



GLOBAL JOURNAL OF SCIENCE FRONTIER RESEARCH
Volume 11 Issue 7 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN : 2249-4626 & Print ISSN:0975-5896

On The Comparison of Two Methods of Analyzing Panel Data Using Simulated Data

By O.D. Ogunwale, A.O. Olubiyi, A.H. Bello.

University of Ado-Ekiti, Nigeria

Abstract - The study focus 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 least square dummy variable (LSDV) model were discussed. The samples of size 60 and 100 were generated and replicated five and ten periods respectively. 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 (LSDV) is superior and better in the analysis of panel data. The study focus 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 least square dummy variable (LSDV) model were discussed. The samples of size 60 and 100 were generated and replicated five and ten periods respectively. 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 (LSDV) is superior and better in the analysis of panel data.

Keywords : R^2 – Coefficient of determination, LSDV- least square dummy variables, Panel Data.

GJSFR-F Classification : FOR Code: 010299



Strictly as per the compliance and regulations of:



On The Comparison of Two Methods of Analyzing Panel Data Using Simulated Data

O.D. Ogunwale^a, A.O. Olubiyi^Q, A.H. Bello.^B

Abstract - The study focus 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 least square dummy variable (LSDV) model were discussed. The samples of size 60 and 100 were generated and replicated five and ten periods respectively. 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 (LSDV) is superior and better in the analysis of panel data.

Keywords : R^2 - Coefficient of determination, LSDV - least square dummy variables, Panel Data.

I. INTRODUCTION

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data. Panel data aggregate the entire individual and analyses them in a period of time. Availability of data on a large number of individuals and on each individual over a very short period of time is becoming increasingly important feature of econometric research.

Very often a researcher would like to use such data to study behavioral relationships that are dynamic in character. Since only a few observations are available overtime, but a great many observations are available for different individual at a point in time. It is exceptionally important to make efficient use of the data across individuals to estimate that part of behavioral relationship containing variables that differ substantially from one individual to another in order that the smaller amount of information overtime can be used to the best advantage in estimation of the relationship studied. As it turns out, the problem is far from simple, or the introduction of dummy variables for individuals produce estimates having serious sample bias.

The term 'panel data' refers to data set where there are observations on the same individual over several periods of time. The main advantage of panel

data as compared to a single cross section or series of cross section with non-overlapping cross-section of units is that:

- They facilitate sting and relaxation of the assumptions that are implicit in cross-sectional analysis.
- They are more informative (more variability, less co-linearity, more degree of freedom) estimates are more efficient.
- They allow studying individual dynamics (e.g separating age and cohort effects)
- They give information on time – ordering of events.
- They allow controlling for individual unobserved heterogeneity and this is a problem of non-experimental research.

Data set that combine time series and cross sections are common in economics. For example the published statistics OECD (Organization Economics Cooperation Development) contain numerous series of economics aggregates observed yearly for many countries. Recently constructed longitudinal data sets contain observation of thousands of individual or families, each observed at several point in time. Other empirical studies have analyzed time-series data on sets of firms, states, countries or industries simultaneously. These data sets provide rich sources of information about the economy. Modeling in this setting however, calls for some complex stochastic specifications.

Basically there are two methods of analyzing panel data which are; the fixed effect and random effect models. One can only use either of the methods.

a) Fixed Versus Random Effects

Attention is now focusing on how to decide on whether to use fixed or random effects model. Mundlak and Hausman (1978) suggest an interpretation of the model which leads to a solution to this problem. He suggested that in both models we should view the effect μ as random. However in the fixed effect model estimation is done conditional on the realized μ in the sample. Estimation of the random effect models is on the other hand based on the assumption that the effect μ are uncorrelated with the regressors X. when this assumption is valid, and then the random effects model uses more information, which makes it a more efficient estimator. However if the assumption of no correlation between μ and X is violated, then the random effect model leads to inconsistent estimate, where as the fixed effect model is still consistent. Thus if there is

^aAuthor : Department of mathematical Sciences, University of Ado-Ekiti, Nigeria.

^BAuthor : Department of Maths/Statistics, Osun State Polytechnic, Iree.

uncertainty about whether the μ effect may be correlated with regressors then the fixed effect model may be a safer choice.

b) Aims and Objectives

The aim and objectives is to compare which of the estimators is better in analyzing panel data. By considering;

- Parameters estimate of the pooled observations using ordinary least squares
- Estimate the individual effect parameters by using least squares dummy variables: fixed effect model
- To compare the results of least square dummy variable and the pooled data estimates so as to determine which is better.

II. RESEARCH METHODOLOGY

a) Data Collection

Data sets play a very significant role in a decision-making situation as powerful ingredients in the choice of appropriate decision support system. This study uses data simulated through the process of Ms-Excel in office 2003. Simulated data are produced by the analyst with a random number generator, usually on the assumed statistical properties.

b) Data Generation Procedure

This research uses simulated data by the use of some computer software. In generating the data we make use of Microsoft Excel 2003, while both SPSS and MS-Excel 2003 were used for analysis.

Data were generated by using the model.

$$Y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

There were two different data sets generated for the purpose of the analysis.

The reason is to make comparison between the result of the first data called data A and second data called data B.

Data A was generated on the number of variable $n=10$ which were observed each for a period of 6 years. Hence the data consists of 60 observations this was replicated five(5) times. Data B was generated on the number of variable $n=10$ which were observed each for a period of 10 years. The data consist of 100 observations which was replicated ten (10) times.

Error ε_{it} data were generated from a normal distribution with mean zero and unit variance i.e $\varepsilon_{it} \sim N(0,1)$.

The random seed was set and updated seed by seed. For this data the following random seed numbers are eligible: 10, 20, 30, ...1000.

The only independent variable x_{it} was generated from a uniform distribution $U(0,1)$ in the range 20 to 30 for replications 1,2,3,4 and 5 and 20 to 40 for replication 6, 7, 8, 9 and 10 respectively.

The same set of variable x_{it} were therefore generated and used for the stated replication above.

Although these choices are arbitrary, they allow for reasonably variable but outlier.

c) Methodology

The simplest estimation method, is that one which proceeds by ignoring the panel structure of the data, it assumes the model:

$$Y = X\beta + U$$

It is assumed that the error term $U_{it} \sim iid(0, \sigma^2)$ for all i and t . That is for a given individuals, observation is serially uncorrelated, and across individuals and time, the error are homoscedastic. By assuming each observation is iid, however we have essentially ignored the panel structure of the data.

The panel data model can be written as

$$Y_{it} = X_{it} \beta + U_{it} \quad i = 1, \dots, n, t = 1, \dots, T$$

Where

Y_{it} is the observation on the dependent variable y for the i^{th} cross-sectional unit in the t^{th} period and

X_{it} is a $1 \times k$ vector observation on k explanatory variables for i^{th} individual in the t^{th} period, and

β is a $k \times 1$ vector of parameters.

U_{it} is a disturbance term and it is assumed that

$$U_{it} = \mu_i + v_{it}$$

So that the error contains an unobservable individual specific effects μ_i and a remainder disturbance v_{it} .

μ_i captures characteristics of the individual i that are not picked up by explanatory variable x_{it} but which are assumed to be time invariant. Stacking the observations first by time and then by individual, the model can be rewritten as

Where Y is an $nT \times 1$ vector, X is an $nT \times k$ matrix and U is $nT \times 1$ vector defined by

$$U = (1_n \quad \Theta \quad I_T) \mu + v$$

Where 1_T is a $T \times 1$ vector of ones, μ is $n \times 1$ vector of specific disturbances and v is $n \times 1$ vector of the remaining disturbances. Two alternative models result from different assumptions about the individual specific effect μ . These are the fixed effects and random effects models.

d) The Fixed Effects Model

The fixed effect model assumes that the individual specific effects, μ are fixed (non-stochastic) parameter to be estimated and the remaining disturbance components are independently and identically distributed with

$$E(v) = 0, \text{var}(v) = \delta^2$$

On these assumptions, the model can be written as

$$Y = D\mu + X\beta + v$$

Where $D = (1_n \theta 1_T)$ is $n \times T$ of dummy variable. This model can be estimated by OLS.

Note that, if X contains an intercept c , then the set of variable D : c will be perfectly collinear. This is the familiar problem of the dummy variable trap and can be solved either by dropping one of the columns of D and the associated element of μ or by imposing the restriction that $1' \mu = 0$.

Often we are only interested in estimating the parameters β from the familiar formula for a partitioned estimator.

e) The Pooled Estimator

Our starting point is by considering the simplest estimation method, which proceeds by ignoring the panel structure of the data.

Most times, the data are stacked to form

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

Each α_i is treated as an unknown parameters to be estimated.

Let Y_i and X_i be the T observation for the i^{th} unit, 1 is a $T \times 1$ column of ones and let ε be the associated $T \times 1$ vector of disturbances.

The model is giving as,

$$Y_i = X_i \beta + \alpha_i + \varepsilon_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \beta + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\text{or } Y = [X_1 \quad d_1 \quad d_2 \quad \dots \quad d_n] \begin{bmatrix} \beta \\ \alpha \end{bmatrix}$$

The model will now be

$$Y = X\beta + D\alpha + \varepsilon$$

Illustration of the LSDV fixed effects model with two variable, two period three from $K=2$, $T=2$, $n=3$
The stacked data will be of the form below :

Where Y is $n \times T \times 1$, X is $n \times T \times k$, ε is $n \times T \times n$ and β is $n \times T \times k$.

When the data is stacked as in the above, the simple linear model can be expressed as :

$$Y = X\beta + \varepsilon$$

We now assumed that $\varepsilon_{it} \sim \text{iid}(0, \sigma^2)$ for all i and t . I. e. for a given individual, observations are serially uncorrelated, and across individuals and time, the errors are homoscedastic.

f) Fixed Effects Model (Least Squares Dummy Variable)

The formulation of the model assumed that differences across unit can be captured in differences in the constant term.

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} X_{111} & X_{211} & 1 & 0 & 0 \\ X_{112} & X_{212} & 1 & 0 & 0 \\ X_{121} & X_{221} & 0 & 1 & 0 \\ X_{122} & X_{222} & 0 & 1 & 0 \\ X_{131} & X_{231} & 0 & 0 & 1 \\ X_{132} & X_{232} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ . \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$\text{i. e } Y = [XD] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \varepsilon$$

This model is usually referred to as the least square dummy variable (LSDV) model (although the "least squares" part of the name refers to the technique used to estimate it, not to the model itself). The model is a classical linear regression model, so no new technique is needed to estimate it. If n is small enough, then the model can be estimated by ordinary least square with k regressors in X and n columns in D , as a multiple regression with $k + n$ parameters.

III. ANALYSIS OF RESULTS

In the analysis it has been shown that there are 60 observations for which we run regression on five different periods, giving a total of 300 observations. The result as stated on the table shows that F - is significant at 95% confidence interval. Also considering the coefficient of determination, R^2 pooled of the model which is generally low meaning that the model explained a very small proportion of the total variation in the dependent variable. The R^2 compute for LSDV-fixed effects explained more of the total variation in the dependent variables.

The standard error computed for the OLS and LSDV indicates that LSDV is better and preferable because of its reduced level of errors in all replications.

Test For Significance

The table on the result contains the estimated parameters of pooled estimation and individual effects.

The F - statistic for testing the joint significant of the fixed effects is

$$F(n-1, nT-n-k) = \frac{(R^2_{\text{LSDV}} - R^2_{\text{Pooled}})/(n-1)}{(1 - R^2_{\text{LSDV}})/(nT-n-k)}$$

$$\begin{aligned} F(9, 80) &= \frac{(0.968547 - 0.613707)}{(1 - 0.968547)/40} \\ &= \frac{0.03942667}{0.0007863} \\ &= 50.41205 \end{aligned}$$

The procedure was used for other in the groups as shown in the result of analysis above

The critical value from the F - distribution table is 3.84, so the evidence is strongly in favor of an individual specific effect in the data. Thus, on this basis, there appear to be significant different across the different period in all groups tested.

On the other hand, the coefficient R^2_{LSDV} of the fixed effect model in high and this implies that the model significantly explain the variation in the dependent variable.

The standard error for each of the group that were replicated shows that standard error for OLS is very high while that of LSDV is low and this is an indication that using the fixed effects model (LSDV) will yield a consistent and efficient estimate.

3.2 Data B

The analysis has shown that there were 100 observations for which we run regression on ten different periods, giving a total of 1000 observation. The result as stated below shows that F - is significant at 95% confidence interval. Also considering the coefficient of determination, R^2 pooled of the model which is generally low meaning that the model explained a very small proportion of total variation in the dependent variables. The R^2 compute for LSDV - Fixed effects model explained more of the total variation in the dependent variables.

The standard error computed for both OLS and LSDV indicates LSDV is better and preferable because of its reduced level of errors in all replications.

Test For The Significance

The table on the result contains the estimated parameters of pooled estimation and individual effects. The F -statistic for testing the joint significant of the fixed effect is

$$\begin{aligned} F(n-1, nT-n-k) &= \frac{(R^2_{\text{LSDV}} - R^2_{\text{LSDV}})/(n-1)}{(1 - R^2_{\text{LSDV}})/(nT-n-k)} \\ F(9, 80) &= \frac{(0.9983 - 0.5609)/9}{(1 - 0.9983)/80} \end{aligned}$$

$$= \frac{0.0486}{0.00002125}$$

$$= 2287.0588$$

The same procedure was used for other in the groups as shown in the result of analysis.

The critical value from the F – distribution is 2.10, so the evidence is strongly in favor of an individual specific effect in the data. Thus, on this basis, there appears to be significant different across the different period in all the groups tested. On the other hand, the coefficient of determination, R^2_{LSDV} of the fixed effects model is high and this implies that the model significantly explain the variation in the dependent Variable

IV. INTERPRETATION OF FINDINGS, CONCLUSION AND RECOMMENDATION

a) Interpretation of Findings

The following were observed when comparing the results of analysis of DATA a and DATA B in respect of OLS and LSDV estimates.

- i. The R^2 and $\bar{R}^2$ are higher for the bigger sample size i. e data B than the smaller sample size i. e data A
- ii. The standard errors term is smaller for the bigger sample size than the smaller sample size.
- iii. The computed F is remarkably higher for the bigger sample size.
- iv. The ratios of the various statistics used for the bigger sample over the smaller sample are remarkably higher for the bigger sample size.

This suggest that the LSDV is not only supervisor and efficient to the OLS , it's superiority increases asymptotically.

b) Conclusion

The results of the analysis have shown that there are significant differences across the different periods in the groups under consideration. Also F is significant at 95% confidence interval by the use of either least squares dummy variable model or pooled model of panel data. The coefficient of Determination R^2 that measure of goodness of fit has shown that the fixed effect model significantly explains the variable in the dependent variation then the OLS model in most cases. Hence, the fixed effects model should be preferred, as it has yielded efficient and more reliable result on the basis of the assumptions under which the small-scale monte-carlo experiment was preferred.

c) Recommendation

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. However, the use of fixed effect model sometimes called least

square dummy variable model is recommended as it gives a consistent and efficient estimate.

There are other methods of analyzing panel data in econometrics depending on the econometric problem to be addressed, such methods include the random effect model, unrelated regression model, dynamic model, unbalance panel data model e.t.c. It is recommended that further study / research work should focus on the use of these methods.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Alan Stuart & Kandall M (1976): The Advanced theory of statistics. Macmillan Publishing New York.
2. Allison, P. D (1994) using panel data to estimate the effects of events. Sociological method and research 23: 174-199.
3. Andrew Donald (1993) "Test for parameter instability and structure change point", Econometrical 821-856.
4. Badi H. Baltagi, (2004) : panel data Econometrics: Theoretical contribution and Empirical application. International journal presentation at Texas University.
5. David Bower(1985): Statistics for Econometrics and business, 2nd edition ELBS Imprint.
6. Finkel, S. (1995) causal analysis with panel data.
7. Hausman & Mundlak (1978): Specification tests in econometrics' Econometrics 46.1251-1271.
8. Humberto Barreto & Frank M. Howland (2006): Using Monte Calo Simulation with Microsoft Excel Published by Cambridge University Press. London .
9. Johnston J. – (1985) Econometrics method, 2nd edition McGraw hill.
10. Microsoft Office 2003 and SPSS version 13.0 for windows www.spss.com.http://www.Google.com.
11. Nickell, S (1981): 'Biases in dynamic model with fixed effect' Econometrics, 14, 1417-1426.
12. Wooldridge, J (2003) Introductory Econometrics. A modern approach. Thomson Chapter 13, 14 (easy introduction).
13. Wooldridge, J (2002) Econometrics Analysis of cross section and panel data. MIT press more advanced, but very thorough.

Summary of Results- DATA A

	R^2			$\overline{R}^2$			SE			F		
REPLI-CAUTION	OLS	LSDV	$\underline{R^2_1}$	OLS	LSDV	$\underline{\overline{R^2_1}}$	OLS	LSDV	$\underline{SE_0}$	OLS	LSDV	$\underline{F_1}$
	R^2_0	R^2_1	R^2_0	$\overline{R^2_0}$	$\overline{R^2_1}$	$\overline{R^2_0}$	SE_0	SE_0	SE_1	F_0	F_1	F_0
1.	0.6137	0.9685	1.578	0.607	0.942	1.552	2.9611	0.919	3.222	92.14	150.88	1.638
2.	0.1118	0.6159	5.509	0.964	0.517	5.36	5.196	3.718	1.398	7.29	7.857	1.078
3.	0.5766	0.9352	1.622	0.569	0.902	1.58	2.890	1.229	2.352	79.008	70.833	0.897
4.	0.7984	0.9798	1.227	0.795	0.955	1.201	3.087	1.060	2.912	229.695	238.667	1.568
5.	0.7943	0.9855	1.245	0.794	0.963	1.213	3.050	0.856	3.563	228.871	358.910	1.568

R.....Coefficient of determination

SE Standard error

 $\overline{R}^2$ Adjusted coefficient of multiple determination

F.....F distribution(significant level)

Summary of Results- DATA B

	R^2			$\overline{R}^2$			SE			F		
REPLI - CAUTION	OLS	LSDV	$\underline{R^2_1}$	OLS	LSDV	$\underline{\overline{R^2_1}}$	OLS	LSDV	$\underline{SE_0}$	OLS	LSDV	$\underline{F_1}$
	R^2_0	R^2_1	R^2_0	$\overline{R^2_0}$	$\overline{R^2_1}$	$\overline{R^2_0}$	SE_0	SE_1	SE_1	F_0	F_1	F_0
1.	0.5609	0.9983	1.7793	0.5564	0.9869	1.7737	2.8659	0.1868	15.3420	125.1933	5243.52	41.883
2.	0.6951	0.9987	1.4368	0.6919	0.9874	1.4271	2.9195	0.1964	14.8650	223.3694	7093.988	31.759
3.	0.6573	0.9978	1.5180	0.6538	0.9863	1.5086	2.8878	0.2455	11.7629	187.9249	3948.485	21.011
4.	0.6578	0.9984	1.5178	0.6544	0.9870	1.5083	2.9293	0.2065	14.1855	188.4185	5752.161	30.529
5.	0.5872	0.9989	1.7011	0.5829	0.9875	1.6941	2.9158	0.1595	18.2809	139.3715	7919.631	56.824
6.	0.5609	0.9983	1.7798	0.5564	0.9869	1.7737	2.8659	0.1868	15.3421	125.1933	5243.52	41.883
7.	0.6951	0.9987	1.4369	0.6919	0.9874	1.4271	2.9195	0.1964	14.8651	223.3694	7093.98	31.759
8.	0.6573	0.9978	1.5180	0.6538	0.9863	1.5086	2.8878	0.2455	17.7629	187.9249	3948.485	21.010
9.	0.6578	0.9984	1.5178	0.6544	0.9870	1.5083	2.9293	0.2065	14.1855	188.4185	5752.161	30.529
10.	0.5871	0.9989	1.7014	0.5829	0.9875	1.6941	2.9158	0.1595	18.2808	139.3715	7919.631	56.823

 R^2 coefficient of determination

SE Standard error

 $\overline{R}^2$ Adjusted coefficient of multiple determination

F F distribution