta$$

This means that as the sample size increases, the distribution of $\hat{\beta}$ degenerates: in the limit, the only possible realization of $\hat{\beta}$ is β.

3.7.1.1 Consistency of the Two Stage Least Square

We now study consistency of 2SLS in single-equation model with several endogenous variables among explanatory variables. From the model as in (3.4) some elements of X may be correlated with error term. To study consistency and normality of the fixed-effect two-stage least square estimator on balanced panel data model, we make the following assumptions.

Assumptions

2SLS.1. $E(\varepsilon_{it}|Z_{it}, \alpha_i) = 0$, $t = 1, 2, \dots, T$

2SLS.2. For some $1 \times L$ vector Z_{it} , $E\,(Z_{it}\varepsilon_{it}) = 0$ the zero conditional mean assumption

(E $(\varepsilon_{it}|Z_{it})$=0)

2SLS.3. Rank condition: (a) rank $(Z'_{it}Z_{it}) = L$ implies rank $\left(\ddot{Z}'_{it}\ddot{Z}_{it}\right) = L$

(b) rank$(Z'_{it}X_{it}) = K$ implies rank$\left(\ddot{Z}'_{it}\ddot{X}_{it}\right) = K$ with $L \geq K$

2SLS.4. $E(\varepsilon_{it}^2 Z'_{it}Z_{it}) = \sigma^2 E(Z'_{it}Z_{it})$, where $\sigma^2 = E(\varepsilon_{it}^2)$ (homoskedasticity)

Theorem 3.1 (Consistency of 2SLS)

Under assumptions 2SLS.3 and 2SLS.4, the 2SLS estimator obtained from a random sample is consistent for β as $n \longrightarrow \infty$.

Proof of Theorem 3.1

Lemma 3.1 (Law of Large Number)

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with common distribution function , each having finite expected value $\mu = E(X_j)$ and $\sigma^2 = var\ (X_j)$. Let we define $S_n = X_1 + X_2 + \ldots + X_n$. Then for any real number $\epsilon > 0$, we have

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. Or equivalently

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1$$

as $n \to \infty$.

Lemma 3.2 (Lindeberg-Levy theorem)

Let $\{ w_i : i = 1, 2, \ldots\}$ be a sequence of independent , identically distributed $G \times 1$ random vectors such that $E\left(w_{ig}^2\right) < \infty$, $g = 1, 2, \ldots G$, and $E\ (w_i) = 0$. Then $\{ w_i : i = 1, 2, \ldots\}$ satisfies the CLT; that is $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i \overset{P}{\to}$ Normal $(0, \mathrm{B})$ where $\mathrm{B} = \mathrm{Var}\ (w_i) = \mathrm{E}(w_i w_i')$ is necessarily positive semidefinite. For our purposes , B is almost always positive definite.

Lemma 3.3 (Slutsky's theorem)

Let $\mathrm{g} : \mathbb{R}^K \longrightarrow \mathbb{R}^{\mathrm{J}}$ be a function continuous at some point $c \in \mathbb{R}^K$. Let $\{ x_N : N = 1, 2, \ldots\}$ be sequence of $K \times 1$ random vectors such that $x_N \overset{P}{\to}$ c. Then $\mathrm{g}\ (x_N) \overset{P}{\to} \mathrm{g}\ (\mathrm{c})$ as $\mathrm{N} \longrightarrow \infty$. In other words, Plim $\mathrm{g}(x_N) =$ g (Plim x_N) if $\mathrm{g}(.)$ is continuous at Plim x_N.

Lemma 3.4 (Continuous mapping theorem)

Let $\{X_N\}$ be a sequence of $K \times 1$ random vectors such that $X_N \overset{d}{\to}$ x. If $g : \mathbb{R}^K \longrightarrow \mathbb{R}^J$ is a continuous function , then $g(x_N) \overset{d}{\to} g(x)$. This lemma tells us that once we know the limiting distribution of x_N, we can find the limiting distribution of many interesting functions of x_N.

Proof of Theorem 3.1

We know that

$\beta = [E\,(X^*X)]^{-1}\; E\,(X^*y)$ and $X^* = E\,(x|z) = Z\Pi$

The fixed effect 2SLS estimator can be written as

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

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

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

But we have $y_{it} = X_{it}\beta + \varepsilon_{it}$

Using straightforward algebra , it can be shown that;

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

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

Applying Law of large number ,

$$P\lim\hat{\beta}_{2SLS} = \beta + Plim\left[\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}X'_{it}Z_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}Z'_{it}Z_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}Z'_{it}X_{it}\right)\right]^{-1}$$

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

If we assume that;

$P\lim \frac{1}{N}Z'Z = \sum_{zz} = E(Z'Z)$, a finite, positive definite matrix (well behaved data),

$P\lim \frac{1}{N}Z'X = \sum_{zx} = E(Z'X)$, a finite, $L \times K$ matrix with rank K(relevance),

$P\lim \frac{1}{N}Z'\varepsilon = 0 = E(Z_{it}\varepsilon_{it})$, (exogeneity).

By the Law of large number, it implies that;

$\lim_{n\rightarrow\infty} P\left(\left|\hat{\beta}_{2SLS} - \beta\right| \geq \epsilon\right) = 0$, for all ϵ

or , equivalently

$$P\lim \hat{\beta}_{2SLS} = \beta$$

Using Slutsky's theorem on probability limit ,we get

If $P\left\{\lim \left(\hat{\beta}_{2SLS}\right)\right\} = \beta$ and $g(.)$ is continuous function , it also holds that

If $P\lim g\left(\hat{\beta}_{2SLS}\right) = g\left(P\lim \hat{\beta}_{2SLS}\right) = g(\beta)$

Therefore , given above assumption 2SLS.3, LLN and Slutsky's theorem,

$P\lim\left(\hat{\beta}_{2SLS} - \beta\right) = [\sum_{zx}, (\sum_{zz})^{-1} \sum_{zx},]^{-1} \sum_{zx} (\sum_{zz})^{-1} E(Z_{it}\varepsilon_{it})$

By assumption 2SLS.2 ($E(Z_{it}\varepsilon_{it}) = 0$);

$$P\left\{\lim \left(\hat{\beta}_{2SLS}\right)\right\} = \beta$$

This implies that $\hat{\beta}_{2SLS}$ estimator is consistent in large samples.

3.7.1.2 Consistency of the Generalized Least Square

In order to get the more consistent estimate of β, we make the following assumptions.

GLS.1 $E(X_i \otimes v_i) = E(X_{it} \otimes v_{it}) = 0$, Kronecker product;

This requires that each elements of v_{it} be uncorrelated with all X_{it} in all equations. Thus, it rules out using lagged y_{it} as X_{it} variable in panel data settings. Note that since we almost always have one of the X_{it} variable be a constant , this implies $E\,(v_{it}) = 0$.

GLS.2 $\Omega = E(v_i\, v_i') = E(v_{it}\, v_{it}')$ is positive definite and $E\,(X_i'\Omega^{-1}X_i) = E\,(X_{it}'\Omega^{-1}X_{it})$ is non singular matrix (has rank K).

GLS.3 $E\,(X_i'\,\Omega^{-1}v_i\,v_i'\,\Omega^{-1}X_i) = E\,(X_i'\Omega^{-1}X_i)$ which also implies
$E\,(X_{it}'\,\Omega^{-1}v_{it}\;v_{it}'\,\Omega^{-1}X_{ti}) = E\,(X_{it}'\Omega^{-1}X_{it})$ where $\Omega = E(v_{it}\;v_{it}')$

This assumption is the consequence of iterated expectations:

$$E\,(X_{it}'\,\Omega^{-1}v_{it}\;v_{it}'\,\Omega^{-1}X_{ti}) = E\left[E\,(X_{it}'\,\Omega^{-1}v_{it}\;v_{it}'\,\Omega^{-1}X_{it}|X_{it})\right]$$

$$= E\left[X_{it}'\,\Omega^{-1}\,E\,(v_{it}\;v_{it}'|X_{it})\Omega^{-1}X_{it}\right]$$

$$= E\left[X_{it}'\,\Omega^{-1}\,\Omega\,\Omega^{-1}X_{it}\right]$$

$$= E\left[X_{it}'\,\Omega^{-1}X_{it}\right]$$

If Ω is diagonal, then panel data model this implies homoskedasticity within each equation (each time periods in the panel data case). The asymptotic results of panel data will mostly focus on the case of fixed T, $N \longrightarrow \infty$.

Theorem 3.2 (Consistency of GLS)

This means that the probability limit $(Plim)$ of $\hat{\beta}_{GLS}$ equals β i.e.

$$\lim_{N\longrightarrow\infty} \Pr\left[\left|\hat{\beta}_{GLS} - \beta\right| < \varepsilon\,\right] = 1 \text{ for any } \varepsilon < 0.$$

Therefore, under assumptions GLS.1 and GLS.2, the GLS estimator $\hat{\beta}_{GLS}$ obtained from a random sample following the population model (3.4) is consistent for β if $E\,(X_{it}'\alpha_i) = 0$.

To prove theorem (3.2), we also use Lemma 3.1, 3.2,3.3 and 3.4

Proof

Recall the definition of random effect estimator

$$\hat{\beta}_{GLS} = (\,X'\Omega^{-1}X\,)^{-1}X'\Omega^{-1}y$$

$$= \left(\sum_{i=1}^{N} X_i' \, \Omega^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i' \, \Omega^{-1} y_i$$

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

But we have $y_{it} = X_{it}'\beta + v_{it}$. Thus;

$$\hat{\beta}_{GLS} = \beta + \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it}$$

Applying law of large number we obtain;

$$P\lim_{NT\to\infty} \hat{\beta}_{GLS} = P\lim_{NT\to\infty} \beta + P\lim_{NT\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} P\lim_{NT\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it}$$

$$= \beta + P\lim_{NT\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} P\lim_{NT\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it}$$

By applying the Law of Law of Number;

$P\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it} = E\left(X_{it}'\Omega^{-1}X_{it}\right) = Q_{X\Omega X}$, a finite positive definite matrix and requires well behaved data.

And

$$P\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it} = E\left(X_{it}'\Omega^{-1}v_{it}\right)$$

a finite positive definite matrix.

It follows from the Slutsky theorem and assumption GLS.2;

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

Therefore,

$$P\lim_{NT\to\infty} \hat{\beta}_{GLS} = \beta + \left[E\left(X_{it}'\Omega^{-1}X_{it}\right)\right]^{-1} E\left(X_{it}'\Omega^{-1}v_{it}\right)$$

The GLS estimator is consistent for β iff $E\left(X_{it}^{'}\Omega^{-1}v_{it}\right)=0$ and $\left(X_{it}^{'}\Omega^{-1}X_{it}\right)=Q_{X\Omega X}$, a positive definite and finite matrix.

Now we must show that plim $\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}^{'}\Omega^{-1}v_{it}\right)=0$. By the LLN, it is sufficient that $E\left(X_{it}^{'}\Omega^{-1}v_{it}\right)=0$. This is where assumption GLS.1 comes in. We can argue that this point informally because $\Omega^{-1}X_{it}$ is a linear combination of X_{it} , and since each element of X_{it} is uncorrelated with each elament of v_{it} , any linear combination of X_{it} is uncorrelated with v_{it}. This can be shown directly using algebra of Kronecker product and vectorization (Theil , 1983). Therefore, under assumption GLS.1,

$$\begin{aligned}\text{vec}\,E\left(X_{it}^{'}\Omega^{-1}v_{it}\right) &= E\left[\left(v_{i}^{'}\otimes X_{it}^{'}\right)\right]\text{vec}\left(\Omega^{-1}\right)\\ &= E\,[(v_{it}\otimes X_{it})^{'}]\,\text{vec}\left(\Omega^{-1}\right)\\ &= 0\,,\ \text{since}\ E\,[(v_{it}\otimes X_{it})^{'}]=E(v_{it}X_{it})=0\end{aligned}$$

The proof of consistency of $\hat{\beta}_{GLS}$ fails if we only make assumption $E(v_{it}X_{it})=0$ does not imply $E\left(X_{it}^{'}\Omega^{-1}v_{it}\right)=0$, except when Ω and X_{it} have special structures. If $E(v_{it}X_{it})=0$ holds , but $E\left(X_{it}^{'}\Omega^{-1}v_{it}\right)=0$ fails , the transformation in equation (3.18) generally induces correlation between X_{it}^{*} and v_{it}^{*} .This can be very important point in panel data applications.

Hence, by assumption GLS.1 , LLN and Slutsky theorem,

$P\lim_{NT\rightarrow\infty}\hat{\beta}_{GLS}=\beta$, this shows that $\hat{\beta}_{GLS}$ is consistent estimator of β .

3.7.2 Asymptotic Normality of the Estimators

3.7.2.1 Asymptotic Normality of Two Stage Least Square

The asymptotic normality of $\sqrt{N}\left(\hat{\beta}_{2SLS}-\beta\right)$ follows from the asymptotic normality of $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Z_{i}^{'}\varepsilon_{i}$ which follows from the central limit theorem under assumption 2SLS.1.

Theorem 3.3 (Asymptotic Normality of 2SLS)

Under assumptions 2SLS.1-2SLS.3 $\sqrt{N}\left(\hat{\beta}_{2SLS}-\beta\right)$ is normally distributed with mean zero and variance matrix $\sigma^{2}\{E(X^{'}Z)[E(Z^{'}Z)]^{-1}E(Z^{'}X)\}^{-1}$.

To prove theorem 3.3 , we make use of the following corollaries in addition to Lemma 3.1, 3.2,3.3 and 3.4

Corollary 3.1

Let $\{ Z_N : N = 1, 2, \ldots \}$ be sequence of random $K \times K$ matrices and let A be a nonradom , invertable $K \times K$ matrix . If $Z_N \overset{P}{\rightarrow}$ A then,

(1) Z_N^{-1} exists with probability approaching one (w.p.a.1)

(2) $Z_N^{-1} \overset{P}{\rightarrow} A^{-1}$ or $Plim\ Z_N^{-1} = A^{-1}$

Proof

Because the determinant is a continuous function on the space of all square matrices , $\det(Z_N) \overset{P}{\rightarrow} \det(A)$. Because A is nonsingular , $\det(A) \neq 0$. Therefore, it follows that $P[\det(Z_N) \neq 0] \longrightarrow 1$ as $N \longrightarrow \infty$. This completes the proof of part 1.

Part 2 requires a convention about how to define Z_N^{-1} and Z_N when Z_N is nonsingular. Let Ω_N be the set of ω (outcomes) such that $Z_N(\omega)$ is nonsingular for $\omega \epsilon \Omega_N$; we just showed that $P(\Omega_N) \longrightarrow 1$ as $N \longrightarrow \infty$. Define a new sequence of matrices by $\tilde{Z}_N(\omega) = Z_N(\omega)$ when $\omega \epsilon \Omega_N$, $\tilde{Z}_N(\omega) = I_K$ when $\omega \notin \Omega_N$. Then, $P(\tilde{Z}_N = Z_N) = P(\Omega_N) \longrightarrow 1$ as $N \longrightarrow \infty$. Then, Because $Z_N \overset{P}{\rightarrow} A$, $\tilde{Z}_N \overset{P}{\rightarrow} A$. The inverse operator is continuous on the space of invertible matrices , so $\tilde{Z}_N^{-1} \overset{P}{\rightarrow} A^{-1}$. This is what we mean by $Z_N^{-1} \overset{P}{\rightarrow} A^{-1}$; the fact that Z_N can be singular with vanishing probability does not affect asymptotic analysis.

Corollary 3.2

If $\{Z_N\}$ is a sequence of $K \times 1$ random vectors such that $Z_N \overset{d}{\rightarrow}$ Normal $(0, V)$ then, for any $K \times M$ nonrandom matrix A, $A Z_N \overset{d}{\rightarrow}$ Normal $(0, A\, V\, A')$.

Proof of Theorem 3.3

To drive the asymptotic normality of 2SLS estimator we write:

$$\hat{\beta}_{2SLS} = \beta + \left[\left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' Z_{it} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} Z_{it}' Z_{it} \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} Z_{it}' X_{it} \right) \right]^{-1}$$

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

The asymptotic distribution of the 2SLS estimators is derived by moving β from LHS to RHS and multiplying by $\sqrt{N}$

$$\sqrt{N}\,(\hat{\beta}_{2SLS} - \beta) = \left[\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}Z_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}X_{it}\right)\right]^{-1}$$

$$\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}Z_{it}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it}\right)^{-1}\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}\varepsilon_{it}\right)$$

Now , by the law of large number we have;

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}Z_{it} \xrightarrow{p} E\left(X'_{it}Z_{it}\right)$$

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it} \xrightarrow{p} E\left(Z'_{it}Z_{it}\right)$$

Thus,

$$Plim\left[\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}Z_{it}\right)\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it}\right)^{-1}\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}X_{it}\right)\right]^{-1}\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}Z_{it}\right)$$

$$\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it}\right)^{-1} = M$$

where $M \equiv \left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1}E\left(Z'_{it}X_{it}\right)\right]^{-1}\ E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1}$ (by corollary 3.1)

Then ,

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

$$\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}Z_{it}\right)^{-1} - M = O_p(1)$$

Further , under assumption 2SLS.1

$$\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}\varepsilon_{it}\right) = E\left(Z'_{it}\varepsilon_{it}\right) = 0$$

And by the central limit theorem (Lemma 3.2)

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}\varepsilon_{it} \overset{d}{\rightarrow} \text{Normal}\,(0, \text{var}\,[Z_{it}\varepsilon_{it}]) \sim \text{Normal}\,(0, W)$$

where W is the $K \times K$ matrix and $W = E\left[\varepsilon_{it}^2 Z'_{it} Z_{it}\right]$

This implies $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}\varepsilon_{it} = O_p(1)$ and so we can write

$$\sqrt{N}\,(\hat{\beta}_{2SLS} - \beta) = M \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} Z'_{it}\varepsilon_{it}\right) + O_p(1)$$

since $O_p(1) + O_p(1) = O_p(1)$

Therefore, by LLN , Slutsky and continuity theorems , the limiting distribution of $\hat{\beta}_{2SLS}$ is

$$\sqrt{N}\,(\hat{\beta}_{2SLS} - \beta) = \left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1} E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1}$$

$$\times N\left(0, E\left[\varepsilon_{it}^2 Z'_{it}Z_{it}\right]\right)$$

$$\overset{d}{\rightarrow} N(0, A)$$

where A is defined as

$$A = \left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1} E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left[\varepsilon_{it}^2 Z'_{it}Z_{it}\right]$$

$$E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1} \quad \text{(by corollary 3.2)}$$

In the special case of homoscedastic error , $E\,(\varepsilon_{it}^2 | Z_{it}) = \sigma_\varepsilon^2$ we have

$$E\left[\varepsilon_{it}^2 Z'_{it}Z_{it}\right] = E\left[E\left(\varepsilon_{it}^2 Z'_{it}Z_{it} | Z\right)\right] = E\left[E(\varepsilon_{it}^2 | Z) Z'_{it}Z_{it}\right] = \sigma_\varepsilon^2\, E\left[Z'_{it}Z_{it}\right]$$

Then, the variance part corrupts to

$$\left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1} E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} \sigma_\varepsilon^2\, E\left[Z'_{it}Z_{it}\right]$$

$$E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1}$$

$$= \sigma_\varepsilon^2 \left[E\left(X'_{it}Z_{it}\right)\ E\left(Z'_{it}Z_{it}\right)^{-1} E\left(Z'_{it}X_{it}\right)\right]^{-1} = B$$

Therefore, the asymptotic distribution of $\hat{\beta}_{2SLS}$ is

$\sqrt{N}\left(\hat{\beta}_{2SLS}-\beta\right) \overset{d}{\rightarrow} \mathrm{N}(0, \mathrm{A})$ in the case of heteroscedastic error

$\sqrt{N}\left(\hat{\beta}_{2SLS}-\beta\right) \overset{d}{\rightarrow} \mathrm{N}(0, \mathrm{B})$ in the case of homoscedastic error

Or equivalently ,

$\hat{\beta}_{2SLS} \overset{d}{\rightarrow} \mathrm{N}\left(0, \frac{1}{\mathrm{N}} \mathrm{A}\right)$ in the case of heteroscedastic error

$\hat{\beta}_{2SLS} \overset{d}{\rightarrow} \mathrm{N}\left(0, \frac{1}{\mathrm{N}} \mathrm{B}\right)$ in the case of homoscedastic error

The consistent estimator for the asymptotic covariance matrix can be obtained as follows

By applying LLN we have

$$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}} \overset{p}{\rightarrow} \mathrm{E}\left(\mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)$$

$$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}} \overset{p}{\rightarrow} \mathrm{E}\left(\mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)$$

$$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \hat{\varepsilon}_{\mathrm{it}}^{2} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}} \overset{p}{\rightarrow} \mathrm{E}\left(\varepsilon_{\mathrm{it}}^{2} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)$$

where $\hat{\varepsilon}_{it}=\mathrm{y}_{\mathrm{it}}-\mathrm{X}_{\mathrm{it}} \hat{\beta}_{2SLS}$

Therefore , the consistent estimator for the asymptotic variance is

$$\hat{A}=\left[\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)^{-1}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{X}_{\mathrm{it}}\right)\right]^{-1}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)$$

$$\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)^{-1} \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \hat{\varepsilon}_{\mathrm{it}}^{2} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)^{-1}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{X}_{\mathrm{it}}\right)$$

$$\left[\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)^{-1}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{X}_{\mathrm{it}}\right)\right]^{-1}$$

In the special case of homoscedastic error , the asymptotic variance corrupts

$$\hat{B}=\hat{\sigma}_{\varepsilon}^{2}\left[\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{X}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{Z}_{\mathrm{it}}\right)^{-1}\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{Z}_{\mathrm{it}}^{\prime} \mathrm{X}_{\mathrm{it}}\right)\right]^{-1}$$

$\hat{\sigma}_{\varepsilon}^{2} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^{2} = \left(y_{it} - X_{it}\hat{\beta}_{2SLS}\right)^{-1}$

Here , by LLN , Slutsky and continuity theorems , consistent estimator of variance can be obtained by

$$\hat{A} \overset{p}{\rightarrow} \left[E\left(X_{it}'Z_{it}\right) \; E\left(Z_{it}'Z_{it}\right)^{-1} E\left(Z_{it}'X_{it}\right)\right]^{-1} E\left(X_{it}'Z_{it}\right) \; E\left(Z_{it}'Z_{it}\right)^{-1} E\left[\varepsilon_{it}^{2} Z_{it}'Z_{it}\right]$$

$$E\left(Z_{it}'Z_{it}\right)^{-1} E\left(Z_{it}'X_{it}\right)\left[E\left(X_{it}'Z_{it}\right) \; E\left(Z_{it}'Z_{it}\right)^{-1} E\left(Z_{it}'X_{it}\right)\right]^{-1}$$

And

$$\hat{B} \overset{p}{\rightarrow} \sigma_{\varepsilon}^{2}\left[E\left(X_{it}'Z_{it}\right) \; E\left(Z_{it}'Z_{it}\right)^{-1} E\left(Z_{it}'X_{it}\right)\right]^{-1}$$

Therefore, By the definition of W it follows from lemma 3.7 and corollary 3.2 that

$\sqrt{N}\left(\hat{\beta}_{2SLS} - \beta\right) \overset{a}{\rightarrow}$ Normal $\left(0, \hat{A}\right)$ in the case of heteroscedastic error

$\sqrt{N}\left(\hat{\beta}_{2SLS} - \beta\right) \overset{a}{\rightarrow}$ Normal $\left(0, \hat{B}\right)$ in the case of homoscedastic error

This completes the proves.

3.7.2.2 Asymptotic Normality of Generalized Least Square

To derive the asymptotic distribution of the generalized least squares estimator, We will make use of central limit theorems and assume that the observations are independent. $\hat{\beta}_{GLS}$ is asymptotically normally distributed if the sequence of properly normalized $\hat{\beta}_{GLS}$ converges in distribution to a multivariate normal random vector. It is also normally distributed under the assumption of normal errors. If the error distribution is unknown , we of course don't know the distribution of the estimator. Assuming the distribution of error is unknown , but the other classical assumptions hold.

Theorem 3.4 (**Asymptotic Normality of GLS**)

Under assumptions GLS.1-GLS.3 $\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right)$ is normally distributed with mean zero and variance matrix $\sigma^{2}\left[E\left(X_{it}\,\Omega^{-1}X_{it}'\right)\right]^{-1}$.

To prove this theorem, we make use the following lemma in addition to lemma 3.1,3.2,3.3 and 3.4.

Lemma 3.5 (Asymptotic equivalence lemma)

Let $\{X_N\}$ and $\{Z_N\}$ be sequences of $K \times 1$ random vectors.If $Z_N \overset{\mathrm{d}}{\to} \mathrm{z}$ and $X_N - Z_N \overset{\mathrm{P}}{\to} 0$, then $X_N \overset{\mathrm{d}}{\to} \mathrm{z}$.

Proof of theorem 3.4

To derive asymptotic normality of random effect estimator we write

$$\hat{\beta}_{GLS} = \beta + \left(X'\Omega^{-1}X \right)^{-1} X'\Omega^{-1}\mathrm{v}$$

Transforming the above equation into

$$\left(\hat{\beta}_{GLS} - \beta\right) = \left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}X_{it}\right)^{-1} \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}v_{it}$$

Then, $\hat{\beta}_{GLS}$ has limit distribution with all mass at β (since $\hat{\beta}_{GLS} \overset{P}{\to} \beta$). To get a nondegenerate distribution inflate $\hat{\beta}_{GLS}$ by $\sqrt{N}$. The GLS estimator is asymptotically normally distributed if the limiting distribution of

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) = \left(\frac{X'\Omega^{-1}X}{N}\right)^{-1} \sqrt{N}\ \frac{X'\Omega^{-1}\mathrm{v}}{N}$$

$$= \left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}X_{it}\right)^{-1} \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}v_{it} .$$

Therefore, recalling that $E(v_{it}X_{it}) = 0$ we have:

$\mathrm{var}(x_i v_i) = E\left(v_i^2 X_i X_i'\right)$,

and applying the Central Limit Theorem yields

$$\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}v_{it} \overset{d}{\to} N\left(0, E\left(v_{it}^2 X_{it}\, \Omega^{-1}X'_{it}\right)\right).$$

and

$$\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}X_{it}\right)^{-1} \overset{d}{\to} Plim \left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}X_{it}\right)^{-1}$$

Further , since $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}v_{it} = O_P(1)$ and

$$\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} X'_{it}\Omega^{-1}X_{it}\right)^{-1} - A^{-1} = O_P(1) \quad \text{with } A \equiv E\left(X'_{it}\, \Omega^{-1}X_{it}\right) \text{ by lemma 3.5}$$

Thus, $\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) = A^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it}\right) + O_P(1)$

Then, we can write the above expression as

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}v_{it}$$

$$\overset{d}{\rightarrow} Plim\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} \times N(0,B), \qquad B = E\left(v_{it}^2 X_{it}\,\Omega^{-1}X_{it}'\right)$$

$$\overset{d}{\rightarrow} N\left[0, Plim\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1} B \times Plim\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'\Omega^{-1}X_{it}\right)^{-1}\right]$$

It then follows from the Cramer theorem that

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \overset{d}{\rightarrow} N\left(0, \left[E\left(X_{it}\,\Omega^{-1}X_{it}'\right)\right]^{-1} E\left(v_{it}^2 X_{it}\,\Omega^{-1}X_{it}'\right)\left[E\left(X_{it}\Omega^{-1}X_{it}'\right)\right]^{-1}\right).$$

this also follows from the asymptotic equivalence lemma that;

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \overset{a}{\rightarrow} \text{Normal}\left(0, A^{-1} B A^{-1}\right)$$

where $A = E\left(X_{it}\,\Omega^{-1}X_{it}'\right)$ and $B = E\left(v_{it}^2 X_{it}\,\Omega^{-1}X_{it}'\right) \equiv E\left(X_{it}'\,\Omega^{-1}v_{it}\,v_{it}'\,\Omega^{-1}X_{ti}\right)$

In practice, $\text{Asym. Cov.}\left(\hat{\beta}_{GLS}\right) = \frac{A^{-1} B A^{-1}}{N}$,

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

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

$$= \left(\frac{X'\Omega^{-1}X}{N}\right)^{-1} = \left(X'\Omega^{-1}X\right)^{-1}$$

and always we use data to estimate A and B.

In the case of the classical regression model this expression simplifies , as it follows from assumption 2 and the law of iterated expectations that ;

$$E\left(v_{it}^2 X_{it}\,\Omega^{-1}X_{it}'\right) = E\left(v_{it}^2 X_{it}\,\Omega^{-1}X_{it}' \middle| X_{it}\right) = E\left[E(v_{it}^2|X_{it})X_{it}\,\Omega^{-1}X_{it}'\right] = \sigma^2 E\left(X_{it}\,\Omega^{-1}X_{it}'\right).$$

So that

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \overset{d}{\rightarrow} N\left(0, \sigma^2\left[E\left(X_{it}\,\Omega^{-1}X_{it}'\right)\right]^{-1}\right)$$

It easily follows that

$$\hat{\beta}_{GLS} \xrightarrow{d} N\left(\beta , \sigma^2 \left[E\left(X_{it}\, \Omega^{-1} X_{it}'\right)\right]^{-1}\right)$$

The robust variance estimation of random effects model is given as follows;

I. Variance estimation under conditional homoskedasticity

We have under the assumption of random sampling and lack of contemporaneous correlation :

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \xrightarrow{d} N\left(0, \left[E\left(X_{it}\, \Omega^{-1} X_{it}'\right)\right]^{-1} E\left(v_{it}^2 X_{it}\, \Omega^{-1} X_{it}'\right) \left[E\left(X_{it} \Omega^{-1} X_{it}'\right)\right]^{-1}\right).$$

If in addition we assume conditional homoskedasticity , i.e. $var\ (v_i|X_i)$ does not depend on X_i , we have

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2 \left[E\left(X_{it}\, \Omega^{-1} X_{it}'\right)\right]^{-1}\right)$$

The asymptotic variance of $\hat{\beta}_{GLS}$ is thus :

$$\frac{V}{N} = \frac{\sigma^2 \left[E\left(X_{it}\, \Omega^{-1} X_{it}'\right)\right]^{-1}}{N}$$

where σ^2 is consistently estimated by $\hat{\sigma}^2$;

$$\hat{\sigma}^2 = \frac{\hat{v}_{it}'\hat{v}_{it}}{NT-K} = \frac{\left(y_{it} - X_{it}\hat{\beta}_{GLS}\right)'\left(y_{it} - X_{it}\hat{\beta}_{GLS}\right)}{N-K}$$, $\hat{\sigma}^2$ is consistent estimator of σ^2.

Proof follows the same procedures as proved under two stage least square estimation.

II. Variance estimation under conditional heteroscedasticity (white formula)

We can calculate the asymptotic variance of $\hat{\beta}_{GLS}$ is

$$\frac{V}{N} = \frac{\left[E\left(X_{it}\, \Omega^{-1} X_{it}'\right)\right]^{-1} E\left(v_{it}^2 X_{it}\, \Omega^{-1} X_{it}'\right) \left[E\left(X_{it} \Omega^{-1} X_{it}'\right)\right]^{-1}}{N}$$

In this case ,a natural estimator of $E\left(v_{it}^2 X_{it}\, \Omega^{-1} X_{it}'\right)$ is ;

$$\frac{1}{N}\sum_{i=1}^{N} \hat{v}_{it}^2\, X_{it}\, \Omega^{-1} X_{it}' = \frac{1}{N}\sum_{i=1}^{N}\left(y_{it} - X_{it}\hat{\beta}_{GLS}\right)^2 X_{it}\, \Omega^{-1} X_{it}'$$

Under the assumption of random sampling and lack of contemporaneous correlation, this estimator is consistent.

Indeed , the LLN implies that ;

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} v_{it}^2 \, X_{it} \, \Omega^{-1} X'_{it} = E\left(v_{it}^2 X_{it} \, \Omega^{-1} X'_{it}\right)$$

Moreover

$$\frac{1}{N} \sum_{i=1}^{N} \hat{v}_{it}^2 \, X_{it} \, \Omega^{-1} X'_{it} = \frac{1}{N} \sum_{i=1}^{N} v_{it}^2 \, X_{it} \, \Omega^{-1} X'_{it} + R$$

where $\lim_{N\to\infty} R = 0$. This comes from consistency of GLS which does not depend on the assumption of homoskedasticity.

A natural estimator of the asymptotic variance of $\hat{\beta}_{GLS}$ is thus ;

$$\frac{\hat{V}_{robust}}{N} = \frac{\left(\frac{1}{N}\sum_{i=1}^{N} X_{it} \, \Omega^{-1} X'_{it}\right)^{-1} \left[\frac{1}{N}\sum_{i=1}^{N}\left(y_{it} - X_{it}\hat{\beta}_{GLS}\right)^2 X_{it} \, \Omega^{-1} X'_{it}\right] \left(\frac{1}{N}\sum_{i=1}^{N} X_{it} \, \Omega^{-1} X'_{it}\right)^{-1}}{N}$$

$$= \left(X'\Omega^{-1}X\right)^{-1} \left[\frac{1}{N}\sum_{i=1}^{N}\left(y_{it} - X_{it}\hat{\beta}_{GLS}\right)^2 X_{it} \, \Omega^{-1} X'_{it}\right] \left(X'\Omega^{-1}X\right)^{-1}$$

This is the robust or white formula. Researchers that are concerned with the possibility of conditional heteroscedasticity usually report this expression. Here also , it is easy to check that $\hat{V}$ is consistent estimator of V.

3.8 The Hausman's Specification Test

Hausman (1978) proposes a general test of specification, that can be applied in the specific context of linear panel models to the issue of specification of individual effects (fixed or random).

The general idea of the an Hausman's test is the following. Let us consider a particular model $f(x;\beta) + \varepsilon$ and particular hypothesis H_0 on this model (parameter, error term etc.). Let us consider two estimators of the K -vector β, denoted $\hat{\beta}_1$ and $\hat{\beta}_2$, both consistent under H_0 and asymptotically normally distributed.

1. Under H_0 , the estimator $\hat{\beta}_1$ attains the asymptotic Cramer-Rao bound.
2. Under H_0 , the estimator $\hat{\beta}_2$ is biased.

By examining the distance between $\hat{\beta}_2$ and $\hat{\beta}_1$, it is possible to conclude about H_0 :

a. If the distance is small, H_0 cannot be rejected.

b. If the distance is large, H_0 can be rejected.

This distance is naturally defined as follows;

$$H = (\hat{\beta}_2 - \hat{\beta}_1)' [Var(\hat{\beta}_2 - \hat{\beta}_1)]^{-1} (\hat{\beta}_2 - \hat{\beta}_1)$$

However, the issue is to compute the variance-covariance matrix $Var(\hat{\beta}_2 - \hat{\beta}_1)$ of the difference between both estimators. Generally we know $Var(\hat{\beta}_2)$ and $Var(\hat{\beta}_1)$, but not $ar(\hat{\beta}_2 - \hat{\beta}_1)$.

Lemma 3.6 (Hausman, 1978)

Based on a sample of N observations, consider two estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ that are both consistent and asymptotically normally distributed, with $\hat{\beta}_1$ attaining the asymptotic Cramer-Rao bound so that $\sqrt{N}(\hat{\beta}_2 - \beta)$ is asymptotically normally distributed with variance-covariance matrix V_1. Suppose $\sqrt{N}(\hat{\beta}_2 - \beta)$ is asymptotically normally distributed, with mean zero and variance-covariance matrix V_2. Let $\hat{q} = \hat{\beta}_2 - \hat{\beta}_2$. Then the limiting distribution [under the null] of $\sqrt{N}(\hat{\beta}_2 - \beta)$ and $\sqrt{N}\,\hat{q}$ has zero covariance, $E(\hat{\beta}_1 \hat{q}') = 0_k$.

Theorem 3.5

From this lemma, it follows that;

$$Var(\hat{\beta}_2 - \hat{\beta}_1) = Var(\hat{\beta}_2) - Var(\hat{\beta}_2)$$

Thus, Hausman suggests using the statistic;

$$H = (\hat{\beta}_2 - \hat{\beta}_1)' [Var(\hat{\beta}_2) - Var(\hat{\beta}_1)]^{-1} (\hat{\beta}_2 - \hat{\beta}_1)$$

or equivalently

$$H = \hat{q}' [Var(\hat{q})]^{-1} \hat{q}$$

Under the null hypothesis, this statistic is distributed asymptotically as central chi-square, with K degrees of freedom.

$$H \xrightarrow[N \to \infty]{H_0} \chi^2(K)$$

Under the alternative, it has a noncentral chi-square distribution with noncentrality parameter $\tilde{q}'[\,Var\ (\hat{q})\ ]^{-1}\ \tilde{q}$,where $\tilde{q}$ is defined as follows:

$$\tilde{q} = P \lim_{H_{1/N \to \infty}} \left(\hat{\beta}_2 - \hat{\beta}_1\right)$$

Now, apply the Hausman's test to discriminate between fixed effects methods and random effects methods. We assume that α_i are random variable and the key assumption tested is here defined as:

$$H_0 : \ E\,(\alpha_i | X_{it}) = 0$$

$$H_1 : \ E\,(\alpha_i | X_{it}) \neq 0$$

This test can be interpreted as a specification test between fixed effect methods and random effect methods.

If the null is rejected, the correlation between individual effects and the explicative variables induces a bias in the GLS estimates. So, a standard approach fixed effect has to be privilegiated. If the null is not rejected, we can use a GLS estimator (random effect method) and specify the individual effects as random variables (random effects model).

According to the Hausman's lemma, we have;

$$cov\ \left(\hat{\beta}_{RE}\ ,\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)\right) = 0 \qquad \Leftrightarrow \quad cov\ \left(\widehat{\beta}_{FE}\ ,\hat{\beta}_{RE}\right) = \ var\ \left(\hat{\beta}_{RE}\right)$$

Since

$$var\ \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) = var\ \left(\hat{\beta}_{FE}\right) + var\ \left(\hat{\beta}_{RE}\right) - \ 2\mathrm{cov}\ \left(\widehat{\beta}_{FE}\ ,\hat{\beta}_{RE}\right)$$

We have

$$Var\ \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) = Var\ \left(\hat{\beta}_{FE}\right) - Var\ \left(\hat{\beta}_{RE}\right)$$

The general Hausman's specification test statistic of individual effect can be defined as follows;

$$H = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\ \right)' \left[\ Var\ \left(\hat{\beta}_{FE}\right) - Var\ \left(\hat{\beta}_{RE}\right)\ \right]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)$$

Or equivalently

$$H = \left(\hat{\beta}_{2SLS} - \hat{\beta}_{GLS}\ \right)' \left[\ Var\ \left(\hat{\beta}_{2SLS}\right) - Var\ \left(\hat{\beta}_{GLS}\right)\ \right]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{GLS}\right)$$

Under H_0 , we have:

$$H \xrightarrow[N \to \infty]{H_0} \chi^2 (K)$$

In the literature, a Hausman test (1978) has been widely used for this purpose. The statistic used for this test is a distance measure between the 2SLS and GLS estimators of β as stated above.

For the cases in which T is fixed and $N \to \infty$, the RE assumption warrants that the Hausman statistic H_{NT} is asymptotically χ^2 - distributed with degrees of freedom equal to k. This result is a direct outcome of the fact that for fixed T and strict exogenous variables, the GLS estimator $\hat{\beta}_{GLS}$ is asymptotically more efficient than the two - stage least square estimator $\hat{\beta}_{2SLS}$, and that the difference between the two estimators is asymptotically normal; specifically, as $N \to \infty$,

$$\sqrt{NT}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \Rightarrow N\left(0, \underset{N \to \infty}{\mathrm{P\,lim}} \sqrt{NT}\left[Var\left(\hat{\beta}_{FE}\right) - Var\left(\hat{\beta}_{FE}\right)\right]\right)$$

where • " ⟹ " means "converges in distribution."

An important condition that guarantees (6) is that $\theta_T > 0$, $\theta_T = \sqrt{\frac{\sigma_v^2}{T\sigma_u^2 + \sigma_v^2}}$. If $\theta_T = 0$, then the FE and GLS estimators become identical and the Hausman statistic is not defined. Observe now that $\theta_T \to 0$ as T → ∞. This observation naturally raises several issues related with the asymptotic properties of the Hausman test as T → ∞. Assume that xit contains a single time-varying regressors which is independently and identically distributed over different i and t. For this simple model, we can easily show that

$$\underset{N \to \infty}{\mathrm{P\,lim}} \sqrt{NT}\, Var\left(\hat{\beta}_{FE}\right) = \underset{N \to \infty}{\mathrm{P\,lim}} \sqrt{NT}\, Var\left(\hat{\beta}_{RE}\right)$$

,using the fact that $\theta_T \to 0$ as T → ∞. This asymptotic equality immediately implies that the 2SLS and GLS estimators of β are asymptotically equivalent; that is ,

$$\underset{N \to \infty}{\mathrm{P\,lim}} \sqrt{NT}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) = 0_{K \times 1}$$

Our equivalence result implies that between variations in data become less informative for the GLS estimation of β as T → ∞. Then, the GLS estimator of β may remain consistent even if the RE assumption is violated. If this is the case, the power of the Hausman test might be inversely related to the size of T.

What makes it complex to investigate the asymptotic properties of the within ,2SLS, GLS estimators and the Hausman statistic is that their convergence rates crucially depend on data generating processes.

3.9 Conclusion

In this chapter, we gave estimation of panel data regression model with fixe and random effects and estimation procedures of each models .We also derived and investigated some asymptotic properties of panel data model estimators. In particular, we have investigated consistency and asymptotic normality of two-stage least square and generalized least square under specified conditions. In the next chapter, we give simulation study to see the performance of the model estimators.

CHAPTER 4

SIMULATION STUDY

4.1 Introduction

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

4.2 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, and three time-variant exogenous variables; probability of conviction given arrest, probability of prison sentencing given conviction and average duration of a prison sentence. Further, we generate an instrument, Z , for endogenous regressor which is standard normal distributed.

As described in the previous chapter, 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 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.

The Simulation of the panel data is based on two different specifications; fixed effects and random effects. In the fixed effect specification, it is we first perform within-transformation to

eliminate individual effects then proceed to estimate the model. In the random effects specification individual specific effects are not estimated and then treat them as the error term.
The data-generating model is defined in the above model. The standard panel framework is determined by the following data-generating process with unobserved unit-specific effects. This DGP has the following properties: the overall constant term is set to zero and therefore vanishes from the (3.4) and the coefficient of the scalar regressor X_{it} is normalized to five. The cross-sectional means of the regressor and error terms follow a standard normal distribution (with expectation zero), implying that its variance is normalized to one.

For benchmark design, we consider three exogenous regressor and one endogenous with coefficient $\beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = 1.5,\ \beta_4 = 2$ enters 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. 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.

The four explanatory variables are denoted x_1, x_2, x_3 and x_4 . These variables are described in table 1. The variables are all normally distributed with different means and standard deviations. All Variables are 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, $\beta_1, \beta_2,\ \beta_3,$ and β_4 were sampled from a normal distribution. The mean of the constants, is assumed to be 10 and the variance of its normal distribution 2. Having determined these variables, the dependent variable, y, is calculated. The settings of the model variables of the simulation study are given in the following table.

Table 1: Description of Variables

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

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	Description	Assumed values
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 (3.4) 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 FE and RE approaches, we generate our data as typical empirical crime problem.

Since repeated measures are used, we can estimate causal relationship rather than mere correlation. To see for potential individual-specific fixed effects, we create a box plot of the response grouped by units .We now plot the considerable heterogeneity of fixed effects across individuals and years of our simulated data as it appear in panel data.

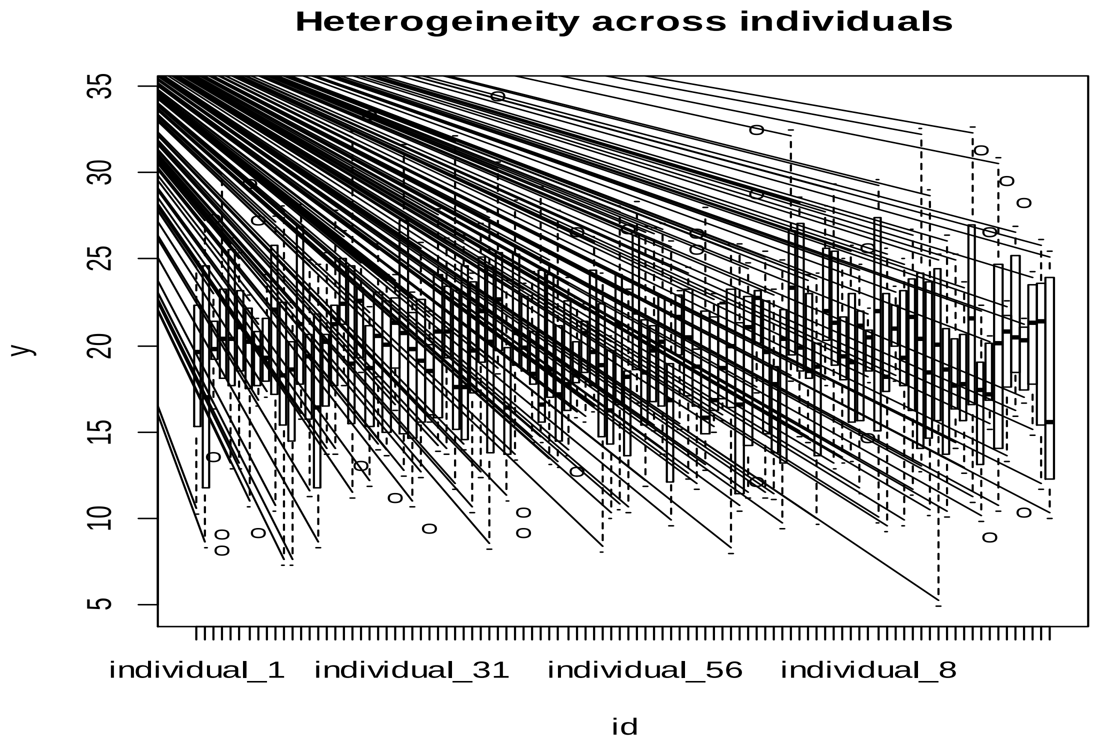

Fig 4.1: FE Heterogeneity of Simulated Row Data Across Individuals for N=100, T=10.

Figure 4.1 shows evidence of heterogeneity in the response among cross sectional units and unobservable variables does not change over time. The model with individual-specific intercepts and common slope appears to fit the simulated data well.

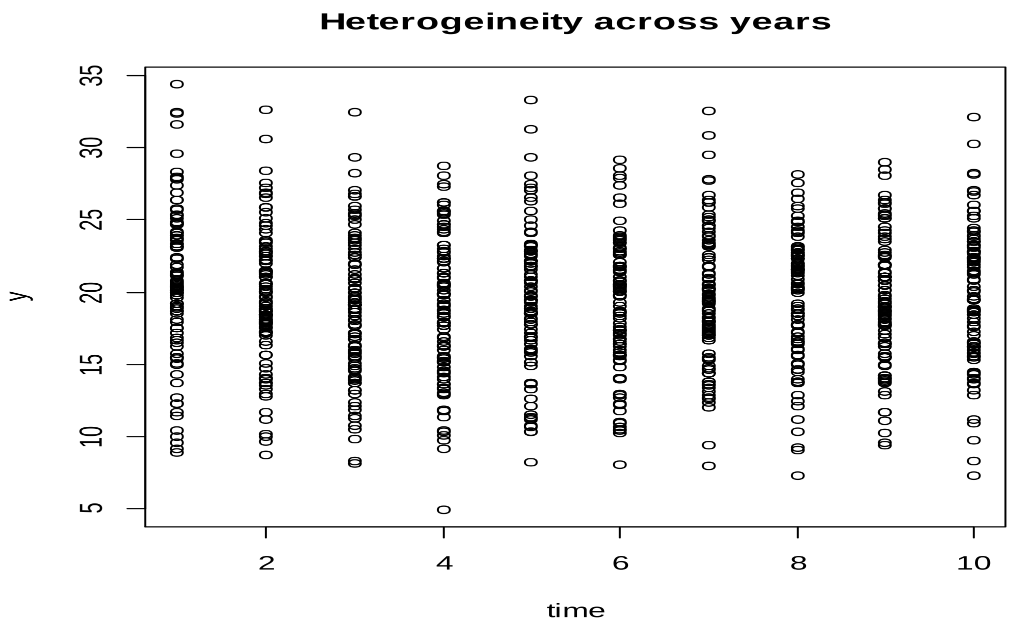

Fig.4.2: FE Heterogeneity Across Years from Simulated Data for N=100 ,T=10

Figure 4.2 suggests there is no heterogeneity in the response between years. However, there are no systematic individual-specific effects over time.

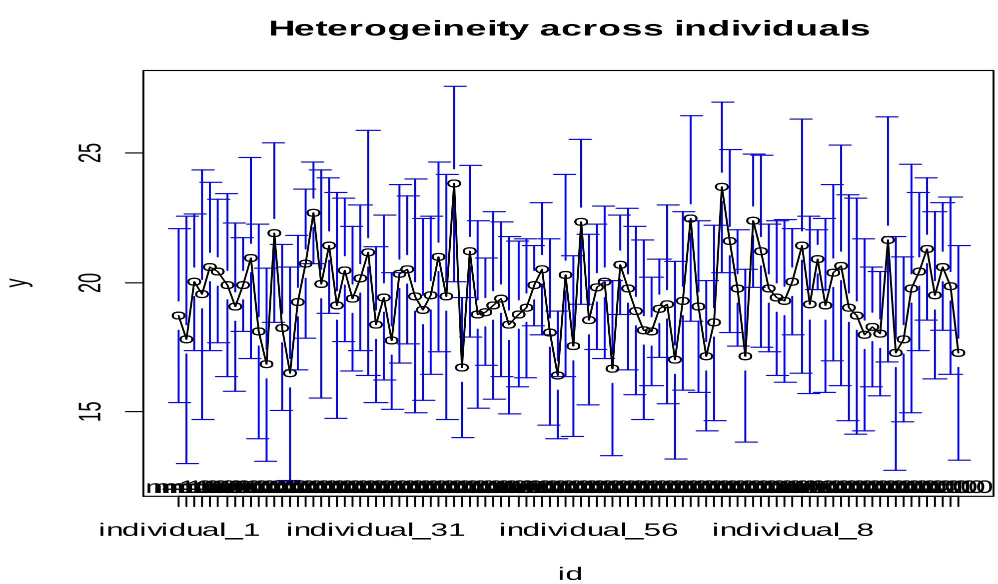

Fig 4.3: FE Heterogeneity of Simulated Data Across mean of Individuals for N=100 ,T=10

The box plot displayed in figure 4.3 represents summary statistics for the analysis of heterogeneity across individuals. Thus, we have Some evidence as there is systematic differences in the mean response among cross sectional units and unobservable variables does not change over time. As seen in the exploratory plots, there are no systematic individual-specific effects over time.

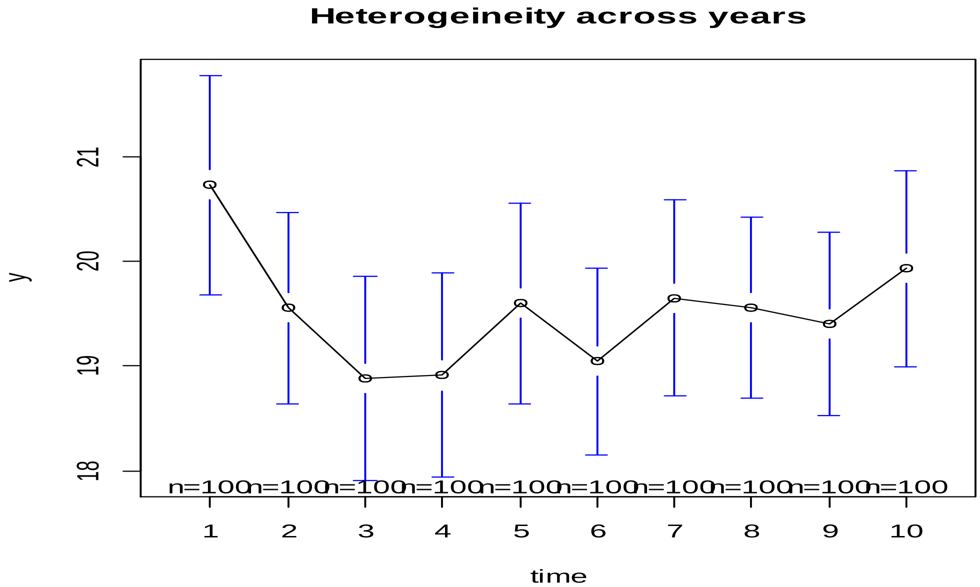

Fig 4.4 : FE Heterogeneity of Simulated Data Across of Years for N=100 ,T=10.

Figure 4.7 represents summary statistics for the analysis heterogeneity across mean of years. Thus, there does not appear to be any systematic differences in the mean response between years and unobservable variables does not change over time. As we can see in the exploratory plots, there are no systematic individual-specific effects over time.

In our simulation design , we have included one endogenous variable in the panel data regression model. Therefore, we plot endogenous varible x3 againist its instrument z for N=100,T=10 to see whether there is a relationship between them.

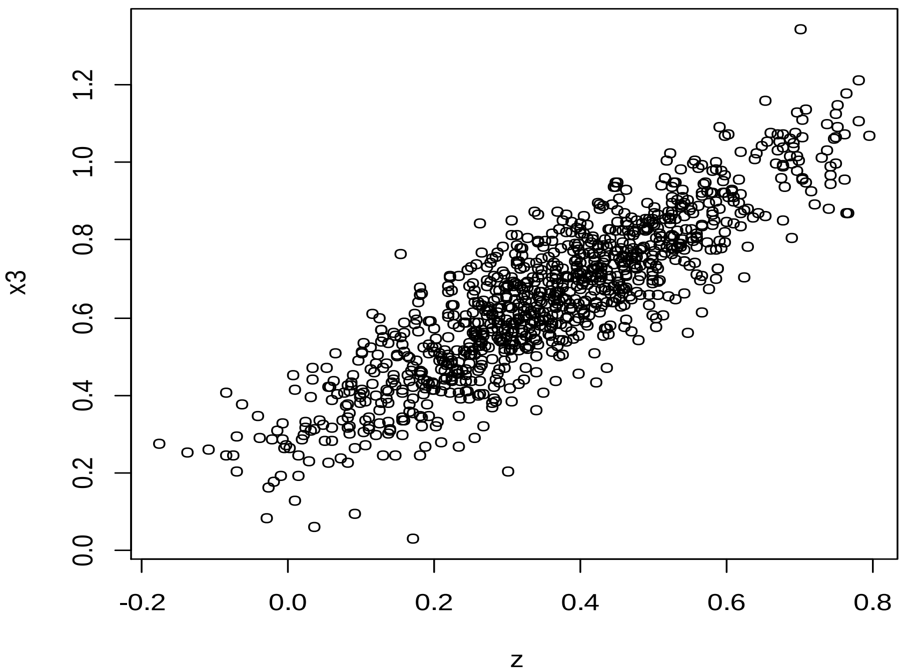

Fig 4.5 : Simulated Data that Shows Relationship Between Endogenous Variable x3 and its Instruments z for N=100,T=10.

From figure 4.5 we have evidence of there is a relationship between endogenous variable x3 and its instruments z. Note that this regression leads to that increases in the instrument z , then the endogenous variable x3 also increase. We calculate this correlation in the simulation which is 0.868.

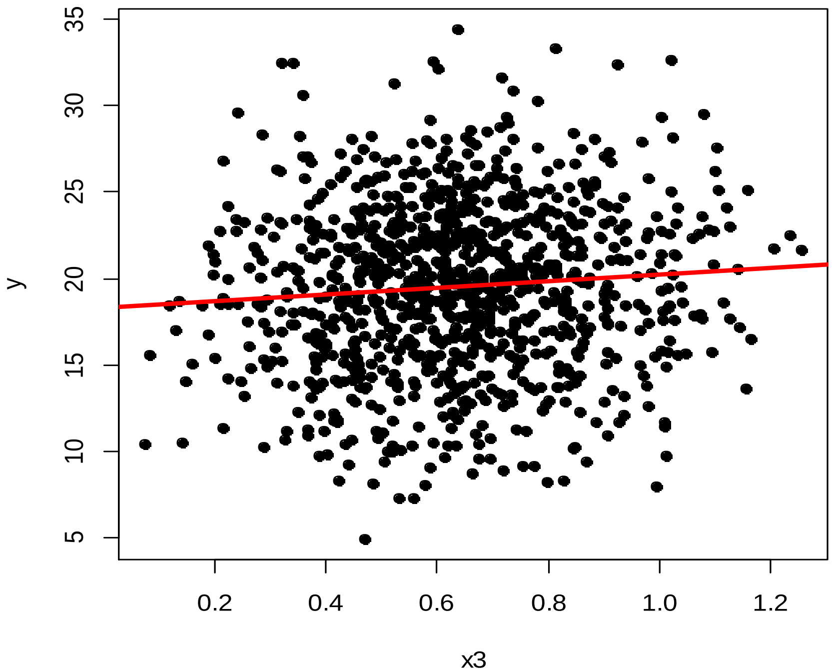

Fig 4.6: Relationship Between Response Variable and Endogenous Variable x3 for N=100,T=10.

The graphical exploration of the simulated data in figure 4.6 suggests that there is considerable heterogeneity in the individual units, which we have been modeled in chapter 3. That is, we discussed a model in which intercepts vary across entities. This model is cannot be estimated via OLS. Thus, from figure 4.6, we have some evidence that regular OLS regression does not consider heterogeneity across groups or time.

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 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. 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 tables 3 consisting 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
R square	0.96916		0.87294				0.96585	
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
R square	0.97371		0.87796				0.97396	
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
R square	0.96023		0.8663				0.96077	
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
R square	0.9600		0.86752				0.96373	
σ_ε^2							0.6617	
σ_α^2							0.01866	
θ							0.117	

The results in table 3 presents that method of estimation, mean ,standard error of estimate of β and adjusted R square. While 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 base on result 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 the table 3 indicate that in the presence of endogeneity, 2SLS estimator has lower standard error than 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 endogenous variable and while other methods are efficient. 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 small standard error. In general, based on our simulation results, pooled OLS has high standard error and 2SLS has smaller standard error compared to all other estimators.

For instance, as N= 100,T=10 2SLS estimator of $\beta(=5)$ converges to 4.2593 and GLS estimator converges to 4.0376.Thus , 2SLS is more consistent estimator than within and GLS in the presence of endogeneity problem.

And the averages mean for pooled OLS and within estimator are 3.98071 and 4.1849 respectively, while their true coefficients values for mean 5 for N=100,T=10. As we can see within fixed estimator is outperformed relative to pooled OLS. For all, 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 indicate 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, 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 is smaller than the random- effects 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 are given below.

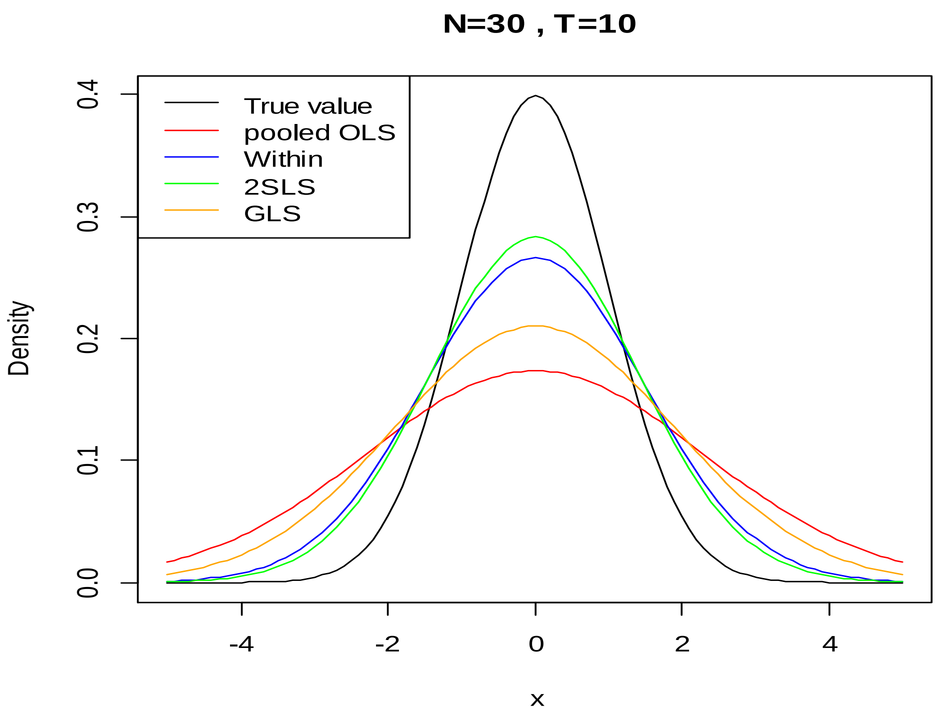

Fig 4.7 a : Distribution of Estimators Using Standard Deviation from Simulated data for N=30 ,T=10.

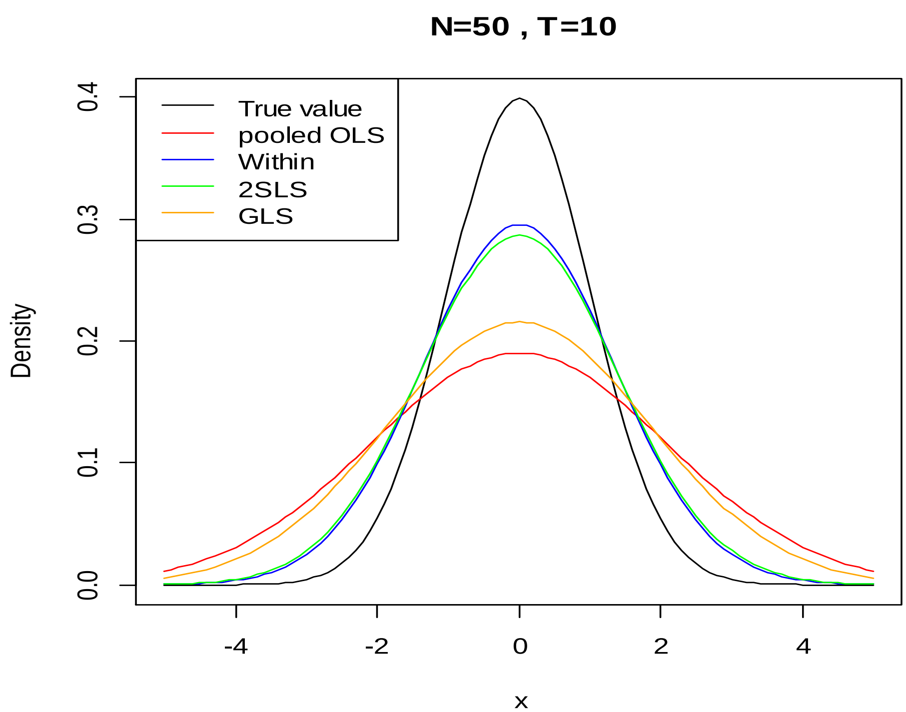

Fig 4.7 b : Distribution of Estimators using Standard Deviation from Simulated Data for N=50 ,T=10.

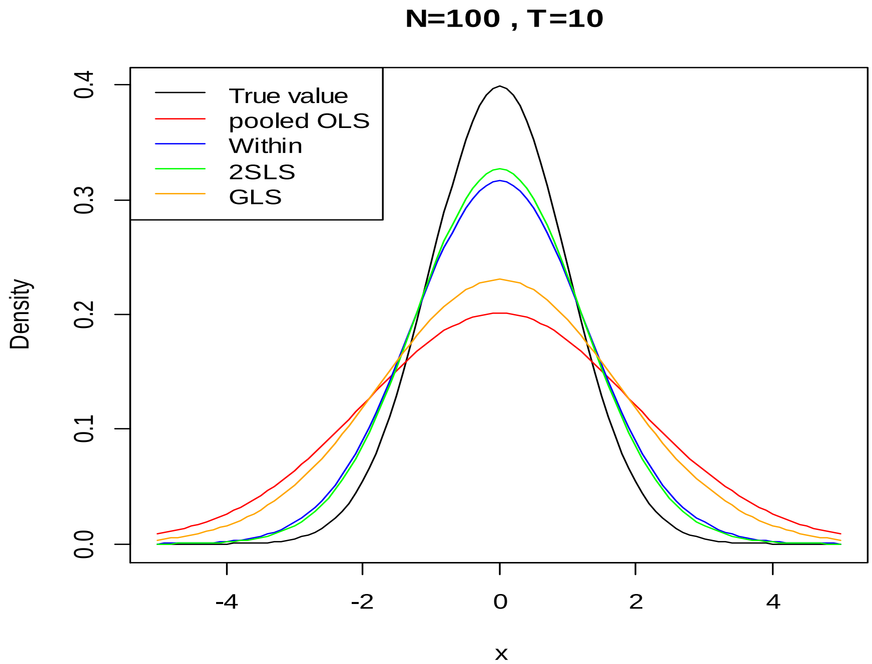

Fig 4.7 c : Distribution of Estimators Using Standard Deviation from Simulated Data for N=100 ,T=10.

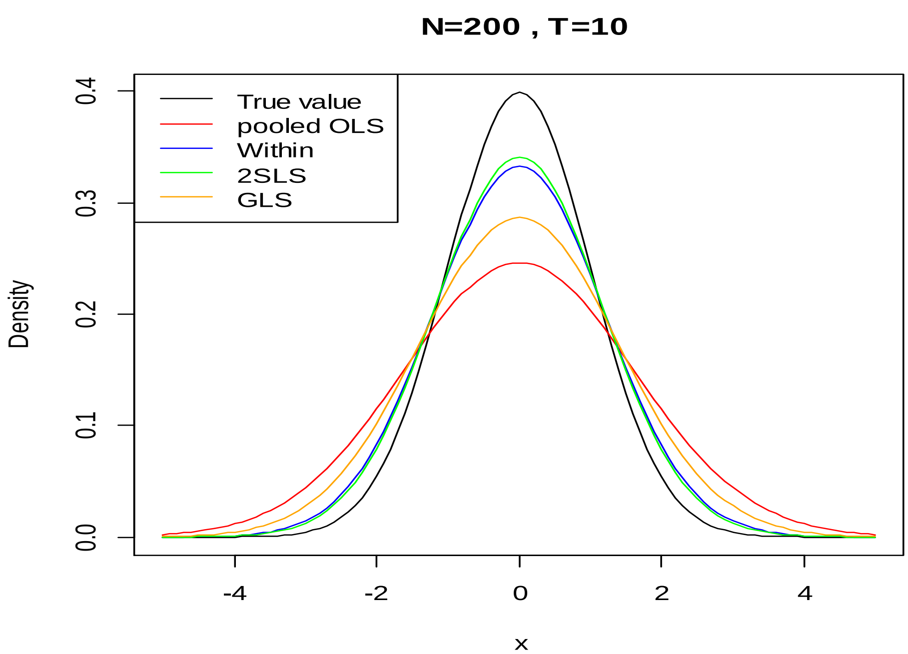

Fig. 4.7 d : Distribution of Estimators Using Standard Deviation from Simulated Data for N=200,T=10.

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, 2SLS estimator is outperformed than Within, GLS and Pooled OLS estimators in terms of mean of estimates. For N=50,T=10,Within estimator is better than the other estimators. In the figure 4.7 c and d 2SLS estimator more close to true value as N increases. In general as number of cross sections unit increase and time dimension is fixed, the panel data estimators closer and closer to true value.

Figures 4.8 show the distributions of Pooled OLS, Within, 2SLS and random effects coefficients estimated from a data generating process with four RHS variables and it's standard deviation.

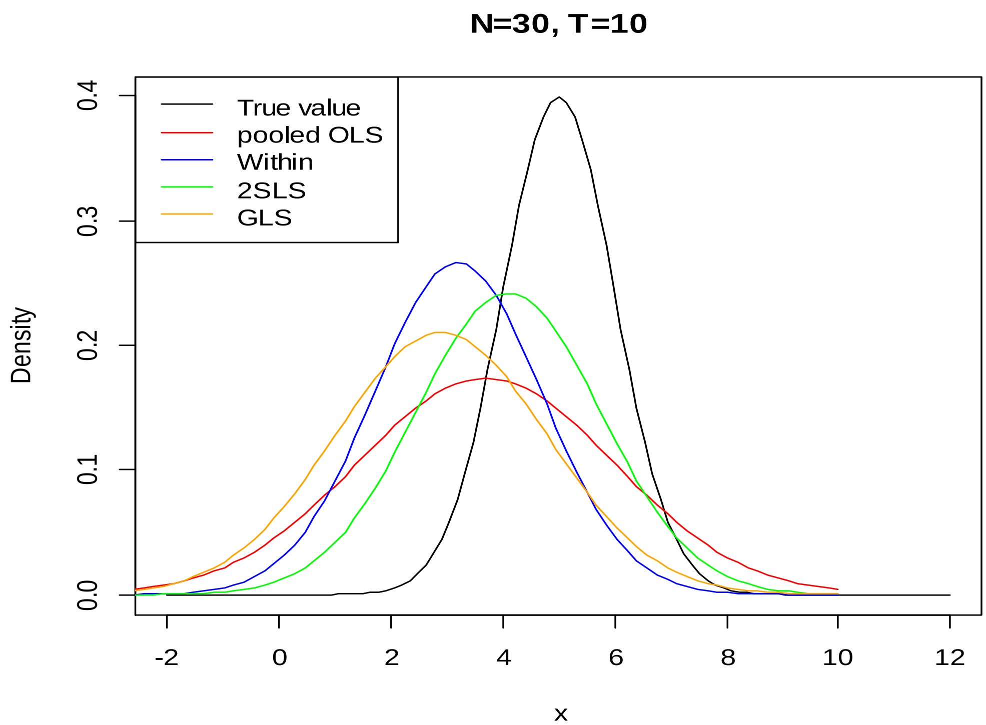

Fig 4.8 a : Distribution of Estimators using Mean and Standard Deviation from Simulated Data for N=30,T=10.

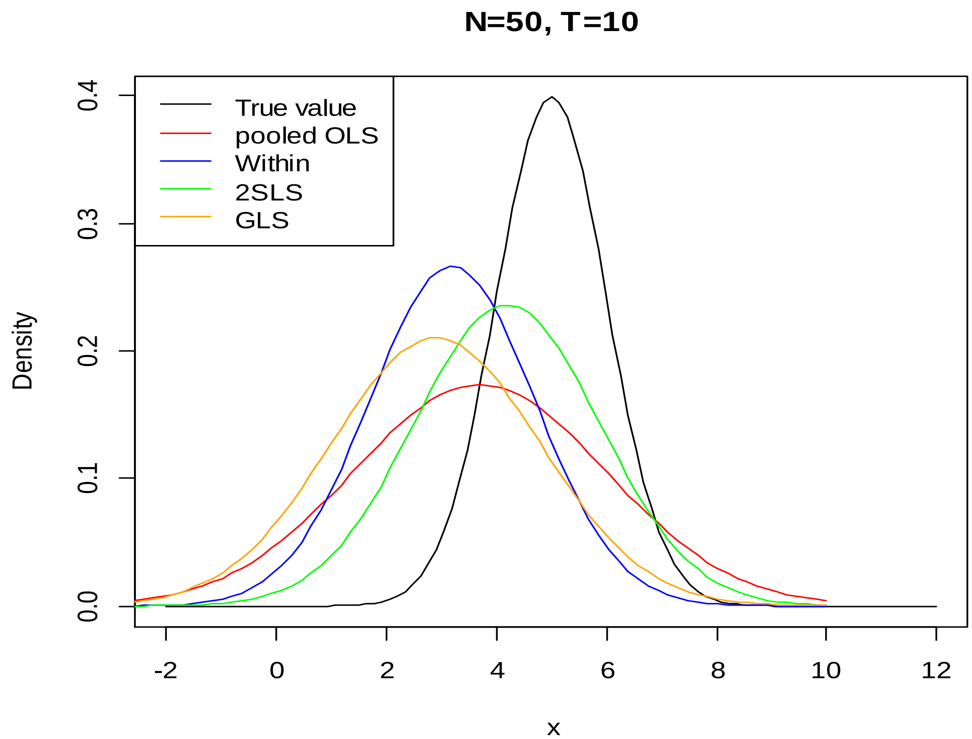

Fig 4.8 b : Distribution of Estimators using Mean and Standard Deviation from Simulated Data for N=50,T=10

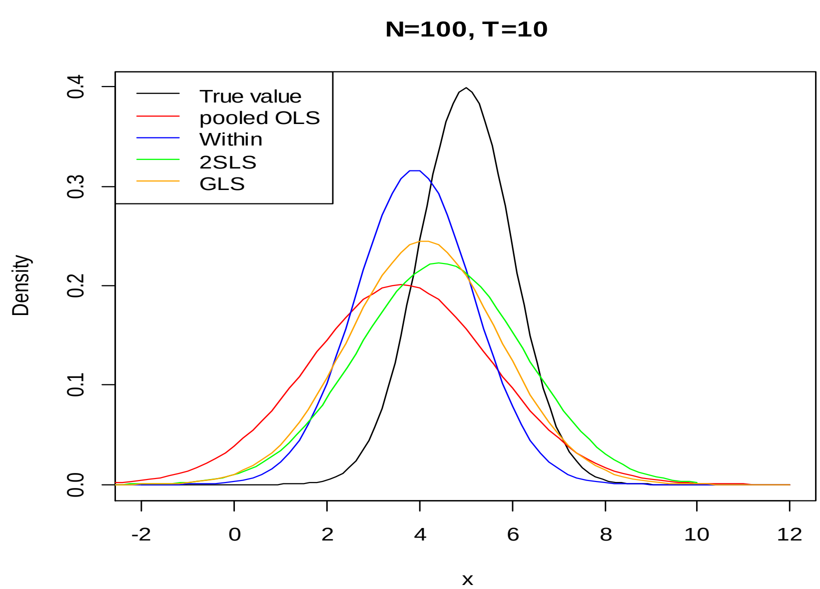

Fig 4.8 c : Distribution of Estimators using Mean and Standard Deviation from Simulated Data for N=100,T=10

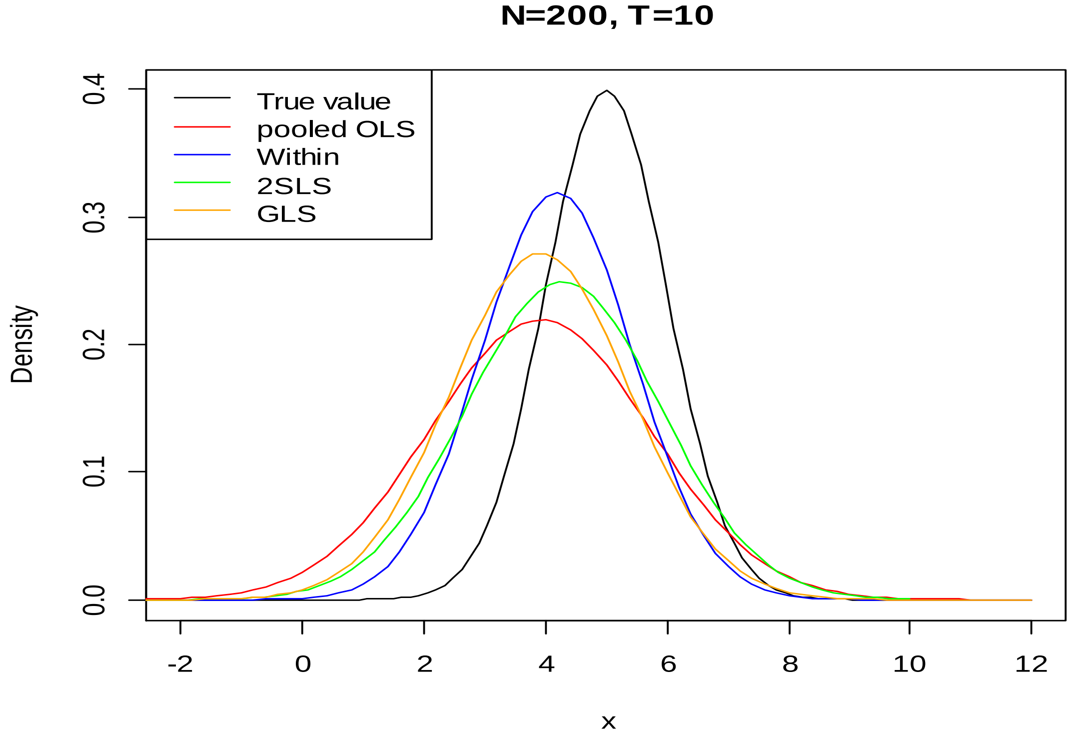

Fig 4.8 d : Distribution of Estimators using Mean and Standard Deviation from simulated data for N=200,T=10

As we can see from figure 4.8 variation across cross sectional units gets smaller as the number of individuals gets larger and 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 endogenous variable and individual effects are correlated with regressors. The pooled OLS estimator far away from the true relationship and its distribution is wider than all other estimators. Overall, our simulation results clearly show that as N increase and T is fixed , then, the standard error of Pooled OLS , within , 2SLS and GLS estimators gets decreases.

CHAPTER 5

CONCLUSION AND RECCOMMENDATION

5.1 Conclusions

Panel data, by blending the inter-individual differences and intra-individual dynamics have advantages over cross-sectional or time series data. It 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 arises 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 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 runs to compute the mean and standard errors of the estimator within, 2SLS , OLS and GLS. 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, 2SLS performs better relative to within estimator with large cross-sections. The standard error of GLS are smaller than OLS, which is consistent with the theoretical result under exogeneity between individual effects and regressors. The pooled OLS estimator far away from the true relationship and its distribution is wider than all other estimators in large sample size. Overall, our simulation results show that the estimated standard error of estimators gets 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.

5.2 Recommendations

There are a number of possible extensions for estimation of panel data model;

1. Previous studies have often assumed that data are cross-sectionally independently and identically distributed. Our findings suggest that future studies should pay more attention to cross-sectional heterogeneity.
2. There are also a number of possible extensions to fixed and random-effects models that can be entertained through the use of a covariance structure approach. While we did not illustrate these extensions , they have potentially important theoretical applications. Therefore , a much wider range of interesting questions can be addressed.
3. For an intermediate model between fixed effects and random effects, these studies propose several instrumental variables estimators by which both the coefficients on time-varying and time invariant regressors can be consistently estimated. It would be interesting to investigate the large N and large T properties of these instrumental variables estimators and the Hausman tests based on these estimators asymptotically.
4. Two widely-used methods are the use of either fixed or random effects models. However How best to choose between fixed and random effects remains unclear in the applied literature. Therefore, researchers and analysts should also pay attention to how one can choose random and fixed model to come up with innovative method like likelihood ratio test , RMSE of estimates , etc and compare with Hausman test.
5. 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. Therefore, there are many important issues such as modeling of joint dependence, simultaneous equations models, the random intercept model ,varying parameter models (e.g., Hsiao 1992, 2003; Hsiao and Pesaran 2006), unbalanced panel, measurement errors (e.g., Griliches and Hausman 1986;Wansbeek and Koning 1989), 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.

6. Finally, asymptotic as $(N, T \to \infty)$ are much more sensitive to data generating processes than asymptotic as either $N \to \infty$ or $T \to \infty$ are. Future studies can avoid making any particular restriction on the relative sizes of N and T and then theoretical results apply to a broader range of panel data.

Publication and Paper sent for publication under this thesis

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

Megersa T.J, Chelule J.C and Odhiambo R.O (2014), On Estimation of Panel Data Regression Model In The Presence of Heterogeneity and Endogeneity. CESER PUBLICATION, International Journal of Mathematics and Computations Vol.26, 2014 (Published).

Megersa T.J, Chelule J.C and Odhiambo R.O (2014), Investigating the Asymptotic Properties of Some Panel Data Regression Model Estimators with Individual Effects. Science Publishing Group , American Journal of Theoretical and Applied Statistics,Vol.3,2014 (Accepted).

Megersa T.J, Chelule J.C and Odhiambo R.O (2014), Estimation of Panel Data Regression Models with Individual Effects. Journal of Statistical Theory and Application ,Vol.13, 2014 (Under review).

References

Ahn and Moon (2001), Large-N and Large-T Properties of Panel Data Estimators and the Hausman Test. *USC CLEO Research Paper No. C01-20*.

Ahn and Schmidth (1995), Efficient estimation of models for dynamic panel data , Journal of Economics 68, 5-27.

Ahn, S.C. and S. Low, (1996), A reformulation of the Hausman test for regression models with pooled cross-section time-series data, Journal of Econometrics 71, 309–319.

Arellano M (2003), Panel data econometrics. Oxford University Press, Oxford.

Arellano, M. and O. Bover, (1995), Another look at the instrumental variables estimation of error-component models, Journal of Econometrics 68, 29-51.

Arellano, M.,(1993), On the testing of correlated effects with panel data, Journal of Econometrics 59, 87-97.

Baillie, R. and B.H. Baltagi,(1995), Prediction from the regression model with one-way error components,working paper, Department of Economics, Texas A&M University, College station, Texas.

Baltagi, B. and S. Khanti-Akom, (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.

Baltagi,B. H. and J. M.Griffin(1988), A generalized error component model with heteroscedastic disturbances, International Economic Review 29, 745-753.

Baltagi, B. H., B. C. Jung and S. H. Song (2008),Testing for heteroscedasticity and serial correlation in a random effects panel data model. Working paper No. 111, Syracuse University, USA.

Baltagi, B.H. and Q. Li, (1992), Prediction in the one-way error component model with serial correlation, Journal of Forecasting 11, 561-567.

Baltagi, B.H., (1981a), Pooling: an experimental study of alternative testing and estimation procedures in a two-way error components model, Journal of Econometrics 17, 21-49.

Baltagi, B.H., (1995b), Econometric analysis of panel data (Chichester: Wiley).

Baltagi, B.H., (2001), Econometric Analysis of Panel Data. Wiley, Chichester.

Bates D (2007), lme4: Linear Mixed E_ects Models Using S4 Classes. R package version 0.99875-9, URL http://CRAN.R-project.org.

Bhargava, A.,(1991),Identification and panel data models with endogenous regressors, Review of Economic Studies 58, 129-140.

Bollen KA, Brand JE (2008), Fixed and Random Effects in Panel Data Using Structural Equations Models.

Bresson, G.,C. Hsiao and A.Pirotte (2006), Heteroskedasticity and random coefficient model on panel data, Working Papers ERMES 0601, ERMES, University Paris 2.

Cameron, A. and Trivedi, P. (2005) Microeconometrics. Methods and Applications. Cambridge University Press,

Cameron, A. Colin, and Pravin K. Trivedi (2009), Microeconometrics Using Stata. TX: Stata Press.

Cardellichio, P.A., (1990), Estimation of production behavior using pooled microdata, Review of Economics and Statistics 72, 11–18.

Choi, I., (1998), Asymptotic analysis of a nonstationary error component model, mimeo, Kookmin University, Korea.

Choi, I.,(2002), Instrumental variables estimation of a nearly nonstationary, heterogeneous error component model, *Journal of Econometrics* 109, 1–32.

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

Croissant Y, Millo G (2008). \Panel Data Econometrics in R : The plm Package." Journal of Statistical Software , 27 (2). URL http://www.jstatsoft.org/v27/i02/.

Croissant and Millo (2011) , Panel Data Econometrics in R: The plm Package working paper University of Trieste and Generali SpA.

Croissant Y, Millo G (2008), Panel Data Econometrics in R: The plm Package." Journal of Statistical Software, 27(2). URL http://www.jstatsoft.org/v27/i02/.

Elhorst JP (2010), Dynamic Panels with Endogenous Interactions Effects when T Is Small Regional Science and Urban Economics, 40, 272-282.

Elhorst JP, Piras G and Arbia G (2010), Growth and Convergence in a Multi-Regional Model with Space-Time Dynamics. Geographical Analysis, 42, 338-355.

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

Greene (2008), Econometric analysis 6th ed., Upper Saddle River, N.J. : Prentice Hall

Greene, W.H. (2002), The Behavior of the Fixed Effects Estimator in Nonlinear Models, unpublished manuscript, New York University.

Hahn, J., and W. Newey (2004), Jackknife and Analytical Bias Reduction for Nonlinear Panel Models, Econometrica 72, 1295-1319.

Halaby, C. (2004), Panel Models in Sociological Research. Annual Rev. of Sociology 30: 507-544.

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

Hsiao, C.(2003), Analysis of panel data , 2nd edition. Econometric society monographs, Cambridge University Press, Cambridge.

Hsiao C (2005) , Why panel data? Singap Econ Rev 50(2):1–12

Kao and Chiang, (2000), On The Estimation And Inference Of A Cointegrated Regression In Panel Data. Elsevier Science Inc. Volume 15, 179–222.

Kao, C., & Chiang, M. (1997), On the Estimation and Inference of a Cointegrated Regression In Panel Data. Working paper, Department of Economics, Syracuse University. Simulation

Kao, C., (1999), Spurious regression and residual-based tests for cointegration in panel data, Journal of Econometrics, 90, 1-44.

Katz, E. (2001), Bias in Conditional and Unconditional Fixed Effects Logit Estimation," Political Analysis 9(4), 379-384.

Kyriazidou E. (2001). Estimation of dynamic panel data sample selection models. Review of Economic Studies 68: 543-572.

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

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

Lancaster (2004), An Introduction to Modern Bayesian Econometrics. Blackwell Publishing Ltd.

Lee LF, Yu J (2010a), A Spatial Dynamic Panel Data Model with both Time and Individual Fixed Effects. Econometric Theory, 26, 564-597.

Lee LF, Yu J (2010b), A United Transformation Approach to the Estimation of Spatial Dynamic Panel Data Models: Stability, Spatial Cointegration and Explosive Roots. In A Ullah, DEA Giles (eds.), Handbook of Empirical Economics and Finance, pp. 397-434. Chapman & Hall/CRC.

Lee LF, Yu J (2010c), Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects. Journal of Econometrics, 154, 165-185.

Lee LF, Yu J (2010d), Some Recent Development in Spatial Panel Data Models." Regional Science and Urban Economics, 40, 255-271.

Li, Q. and T. Stengos (1994), Adaptive estimation in the panel data error component model with heteroscedasticity of unknown form, International Economic Review 35, 981-1000.

Maddala, G.S. (2008), Introduction to econometrics, 3rd edition, John Wiley & Sons, Ltd, Chichester, UK.

Maddala, G.S. and T.D. Mount, (1973), A comparative study of alternative estimators for variance components models used in econometric applications, Journal of the American Statistical Association 68, 324-328.

Magnus, J.R. (1982), Multivariate error components analysis of linear and non-linear regression models by maximum likelihood, Journal of econometrics 19, 239-285.

Mazodier, P. and A. Trognon (1978), Heteroskedasticity and stratification in error components models, Annales de l'INSEE 30-31, 451-482.

Metcalf , G.E (1996) , Specification Testing in Panel data with Instrumental variables , Journal of Econometrics 71,291-307.

Moulton, B.R., (1986), Random group effects and the precision of regression estimates, Journal of Econometrics 32, 385–397.

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

Mutl J (2006), Dynamic Panel Data Models with Spatially Autocorrelated Disturbances. Ph.D. thesis, University of Maryland, College Park.

Mutl J, Pfa_ermayr M (2011), The Hausman Test in a Clif and Ord Panel Model. Econo- metrics Journal, 14, 48-76.

Neyman and Scott, (1948) , consistent Estimates based on Partially consistent observations , Econometric 16, 1-32.

Ogunwale, et al (2011), On The Comparison of Two Methods of Analyzing Panel Data Using Simulated Data, University of Ado-Ekiti, Nigeria.

Owusu-Gyapong, A.,(1986), Alternative estimating techniques for panel data on strike activity, Review of Economics and Statistics 68, 526–531.

Pesaran HM, Tosetti E (2011), Large Panels with Common Factors and Spatial Correlations. Journal of Econometrics, 161(2), 182 -202.

Phillips, P.C.B., and H. Moon, (1999), Linear regression limit theory for nonstationary panel data, Econometrica, 67, 1057-1111.

Pinheiro J, Bates D, DebRoy S, the~R Core~team DS (2007), nlme: Linear and Nonlinear Mixed E_ects Models. R package version 3.1-86, URL http://CRAN.R-project.org.

R Development Core Team (2008), R: A Language and Environment for Statistical Computing.

R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

R. Frisch and F.V. Waugh (1989), Partial time regressions as compared with individual trends. Econometrica, 1:387–401.

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

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

Roy, N. (2002), Is adaptive estimation useful for panel models with heteroscedasticity in the individual specific error component? Some Monte Carlo evidence, Econometric Reviews 21, 189-203.

Schmidt (2005), Econometrics, McGraw-Hill/Irwin, New York.

Semykina A, Wooldridge JM. (2010), Estimating panel data models in the presence of endogeneity and selection. Journal of Econometrics 157: 375-380.

Taub, A.J., (1979), Prediction in the context of the variance-components model, Journal of Econometrics 10, 103-108.

Taylor, W. E. (1980). Small sample considerations in estimation from panel data, Journal of Econometrics, 13, 203–223.

Wallace, T.D. and A. Hussain, (1969), The use of error components models in combining crosssection and time-series data, Econometrica 37, 55-72.

Wansbeek, T.J. (1989), An alternative heteroscedastic error component model, Econometric Theory 5, 326.

Wansbeek, T.J. and A. Kapteyn, (1978), The separation of individual variation and systematic change in the analysis of panel data, Annales de l'INSEE 30-31, 659-680.

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

Wooldridge, J. M. (2012), Introductory Econometrics: A Modern Approach, 5th edition, South-Western College.

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

Yaffee, R. (2003), Primer for Panel Data Analysis. Connect: Information Technology at NYU.Available online at:http://www.nyu.edu/its/pubs/connect/fall03/yaffee_primer.html.

Appendix

```
#Revised R code for simulation of panel data regression model and estimators
For further enquiry contact magetade2003@gmail.com
library(mvtnorm)
N <- 200
T<- 10
NT<- N*T
nSims <- 1000
b1=0.5
b2=0.75
b3=1
b4=1.5
sigmaAlpha <- 0.7
sigma <- 0.2
rXm <- c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1 )
alphai=rnorm(N,10,2)  #alphai simulation
id <- rep(1:N,T)
f <- rep(rnorm(N),T)  # is individual specific
sig <- diag(rep(1,T))
z <- rnorm (N*T , 0.35,0.18)
u <-rnorm(N*T,0.29,0.1)
x1 <- rnorm(N*T, 0.37,0.3)
x2 <- rnorm(N*T,0.65,0.05)
x3 <- z + u
x4 <- rnorm(N*T,12,3)
X = cbind(x1,x2,x3,x4)
y <- b1*x1 + b2*x2 + b3* x3 + b4* x4+ alphai + u
pdata <- data.frame(id = rep(paste("individual",1:N, sep="_"),each =T),time = rep(1:T,N),alphai
= rnorm(N), y, x1,x2,x3,x4,z,u )
library("plm")
panel <- plm.data(pdata, index = c("id","time"))
```

```
#panel <- pdata.frame(pdata,c("id","time"))
 eq <- y ~ x1 + x2 + x3 + x4
# Pooled OLS estimator
pooling  <- plm(eq  , model = "pooling",  data=panel)
summary(pooling  )
# Between estimator
between <- plm(eq  , model = "between",  data=panel)
summary(between)
# Fixed effects  or within  estimator
fixedeffects  <- plm(eq  , model = "within",  data=panel,effects="individual")
summary(fixedeffects)
# Random effects  estimator
random <- plm(eq,  model = "random",  data=panel,effects  = "individual")
summary(random)
# Pooled OLS estimator
pooling  <- plm(y  ~ x3, model  = "pooling",  data=panel)
summary(pooling  )
# Between estimator
between <- plm(y  ~ x3 , model = "between",  data=panel)
summary(between)
# Fixed effects  or within  estimator
fixedeffectx3  <- plm(y  ~ x3, model = "within",  data=panel,   effect="individual")
summary(fixedeffectx3)
library("sem")
# Two-stage  least square estimator
twostageleastsquare  <- tsls(y ~ x3, instruments  = ~ z,data=panel)
summary(twostageleastsquare)
fit2sls  <- plm(y  ~ x3,data=panel,method="2SLS",instrument=~z)
summary(fit2sls)
twostageleastsquare3  <- tsls(y ~ x1 + x2 + x4, instruments  = ~ z,,data=panel)
summary(twostageleastsquare3)
```

```
library("AER")
 iv <- ivreg(y  ~ x3 | z, data=panel)
summary(iv)
# Random effects estimator
random <- plm(eq, model = "random", data=panel,effects = "individual")
summary(random)
# Hausman test for fixed versus random effects model
phtest(random, fixed)
phtest(fe, re)
phtest(re,iv)
phtest(re,twostageleastsquare)
plot(density(fixef(fixedeffects)))
plot(x3,z)
plot(y ~ id, main="Heterogeineity across individuals", data=panel)
plot(y ~ time, main="Heterogeineity across years", data=pdata)
plot(y ~ x3,main="Heterogeineity across individuals", data=pdata)
library(gplots)
plotmeans(y ~ id, main="Heterogeineity across individuals", data=pdata)
plotmeans(y ~ time, main="Heterogeineity across years", data=pdata)
plot(pdata$x3, pdata$y, pch=19, xlab="x3", ylab="y")
abline(lm(pdata$y~pdata$x3),lwd=3, col="red")
```

good!

```
#Simulation for comparision of estimators using their standard deviation
plot(function(x) {dnorm(x)}, -5, 5, ylab="Density")
plot(function(x) {dnorm(x, sd=2.3)}, -5, 5, col="red", add=TRUE)
plot(function(x) {dnorm(x, sd=1.5)}, -5, 5, col="blue", add=TRUE)
plot(function(x) {dnorm(x, sd=1.41)}, -5, 5, col="green", add=TRUE)
plot(function(x) {dnorm(x, sd=1.9)}, -5, 5, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"), lty =
c(1, 1, 1), col = c("black", "red", "blue","green","orange"))
```

```
title("  N=30 , T=10")
plot(function(x)  {dnorm(x)},  -5, 5, ylab="Density")
plot(function(x)  {dnorm(x,  sd=2.1)}, -5, 5, col="red",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.35)}, -5, 5, col="blue",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.39)}, -5, 5, col="green",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.85)}, -5, 5, col="orange",  add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),   lty =
c(1, 1, 1), col = c("black",  "red",  "blue","green","orange"))
title("  N=50 , T=10")
plot(function(x)  {dnorm(x)},  -5, 5, ylab="Density")
plot(function(x)  {dnorm(x,  sd=1.98)}, -5, 5, col="red",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.26)}, -5, 5, col="blue",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.22)}, -5, 5, col="green",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.73)}, -5, 5, col="orange",  add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),   lty =
c(1, 1, 1), col = c("black",  "red",  "blue","green","orange"))
title("  N=100 , T=10")
plot(function(x)  {dnorm(x)},  -5, 5, ylab="Density")
plot(function(x)  {dnorm(x,  sd=1.62)}, -5, 5, col="red",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.20)}, -5, 5, col="blue",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.17)}, -5, 5, col="green",  add=TRUE)
plot(function(x)  {dnorm(x,  sd=1.39)}, -5, 5, col="orange",  add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),   lty =
c(1, 1, 1), col = c("black",  "red",  "blue","green","orange"))
title("  N=200 , T=10")
# simulation  for comparision  of estimators  using  betas and sigmas
plot(function(x)  {dnorm(x,mean=5,sd=1)},  -2, 12, ylab="Density")
plot(function(x)  {dnorm(x,  mean=3.7,sd=2.3)}  ,  -8, 10, col="red",  add=TRUE)
plot(function(x)  {dnorm(x,  mean=3.2,sd=1.5)}  , -8, 10, col="blue",  add=TRUE)
plot(function(x)  {dnorm(x,  mean=4.1,  sd=1.65)}, -8, 10, col="green",  add=TRUE)
plot(function(x)  {dnorm(x,  mean=2.9,sd=1.9)}  , -8, 10, col="orange",  add=TRUE)
```

```
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),    lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=30, T=10")

plot(function(x) {dnorm(x,mean=5,sd=1)}, -2, 12, ylab="Density")
plot(function(x) {dnorm(x, mean=3.7,sd=2.3)} , -8, 10, col="red", add=TRUE)
plot(function(x) {dnorm(x, mean=3.2,sd=1.5)} , -8, 10, col="blue", add=TRUE)
plot(function(x) {dnorm(x, mean=4.193, sd=1.69)}, -8, 10, col="green", add=TRUE)
plot(function(x) {dnorm(x, mean=2.9,sd=1.9)} , -8, 10, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),    lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=50, T=10")
plot(function(x) {dnorm(x,mean=5,sd=1)}, -2, 12, ylab="Density")
plot(function(x) {dnorm(x, mean=4.36,sd=2.1)} , -2, 12, col="red", add=TRUE)
plot(function(x) {dnorm(x, mean=3.5,sd=1.35)} , -8, 12, col="blue", add=TRUE)
plot(function(x) {dnorm(x, mean=4.41, sd=1.72)}, -8, 10, col="green", add=TRUE)
plot(function(x) {dnorm(x, mean=3.7,sd=1.85)} , -8, 12, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),    lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=50, T=10")
plot(function(x) {dnorm(x,mean=5,sd=1)}, -2, 12, ylab="Density")
plot(function(x) {dnorm(x, mean=3.6,sd=1.98)} , -8, 12, col="red", add=TRUE)
plot(function(x) {dnorm(x, mean=3.9,sd=1.26)} , -8, 12, col="blue", add=TRUE)
plot(function(x) {dnorm(x, mean=4.458, sd=1.79)}, -8, 10, col="green", add=TRUE)
plot(function(x) {dnorm(x, mean=4.1,sd=1.63)} , -8, 12, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"),    lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=100, T=10")
plot(function(x) {dnorm(x,mean=5,sd=1)}, -2, 12, ylab="Density")
plot(function(x) {dnorm(x, mean=4.9,sd=1.62)} , -8, 12, col="red", add=TRUE)
plot(function(x) {dnorm(x, mean=4.1,sd=1.23)} , -8, 12, col="blue", add=TRUE)
```

```
plot(function(x) {dnorm(x, mean=4.497, sd=1.58)}, -8, 10, col="green", add=TRUE)
plot(function(x) {dnorm(x, mean=4.39,sd=1.37)} , -8, 12, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"), lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=200, T=10")
plot(function(x) {dnorm(x,mean=5,sd=1)}, -2, 12, ylab="Density")
plot(function(x) {dnorm(x, mean=3.916,sd=1.82)} , -8, 12, col="red", add=TRUE)
plot(function(x) {dnorm(x, mean=4.19,sd=1.25)} , -8, 12, col="blue", add=TRUE)
plot(function(x) {dnorm(x, mean=4.297, sd=1.602)}, -8, 10, col="green", add=TRUE)
plot(function(x) {dnorm(x, mean=3.92,sd=1.47)} , -8, 12, col="orange", add=TRUE)
legend(x = "topleft", legend = c("True value", "pooled OLS", "Within","2SLS","GLS"), lty =
c(1, 1, 1),col = c("black", "red", "blue","green","orange"))
title("N=200, T=10")
```

The end !