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

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 (R², 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 (R², 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: R² – Coefficient of determination, LSDV- least square dummy variables, Panel Data.

GJSFR-F Classification: FOR Code: 010299

© 2011 . O.D. Ogunwale, A.O. Olubiyi, A.H. Bello. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
On The Comparison of Two Methods of Analyzing Panel Data Using Simulated Data

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

Abstract - The study focuses 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 (R², Standard error, t and t 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: R² – 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 over time, 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 that is available 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 collinearity, 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 for Economic 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
uncertainty about whether the $\mu$ 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;

i. Parameters estimate of the pooled observations using ordinary least squares

ii. Estimate the individual effect parameters by using least squares dummy variables: fixed effect model

iii. 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.

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

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

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 $x_t$ 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 $x_t$ 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 giving 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

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

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 1xk vector observation on k explanatory variables for $i^{th}$ individual in the $t^{th}$ period, and

$B$ is a $kx1$ 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 $\mu_i$ and a remainder disturbance $v_{it}$.

$\mu_i$ captures characteristics of the individual I that are not picked up by explanatory variable $x_{it}$ but which are assumed to be time invariable. 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

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

Where $l_T$ is a $T \times 1$ vector of ones, $\mu$ 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 $\mu$. These are the fixed effects and random effects models.

d) The Fixed Effects Model

The fixed effects model assumes that the individual specific effects $\mu$ 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 \( D = \{1_n \Theta l_T \} \) is \( n \times n \) 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 \( \mu \) or by imposing the restriction that \( l'\mu = 0 \).

Often we are only interested in estimating the parameters \( \beta \) 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 \( \alpha_i \) is treated as an unknown parameters to be estimated.

Let \( Y_i \) and \( X_i \) be the \( T \) observation for the \( i \)th unit, \( l \) is a \( T \times 1 \) column of ones and let \( \varepsilon \) be the associated \( T \times 1 \) vector of disturbances.

The model is giving as,

\[
Y_i = X_i \beta + \alpha_i + \varepsilon_i
\]

or

\[
Y = [X1 \ d1 , \ d2 \ \ldots \ \ldots \ dn] [\beta \ \alpha] + \varepsilon
\]

The model will now be

\[
Y = X\beta + D\alpha + \varepsilon
\]
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.

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 if it’s 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

\[
F(9,80) = \frac{(0.9983 - 0.5609)/9}{(1 - 0.9983)/80} = 50.41205
\]
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, its 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

### Summary of Results - DATA A

<table>
<thead>
<tr>
<th>REPLICATION</th>
<th>OLS $R^2$</th>
<th>LSDV $R^2$</th>
<th>OLS $\bar{R}^2_0$</th>
<th>LSDV $\bar{R}^2_1$</th>
<th>SE $\bar{R}^2_0$</th>
<th>SE $\bar{R}^2_1$</th>
<th>F</th>
<th>F$_0$</th>
<th>F$_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.6137</td>
<td>0.9685</td>
<td>1.578</td>
<td>0.607</td>
<td>0.942</td>
<td>1.552</td>
<td>2.9611</td>
<td>0.919</td>
<td>3.222</td>
</tr>
<tr>
<td>2</td>
<td>0.1118</td>
<td>0.6159</td>
<td>5.509</td>
<td>0.964</td>
<td>0.517</td>
<td>5.36</td>
<td>5.196</td>
<td>3.718</td>
<td>1.398</td>
</tr>
<tr>
<td>3</td>
<td>0.5766</td>
<td>0.9352</td>
<td>1.622</td>
<td>0.569</td>
<td>0.902</td>
<td>1.58</td>
<td>2.890</td>
<td>1.229</td>
<td>2.352</td>
</tr>
<tr>
<td>4</td>
<td>0.7984</td>
<td>0.9798</td>
<td>1.227</td>
<td>0.795</td>
<td>0.955</td>
<td>1.201</td>
<td>3.087</td>
<td>1.060</td>
<td>2.912</td>
</tr>
<tr>
<td>5</td>
<td>0.7943</td>
<td>0.9855</td>
<td>1.245</td>
<td>0.794</td>
<td>0.963</td>
<td>1.213</td>
<td>3.050</td>
<td>0.856</td>
<td>3.563</td>
</tr>
</tbody>
</table>

- $R^2$ .......... Coefficient of determination
- $\bar{R}^2$ .......... Adjusted coefficient of multiple determination
- SE .......... Standard error
- F ............... F distribution (significant level)

### Summary of Results - DATA B

<table>
<thead>
<tr>
<th>REPLICATION</th>
<th>OLS $R^2$</th>
<th>LSDV $R^2$</th>
<th>OLS $\bar{R}^2_0$</th>
<th>LSDV $\bar{R}^2_1$</th>
<th>SE $\bar{R}^2_0$</th>
<th>SE $\bar{R}^2_1$</th>
<th>F</th>
<th>F$_0$</th>
<th>F$_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.5609</td>
<td>0.9983</td>
<td>1.7793</td>
<td>0.5564</td>
<td>0.9869</td>
<td>1.7737</td>
<td>2.8659</td>
<td>0.1868</td>
<td>15.3420</td>
</tr>
<tr>
<td>2</td>
<td>0.6951</td>
<td>0.9987</td>
<td>1.4368</td>
<td>0.6919</td>
<td>0.9874</td>
<td>1.4271</td>
<td>2.9195</td>
<td>0.1964</td>
<td>14.8650</td>
</tr>
<tr>
<td>3</td>
<td>0.6573</td>
<td>0.9978</td>
<td>1.5180</td>
<td>0.6538</td>
<td>0.9863</td>
<td>1.5086</td>
<td>2.8878</td>
<td>0.2455</td>
<td>11.7629</td>
</tr>
<tr>
<td>4</td>
<td>0.6578</td>
<td>0.9984</td>
<td>1.5178</td>
<td>0.6544</td>
<td>0.9870</td>
<td>1.5083</td>
<td>2.9293</td>
<td>0.2065</td>
<td>14.1855</td>
</tr>
<tr>
<td>5</td>
<td>0.5872</td>
<td>0.9989</td>
<td>1.7011</td>
<td>0.5829</td>
<td>0.9875</td>
<td>1.6941</td>
<td>2.9158</td>
<td>0.1595</td>
<td>18.2809</td>
</tr>
<tr>
<td>6</td>
<td>0.5609</td>
<td>0.9983</td>
<td>1.7798</td>
<td>0.5564</td>
<td>0.9869</td>
<td>1.7737</td>
<td>2.8659</td>
<td>0.1868</td>
<td>15.3421</td>
</tr>
<tr>
<td>7</td>
<td>0.6951</td>
<td>0.9987</td>
<td>1.4369</td>
<td>0.6919</td>
<td>0.9874</td>
<td>1.4271</td>
<td>2.9195</td>
<td>0.1964</td>
<td>14.8651</td>
</tr>
<tr>
<td>8</td>
<td>0.6573</td>
<td>0.9978</td>
<td>1.5180</td>
<td>0.6538</td>
<td>0.9863</td>
<td>1.5086</td>
<td>2.8878</td>
<td>0.2455</td>
<td>17.7629</td>
</tr>
<tr>
<td>9</td>
<td>0.6578</td>
<td>0.9984</td>
<td>1.5178</td>
<td>0.6544</td>
<td>0.9870</td>
<td>1.5083</td>
<td>2.9293</td>
<td>0.2065</td>
<td>14.1855</td>
</tr>
<tr>
<td>10</td>
<td>0.5871</td>
<td>0.9989</td>
<td>1.7014</td>
<td>0.5829</td>
<td>0.9875</td>
<td>1.6941</td>
<td>2.9158</td>
<td>0.1595</td>
<td>18.2808</td>
</tr>
</tbody>
</table>