Journal of Econometrics and Statistics
Vol. 3, Issue 1, 2023, pp. 91-105
https://doi.org/10.47509/JES.2023.v03i01.06

A Comprehensive Study of Model Comparisons in Panel Data with Non-Linearity Between Outcome and Explanatory Variables

Meghna Athwani[a]* and Rajeev Pandey[b]

[a]*Department of Statistics, University of Lucknow, Lucknow, India;* [b]*Department of Statistics, University of Lucknow, Lucknow, India*

***Corresponding Author:** Meghna Athwani, Ph.D. Scholar, Department of Statistics, University of Lucknow, India - 226007; Email: meghnaathwani@gmail.com

The objective of this paper is to provide insights on some widely used Panel Data Econometric Models to determine the best one. With this aim, data of Organized Manufacturing Sector from the Annual Survey of Industries has been used. Pooled OLS, First-Difference, Time-Demeaning and Random Effects models have been estimated using a sample of 26 states, over an 18-year period (2000-2017), where Input Variables are in a non-linear relationship with Total Output. The findings of the study recommend that in a panel setting the most robust model is the Time-Demeaning Fixed Effects model.

Keywords: Annual Survey of Industries; Fixed Effects; Panel Data; Pooled OLS; R; Random Effects

1. INTRODUCTION

A panel dataset is defined as a cross-sectional time-series dataset, which provides repeated measurements of a certain number of variables over a period on observed units, such as individuals households, companies, countries, cities, or states (Miller, 2008). Therefore, panel data observations involve at least two dimensions; a time series dimension, indicated by subscript t and a cross-sectional dimension, indicated by subscript i (Hsiao, 2007).

The Annual Survey of Industries (ASI) is the principal source of industrial statistics in India. It provides statistical information to assess and evaluate, the changes in growth, composition, and structure of organised manufacturing sector of the country comprising of several activities related to manufacturing processes, gas and water supply, repair services, and cold storage (MOSPI, n.d.).

Article History
Received : 11 April 2023; Revised : 02 May 2023; Accepted : 10 May 2023; Published : 27 May 2023

To cite this paper
Meghna Athwani & Rajeev Pandey (2023). A Comprehensive Study of Model Comparisons in Panel Data with Non-Linearity between Outcome and Explanatory Variables. *Journal of Econometrics and Statistics*. 3(1), 91-105. https://doi.org/10.47509/JES.2023.v03i01.06

In the present study, data of ASI has been analysed **using R software (version 4.0.2)** (R Core Team, 2022). Pooled Ordinary Least Square (OLS), Fixed Effects and Random Effects Model have been incorporated and their results have been compared to suggest a suitable fit for the desired Panel Data. The Lagrange's Multiplier and the Wu-Hausman Tests have also been applied to detect the best model for the analysis of Data.

The proceeding sections give a brief review of the database, followed by a description of the methodology used, along with analysis and a summary of the results. The paper completes by summarizing its main conclusions.

2. MATERIAL AND METHODS

2.1. DATA

The present study utilizes yearly data ranging from 2000-01 to 2017-18, compiled from several publications of ASI and obtained from the National Data Archive. It has been contributed by the Central Statistics Office (Industrial Statistics Wing) - Ministry of Statistics & Programme Implementation (MOSPI), Govt. of India. It consists of Production in Organized Manufacturing Sector in India with respect to 26 states, viz: Assam, Andhra Pradesh, Bihar, Chhattisgarh, Delhi, Gujarat, Goa, Himachal Pradesh, Haryana, Jharkhand, Jammu & Kashmir, Kerala, Karnataka, Maharashtra, Madhya Pradesh, Meghalaya, Manipur, Nagaland, Odisha, Punjab, Rajasthan, Tripura, Tamil Nadu, Uttarakhand, Uttar Pradesh, West Bengal. The states have been shortlisted for the study as there were no gaps in data with respect to the chosen time period.

Six key input (explanatory) variables have been considered for this study. **Fixed Capital** has been included as it aids firms in expansion and diversification and to replace dated and scrapped assets, while **Working capital** has been included because manufacturing companies are subject to challenges as supplier and production expenses frequently require payment several months before goods are sold to customers. Wage related two variables namely, **Wages and Salaries of Workers**; **Provident and Other Funds, Workmen and Staff Welfare Expenses** have been examined as Endogenous Growth literature highlights that higher wage growth may be one of the factors stimulating capital investment in new technology. Other factors of Production like **Rent Paid** and **Interest Paid** have also been scrutinised.

The sole **output (outcome) variable** is the value of **Total Output** which entails total ex-factory value of products and by-products manufactured as well as other receipts from non-industrial services rendered to others, work done for others on material supplied by them, value of electricity produced and sold, sale value of goods sold in the same conditions purchased, addition in stock of semi- finished goods and value of own construction. All the values are expressed in Rupees lakhs.

2.2. METHODOLOGY

Four Panel Data Models have been considered, viz: **Pooled OLS model**, **Fixed Effects** (**First Difference** and **Time Demeaning**) models and **Random Effects Model**.

For diagnosis, **Lagrange's Multiplier test** and **Wu-Hausman's Specification test** have been applied to detect the best fit.

General Regression Model for Panel Data

Panel data contains information on temporal and spatial dimensions (Xu, et al. 2007). Here, the temporal dimension is the period in which repeated measurements are taken over the years and the spatial dimension are the states which is the unit of observation.

The general regression model for panel data can be expressed as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} + v_{it} \; ; \qquad (1)$$

$$i = 1,2, \dots , N; \; t = 1,2, \dots , T; \; k = 1,2, \dots , K$$

Where i is the unit of observation, N is the spatial dimension, t is the period of time, T is the temporal dimension, k indicates the k[th] independent (explanatory) variable, β_0 is the intercept term, β_k is the coefficient of each independent variable, K is the number of explanatory variables and v_{it} is the random noise.

v_{it} is the composite error term in Equation (1) and can be decomposed into two components, viz: an idiosyncratic error, u_{it} and a cross-sectional unit-specific error, a_i.

$$v_{it} = a_i + u_{it} \tag{2}$$

The error a_i, does not change over time but, u_{it}, varies over the cross-sectional units and time (Baltagi, 2001; Greene, 2003; Gujarati, 2003). The error term is decomposed into two parts so that if we eliminate some part of it using panel data, we would be able to minimize concerns for omitted variable bias caused by unmeasured unit-specific factors.

If we incorporate Equation (2) in Equation (1), we can get an error component model as:

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} + a_i + u_{it} \tag{3}$$

The time-constant and unit-specific errors, a_i, are unobserved factors. The estimation methods of error component models are classified based on how to treat a_i. The pooled OLS model does not distinguish it from other types of errors, whereas the fixed effects model regards it as coefficients to be estimated, and the random effects model treats it as random variables (Baltagi, 2001; Greene, 2003; Maddala, 2001).

To estimate the Total Output, we specify the following equation:

$$VO_{it} = \beta_0 + \beta_1 FC_{it} + \beta_2 WC_{it} + \beta_3 WW_{it} + \beta_4 PFO_{it} + \beta_5 RP_{it} + \beta_6 IP_{it} + v_{it} \tag{4}$$

Where,

Outcome Variable	Explanatory Variables
$\boldsymbol{VO_{it}}$ is the value of **Total Output**	$\boldsymbol{FC_{it}}$ is the value of **Fixed Capital**
	$\boldsymbol{WC_{it}}$ is the value of **Working Capital**
	$\boldsymbol{WW_{it}}$ is the value of **Wages and Salaries of workers**
	$\boldsymbol{PFO_{it}}$ is the value of **Provident Fund, Other Funds, Workmen and Staff Welfare Expenses**
	$\boldsymbol{RP_{it}}$ is the value of **Rent Paid**
	$\boldsymbol{IP_{it}}$ is the value of **Interest Paid**

2.2.1. The Pooled OLS Model

OLS is used to estimate Equation (1) after pooling the data. It is assumed that v_{it} is not correlated with the independent variables $x_{it,k}$, i.e., only when there are neither cross-sectional nor temporal effects. The subscripts *i* and *t* will not appear in the model.

The pooled OLS version of Equation (1) is expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + v \qquad (5)$$

2.2.2. The Fixed Effects Model

Equation (3) is estimated using the Fixed Effects model. It is assumed that error a_i is correlated with $x_{it,k}$. The error u_{it} is independent of $x_{it,k}$ is another important assumption (Baltagi, 2001; Wooldridge, 2006).

The standard fixed effects panel data model is:

$$y_{it} = a_i + x'_{it}\beta + u_{it}; \quad i = 1,2, \dots, N; \quad t = 1,2, \dots, T \qquad (6)$$

Where y_{it} an observation on the dependent variable observed for the individual i at time t, x_{it} is a $K\ x\ 1$ vector of observations on K explanatory variables, β is a $K\ x\ 1$ vector of the regression parameters, a_i denotes the individual effect and u_{it} is the error term.

The unobserved effect a_i is eliminated to reduce the omitted variables biases and obtain robust estimates. The two widely used methods, i.e., the first-difference model and the time-demeaning model for eliminating a_i in the panel data analysis are discussed.

The First-Difference (FD) Model

Here we eliminate a_i by differencing the data across two time periods. If y_t denotes the value of *y* at period *t*, then the first difference of *y* at period *t* equals $y_t - y_{t-1}$. Thus, on subtracting Equation (3) at the period of time 1 from Equation (3) at the period of time 2, we get the following equation:

$$(y_{i2} - y_{i1}) = \beta_1(x_{i2,1} - x_{i1,1}) + \cdots + \beta_k(x_{i2,k} - x_{i1,k}) + (u_{i2} - u_{i1}); \qquad (7)$$

$$or\ \Delta y_i = \beta_k \Delta x_{i,k} + \Delta u_i$$

Here Δ indicates the change of time-period from 1 to 2.

The Time-Demeaning (TD) Method

This is used to transform the original variables into the deviations from the group means of each variable. Averaging Equation (3) over time (*t*), we obtain:

$$\bar{y}_i = \beta_0 + \beta_1\bar{x}_{i1} + \beta_2\bar{x}_{i2} + \cdots + \beta_k\bar{x}_{ik} + a_i + \bar{u}_i \qquad (8)$$

Subtracting Equation (8) from Equation (3) we get:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{it,1} - \bar{x}_{i1}) + \beta_2(x_{it,2} - \bar{x}_{i2}) + \cdots + \beta_k(x_{it,k} - \bar{x}_{ik}) + (u_{it} - \bar{u}_i) \quad (9)$$

Where, $(y_{it} - \bar{y}_i)$, $(x_{it,1} - \bar{x}_{i1})$ and $(u_{it} - \bar{u}_i)$ are the time-demeaned values of *y, x,* and *u* respectively.

2.2.3. The Random Effects (RE) Model

The fixed effects model is used to eliminate the unobserved heterogeneity a_i as it is assumed to be correlated with any of the $x_{it,k}$. However, when a_i is independent of each explanatory variable, the fixed effects model will result in estimators that are inefficient (Baltagi, 2001; Greene, 2003). The random effects model regards a_i as random variables rather than fixed ones and is thus appropriate when the cross-sectional units are randomly selected from a large population.

The standard random effects panel data model is:

$$y_{it} = \alpha + a_i + x'_{it}\beta + u_{it}; \quad i = 1,2, \dots, N; \quad t = 1,2, \dots, T \tag{10}$$

Where y_{it} an observation on the dependent variable observed for the individual i at time t, x_{it} is a $K\ x\ 1$ matrix of observations on K explanatory variables, β is a $K\ x\ 1$ matrix of the regression parameters, α denotes the intercept term, a_i are the individual effects and u_{it} is the error term. In random effects model we assume a_i to be uncorrelated with x_{it} whereas in fixed effects model a_i may be correlated with x_{it}.

Tests for Detecting Best Fit.

2.2.4. Lagrange Multiplier (LM) Test

Breusch and Pagan (1980) developed a test to check for the existence of the random effects. The null hypothesis is that individual (or time) specific variance components are zero. $H_0: \sigma_u^2 = 0$. Greene (2003) suggested that if the null hypothesis is rejected, the random effect model is better than the pooled OLS regression model.

The LM statistic follows the chi-squared distribution with one degree of freedom.

$$LM_u = \frac{nT}{2(T-1)}\left[\frac{T^2\bar{e}'\bar{e}}{e'e} - 1\right]^2 \sim \chi^2_{(1)} \tag{11}$$

where *e* is the *n* × *1* vector of the group means of pooled regression residuals, and *e'e* is the SSE of the pooled OLS regression.

2.2.5. Wu-Hausman's (WH) Specification Test

Hausman (1978) provided the specification test often used to compare a random effect model to its fixed counterpart. The test compares the fixed versus random effects in the panel data models under the null hypothesis that the individual effects (a_i) are independent of the $x_{it,k}$ in the model (Baltagi, 2001; Greene, 2003; Maddala, 2001). If the null hypothesis is not rejected, then one should preferably use random effects model as it produces estimators that are more efficient.

This test statistic follows the chi-squared distribution with k degrees of freedom.

$$WH = \left(b_{fixed} - b_{random}\right) W^{-1} \left(b_{fixed} - b_{random}\right) \sim \chi^2_{(k)} \qquad (12)$$

Where, $W = Var\left(b_{fixed} - b_{random}\right) = Var(b_{fixed}) - Var(b_{random})$ is the difference in the estimated covariance matrices of the robust model) and the efficient model.

3. RESULTS

The exploratory data analysis can be gauged through **Figures 1 and 2** which give a brief idea about its heterogeneity across years and states.

Fig. 1: Heterogeneity across Years.

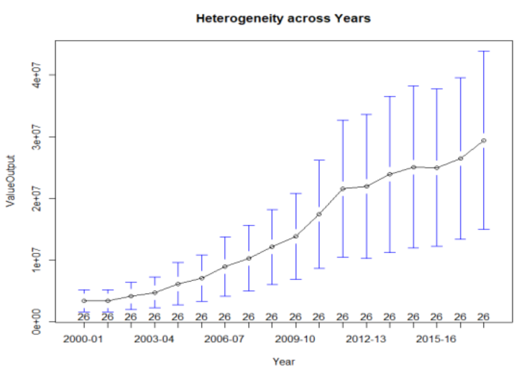

Fig. 2: Heterogeneity across States.

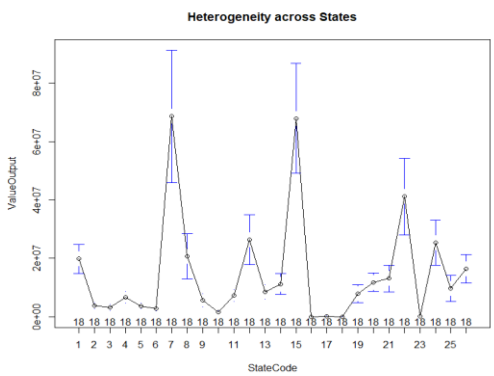

Fig. 3: Scatter Plots of Total Output (Y) vs All explanatory variables (X)

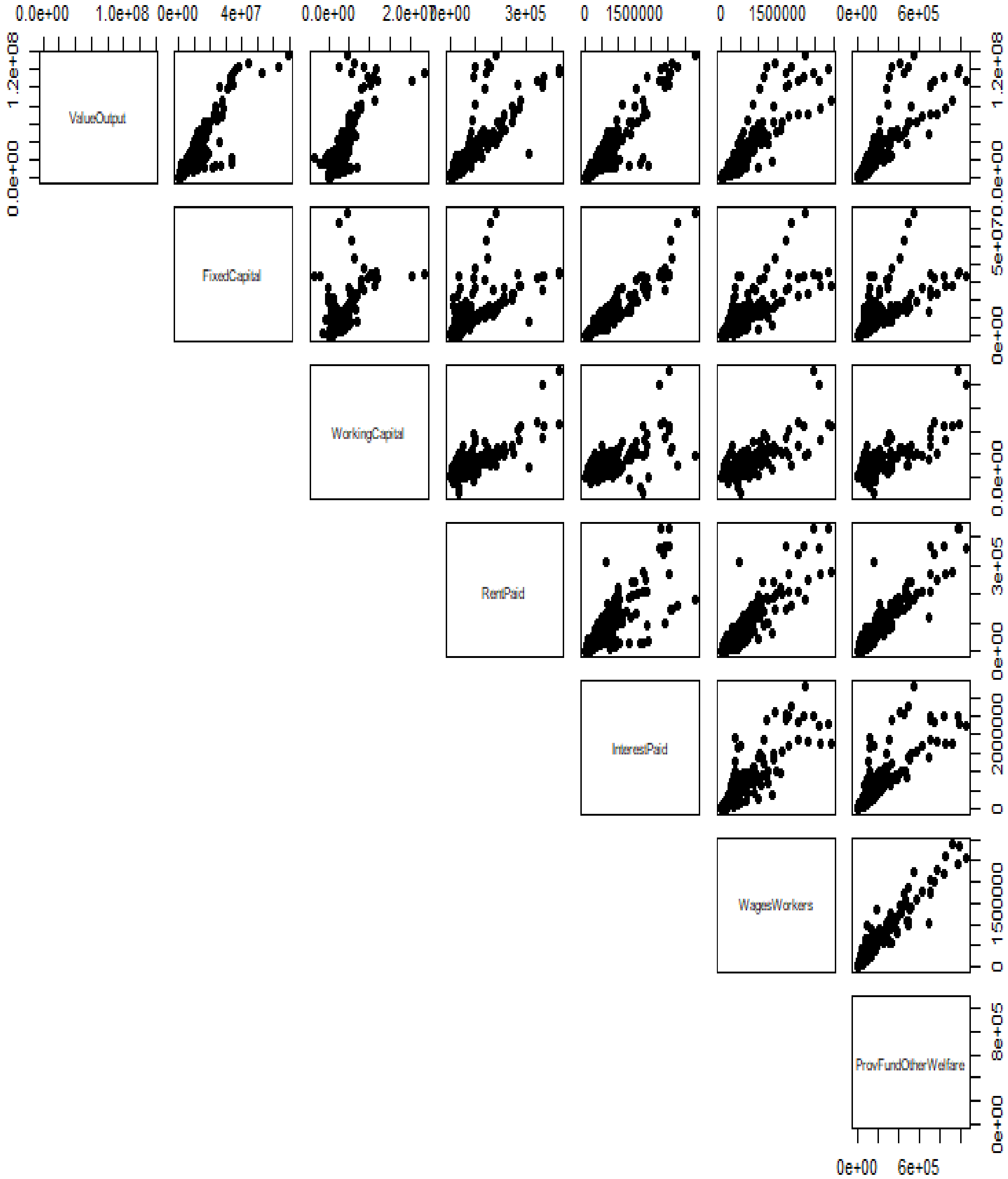

Table 1: Value of nonlinear correlation between variables

Variables		**nlcor value**
Total Output	**Fixed Capital**	0.912
	Working Capital	0.840
	Rent Paid	0.867
	Interest Paid	0.938
	Provident Fund and Other Welfare Fund	0.894
	Wages and Workers	0.916

Table 2: Summary of estimates.

Type of Model	Variable	Coefficient (S.E.)	p-value
Pooled OLS model	**(Intercept)**	-713940 (329000)	**0.031***
	Fixed Capital	0.849 (0.119)	**< 0.001***
	Working Capital	2.447 (0.229)	**< 0.001***
	Rent Paid	19.803 (12.84)	0.125
	Interest Paid	10.821 (2.489)	**< 0.001***
	Wages and Salaries	11.531 (2.468)	**< 0.001***
	Provident Fund, Other Funds	-9.300 (8.204)	0.258
	Adj. R-Squared	0.937	
	p-value	**< 0.001 ***	
First-Difference Fixed Effects model	**(Intercept)**	-631420 (128940)	**< 0.001***
	Fixed Capital	0.293 (0.084)	**< 0.001***
	Working Capital	0.336 (0.117)	**0.004***
	Rent Paid	22.4660 (5.532)	**< 0.001***
	Interest Paid	13.6340 (1.432)	**< 0.001***
	Wages and Salaries	2.7219 (1.018)	**0.008***
	Provident Fund, Other Funds	8.1881 (3.666)	**0.026***
	Adj. R-Squared	0.432	
	p-value	**< 0.001 ***	
Time-Demeaning Fixed Effects model	**Fixed Capital**	0.694 (0.120)	**< 0.001***
	Working Capital	1.809 (0.204)	**< 0.001***
	Rent Paid	14.148 (11.570)	0.222
	Interest Paid	11.557 (2.238)	**< 0.001***
	Wages and Salaries	10.011 (2.259)	**< 0.001***
	Provident Fund, Other Funds	7.071 (7.694)	0.359
	Adj. R-Squared	0.898	
	p-value	**< 0.001 ***	
Random Effects model	**(Intercept)**	-272180 (787060)	0.729
	Fixed Capital	0.720 (0.117)	**< 0.001***
	Working Capital	1.884 (0.201)	**< 0.001***
	Rent Paid	14.691 (11.445)	0.199
	Interest Paid	11.480 (2.211)	**< 0.001***
	Wages and Salaries	10.038 (2.231)	**< 0.001***
	Provident Fund, Other Funds	5.301 (7.527)	0.481
	Adj. R-Squared	0.909	
	p-value	**< 0.001 ***	

*$p<0.05$

We have used the nlcor package in R (Ranjan, 2020) to find the nonlinear correlation between two data vectors. On observing the values from **Table 1** and the scatter plots in **Figure 3** we deduce that the relationship between the Outcome and the Explanatory variables is **Non-Linear** and the main assumption of OLS method does not follow. Owing to which, we are not including OLS model and thus, Pooled OLS is considered.

On observing the p-values of the explanatory variables from **Table 2**, the regressors FC, WC, IP and WW are significant at 5% level. RP and PFO are not significant at the 5% level as their p-value is greater 0.05.

The model can be written as:

$$\widehat{VO_{it}} = -713940 + 0.8496\, FC_{it} + 2.4473\, WC_{it} + 11.531\, WW_{it} + 10.821\, IP_{it}$$

Value of Adjusted R Square is greater than 0.9. So, if we were to assume no dependence within individual groups, our panel data could be treated as one large, pooled dataset. But linear independence within the states of our data is unlikely and thus **pooled OLS Model is not acceptable**.

Below the Pooled OLS method are the results of the FD Method. From the p-values corresponding to the coefficients of the explanatory variables, it is seen that all the regressors, i.e., FC, WC, RP, IP, WW and PFO are significantly different from zero at 5% level. The model can be written as:

$$\widehat{VO_{it}} = -631420 + 0.2934\, FC_{it} + 0.3364\, WC_{it} + 2.7219\, WW_{it} + 8.1881\, PFO_{it} + 22.466\, RP_{it} + 13.634\, IP_{it}$$

But the value of Adjusted R square is very low (0.43). Thus, the **First-Difference model cannot be accepted as a good fit to the data**.

Econometric analysis carried out with the help of TD Method is shown below the FD method in **Table 2**. The p-values of variables FC, WC, IP, and WW prove that they are significantly different than zero. On the other hand, RP and PFO are not significant in this model. The model can be written as:

$$\widehat{VO_{it}} = 0.6937\, FC_{it} + 1.8087\, WC_{it} + 10.0111\, WW_{it} + 11.5571\, IP_{it}$$

The Adjusted R square value is around 0.9 which indicates that the **Time-Demeaning Fixed effects Model is a Good Fit to the data.**

The results of the RE model somewhat resemble that of the TD Fixed Effect model as seen in the same table. FC, WC, IP, and WW are significant explanatory variables whereas RP and PFO are not considered as significant. The model can be written as:

$$\widehat{VO_{it}} = -272180 + 0.72\, FC_{it} + 1.8836\, WC_{it} + 10.038\, WW_{it} + 11.48\, IP_{it}$$

The Adjusted R square value is around 0.9 which **indicates Random Effects Model is a Good Fit.** The only visible difference is in the coefficients of the variables in both the models. Here, we use Hausman test to decide which test should be preferred.

Computation of the **LM test (studentized Breusch-Pagan test)** with the help of R software gives a very small p-value (< 0.0001). Thus, we reject the null hypothesis and conclude that there is a significant random effect in the ASI Panel data, and thus, the **RE model is able to deal with heterogeneity across the States better than the Pooled OLS model**.

The computation of the **Wu-Hausman test** gives a p-value of 0.0283. Thus, the null hypothesis of no correlation is rejected, i.e., individual effects u_i are significantly correlated with at least one explanatory variable in the model. Hence, we conclude that the random effect model is problematic, and there is no sufficient statistical evidence to reject the fixed effect estimators.

Therefore, the **TD fixed effects model is better than the RE model** for ASI Panel Data.

4. CONCLUSION

The objective of this study is to gauge the effect of various Input variables on the Total Output of all the industries in the organized manufacturing sector of India. Consequently, panel data regression models are estimated using these Input and Output variables.

Pooled OLS disregards useful information of both time as well as cross-sectional dimensions and thus, the estimator obtained through this method is biased and inconsistent. Although, pooling cross-sectional data increases the number of observations, secures more variations of the key explanatory variables, and produces significant estimates and therefore, Pooled OLS models are used in public administration literature quite frequently.

Fixed effect models allow us to identify causal effects between the States, and they are constant within the States. The results of the FD model differ quite significantly from that of the TD model even though both are Fixed effect models.

Derivation of two estimators having different properties depends on the correlation between a_i and the explanatory variables. If the effects are not correlated with the regressors, the TD fixed effects estimator will be consistent but not efficient whereas the RE estimator is consistent and efficient. In case the effects are correlated with the regressors, the RE estimator becomes inconsistent but the TD fixed effects estimator will be consistent and efficient. Thus, the TD Fixed Effects Model is correct on the basis of Wu-Hausman test.

ACKNOWLEDGEMENTS

We are extremely grateful to Dr. Shambhavi Mishra and Ms. Kuhu Awasthi for their invaluable assistance and support towards the analysis performed.

DATA AVAILABILITY STATEMENT

The Panel data from Annual Survey of Industries that supports our findings in this study is available in **figshare** at https://doi.org/10.6084/m9.figshare.14241824.v1, reference number 14241824. This data was derived from **National Data Archive**, available in public domain and can be accessed from http://microdata.gov.in/nada43/index.php/home with the permission of MOSPI after registering an account.

FUNDING SOURCES

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATIONS OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

ATHWANI, MEGHNA; Pandey, Rajeev (2021): Meghna Athwani.xlsx. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14241824.v1 .

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*, New York: John Wiley & Sons.

Erica. (2019, November 29). Introduction to the Fundamentals of Panel Data. *Data Analytics Blog*. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-panel-data/.

Greene, W.H. (2003). *Econometric Analysis*, Upper Saddle River, New Jersey: Prentice Hall.

Gujarati, D.N. (2003). *Basic Econometrics*, New York: McGraw Hill.

Hausman, J. (1978). *Specification Tests in Econometrics*. Econometrica, 46(6), 1251-1271. doi:10.2307/1913827.

Hsiao, C (2007). *Panel data analysis—advantages and challenges*. TEST 16, 1–22. https://doi.org/10.1007/s11749-007-0046-x.

Maddala, G.S. (2001). *Introduction to Econometrics*, New York: John Wiley & Sons.

Miller G. J., Yang K. (2008). *Handbook of Research Methods in Public Administration*. (pp. 574 – 594). New York: Taylor and Francis.

Ministry of Statistics and Programme Implementation. (n.d.). *Annual Survey of Industries (ASI)*. Retrieved February 27, 2021, from http://mospi.nic.in/sites/default/files/asi_results/all_about_asi.pdf?download=1.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Ranjan, C. (2020, March). Package 'nlcor': Compute Nonlinear Correlations. doi: http://dx.doi.org/10.13140/RG.2.2.33716.68480

Wooldridge, J.M. (2006). *Introductory Econometrics: A Modern Approach*, Mason, Ohio: Thomson South-Western.

Xu, Hua & Lee, Sock Hwan & Eom, Tae. (2007). *Introduction to Panel Data Analysis*. 10.1201/9781420013276.ch32.