

Project Report  
On

# **Explaining Productivity Differential Among Major States of India : A Panel Data Analysis**

Submitted to



Department of Statistics

**UNIVERSITY OF KALYANI**

Kalyani, Nadia, West Bengal- 741235, India

In partial fulfillment for

Master of Science in Statistics

Submitted by

**U Parimala**

Roll No : 96/STA No. 210032

REG No : 015567 of 2017-2018

Under the Supervision of

**CHIRANJIB NEOGI**

PROFESSOR

DEPARTMENT OF STATISTICS, UNIVERSITY OF KALYANI

# UNIVERSITY OF KALYANI

Kalyani, Nadia, West Bengal- 741235, India

## Department of Statistics



This is to certify that the project entitled as “ ***Explaining Productivity Differential Among Major States of India : A Panel Data Analysis*** ” has been submitted by Miss. *U Parimala* under my supervision and guidance. This project is submitted by her as a part of M.Sc. (4<sup>th</sup> Semester) educational curriculum, 2023.

*This project in part or in full is not submitted anywhere for publication.*

Place - Kalyani

Date : /08/2023

(Prof. Chiranjib Neogi)

## **ABSTRACT**

Discussions on the production function have always taken care of the attention of the economists. A production function is a mathematical function that shows how the quantity of output behaves as a function of the inputs used in production.

This study aims to estimate the Cobb-Douglas production function in different states of India by using 'capital', 'labour' and 'material' as the input factors and investigate the effect of economic input factors on the overall economic growth of the states. In this study linear panel data analysis techniques were used for 17 states of India with the data of 2008-2016 period.

The coefficients thus obtained from panel regression analysis were subjected to further analysis to find out which state wise factors result in output fluctuation in the states.

# **Contents**

S. No.		Page
1.	Introduction.....	3
2.	Theories Related to the Analysis	
	2.1 What is Panel Data.....	5
	2.2 Modelling Panel Data.....	6
	2.2.1 Pooled OLS Model.....	6
	2.2.2 Fixed Effect Model.....	7
	a. LSDV Model.....	7
	b. Fixed Effect within Model.....	8
	2.2.3 Random Effect Model.....	8
	2.3 Multiple Regression.....	9
	2.4 Cobb- Douglas production Function.....	10
3.	Tests / Diagnostics for Panel Data Analysis	
	3.1 Hausman Test.....	11
	3.2 Breush Pagan Test.....	12
	3.3 Poolability Test.....	12
4.	Materials and Methodology used.....	13
	4.1 Data Source.....	13
	4.2 Empirical Research Model.....	13
	4.2.1 Independent variable.....	14
	4.2.2 Dependent variable.....	14
	4.3 Methodology.....	15
5.	Exploratory Data analysis & Interpretation.....	16
	5.1 Preliminary Analysis.....	16
	5.2 Model Fitting & Analysis.....	21
	5.3 Test Results & Interpretation.....	28
6.	Conclusion.....	30
7.	Source Code.....	30
8.	Challenges and Limitations.....	31
9.	Acknowledgement.....	32
10.	References.....	33

# 1. INTRODUCTION

Productivity growth is the basis for improvements in real incomes and welfare. Slow productivity growth limits the rate at which real incomes can improve, and also increases the likelihood of conflicting demands concerning the distribution of income (Englander and Gurney, 1994). Measures of productivity growth and of productivity levels therefore constitute important economic indicators. The conversion of production factors into outputs is represented by mathematical expressions called the production function in the neoclassical tradition. In economic theory, the production function is simply described as the technical relationship between economic inputs and outputs (Cheng & Han, 2017). The most important condition for a country's economic growth is to increase production and to use production factors effectively to ensure this (M Songur & Saraç Elmas, 2017).

There are many different measures of productivity growth. The choice between them depends on the purpose of productivity measurement and, in many instances, on the availability of data. Broadly, productivity measures can be classified as single-factor productivity measures (relating a measure of output to a single measure of input) or multi-factor productivity measures (relating a measure of output to a bundle of inputs).<sup>1</sup> Another distinction, of particular relevance at the industry or firm level is between productivity measures that relate gross output to one or several inputs and those which use a value-added concept to capture movements of output. Table 1 uses these criteria to enumerate the main productivity measures.

Table 1. **Overview of the main productivity measures**

Type of output measure:	Type of input measure			
	Labour	Capital	Capital and labour	Capital, labour and intermediate inputs (energy, materials, services)
<b>Gross output</b>	Labour productivity (based on gross output)	Capital productivity (based on gross output)	Capital-labour MFP (based on gross output)	KLEMS multi-factor productivity
<b>Value-added</b>	Labour productivity (based on value-added)	Capital productivity (based on value added)	Capital-labour MFP (based on value-added)	—
	<b>Single factor productivity measures</b>		<b>Multi-factor productivity (MFP) measures</b>	

Mathematically, there are many forms of production functions in the literature. In literature, very often used production functions are linear production function, Cobb – Douglas production function, Constant Elasticity of Substitution production function (CES), Variable Elasticity Substitution production function (VES), Leontief Production

function, and Translog production function (Cheng & Han, 2017). In this study, Cobb – Douglas production function was used. 56 Hulya BAŞEGMEZ Cobb – Douglas production function has a lot of applications. One of the first studies in this field is Bronfenbrenner and Douglas's (1939) work. Later, Douglas analyzed the production function developed under different names in different years (Daly & Douglas, 1943; Daly, Olson, & Douglas, 1943; Gunn & Douglas, 1941, 1942). Regression analysis was used in all of these studies. Some of the studies using panel data analysis are as follows Wakelin (2001), Cantos, Gumbau-Albert, & Maudos (2005), Çermikli & Tokathog̃lu (2015), Inglesi-Lotz (2016) and Chikabwi, Chidoko, & Mudzingiri (2017). The popular production theory of Ferguson and Pfouts (1962) and Berndt and Christensen (1973) is advanced by Cameron and Schwartz (1980), Field and Grebenstein (1980).

There are many factors that have contributed to GDP, such that capital, labor, energy, optimal allocation of technology sources, innovations, etc. In this study, the Cobb – Douglas production function is used to scale the effect of capital, labour and material consumption on economic growth in the different states of india. The aim of this study is to estimate the Cobb – Douglas production function at the macro level for different states of India.

India is a developing country and it is frequently claimed in the current press and financial markets that capital is needed in developing countries, and external borrowing is mandatory, and there will be no development without foreign capital. Whether this claim is correct or not can be seen by estimating the output elasticity of capital.

In this context, data from 17 states of India were used in our study. In the study, we used three different inputs in the production function. These are capital (K), labor (L), and material consumption (E) inputs that must be involved in the production function. GVA (Gross Value Added) values which is a productivity indicator are used to represent the output. The formula by which the GVA is arrived is the following:

**GVA = value of output (quantity \* Price) - value of intermediary consumption**

Before the analysis data preparation was done by taking input factors (except workers) and output (GVA) at constant prices for removal of inflationary effects. Wholesale Price Index values were used to deflate GVA values and annual GFCF values at national level were used to deflate the capital value and material input (monetary units).

The analysis in the study were carried out with the help of linear panel data analysis techniques.

## **2. Theories Related to the Analysis**

### **What is Panel data :**

Before delving into the theory behind panel data, let us first briefly define the concepts of time series and cross-sectional data and illustrate how these two are connected with the theory of panel data.





Cross-sectional data refer to observations of many different individuals (subjects, objects) at a given time, each observation belonging to a different individual/entity,  $\{X_i$ , where,  $i = 1, 2, 3, \dots, N\}$ . The interest lies in modelling the distinction of single entities. For example, if we have the monthly sales figures of a product over multiple years want to study the average income of people in different cities, we could collect cross-sectional data by surveying individuals from various cities at the same time and recording their respective incomes.

On the other side, Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order. The interest lies in modelling distinction over time,  $\{X_t$ , where,  $t = 1, 2, 3, \dots, T\}$ . For example, daily closing prices of a company's stock over a period of time, hourly temperature measurements recorded over several months, monthly sales figures of a product over multiple years etc.

**Panel data**, also known as longitudinal or cross-sectional time series data, is a type of data that combines elements of both cross-sectional data and time series data. It involves observing multiple individuals, entities, or subjects over multiple time periods. Panel data analysis is particularly useful for studying the relationship between variables while controlling for individual-specific or time-specific effects. It allows researchers to account for individual heterogeneity and capture the dynamics of variables over time.

In contrast to cross-section data where we have observations on  $n$  subjects (entities), panel data has observations on  $n$  entities at  $T \geq 2$  time periods. This is denoted  $(X_{it}, Y_{it})$ ,  $i = 1, \dots, n$  and  $t = 1, \dots, T$  where the index  $i$  refers to the individual while  $t$  refers to the time period.

We can distinguish between balanced, unbalanced panels and also micro and macro panels.

-  A balanced panel is a dataset in which each panel member (i.e., person) is observed every year. Consequently, if a balanced panel contains  $N$  panel members and  $T$  periods, the number of observations ( $n$ ) in the dataset is necessarily  $n = N \times T$ .
-  An unbalanced panel is a dataset in which at least one panel member is not observed every period. Therefore, if an unbalanced panel contains  $N$  panel members and  $T$  periods, then the following strict inequality holds for the number of observations ( $n$ ) in the dataset:  $n < N \times T$ .
-  Macro panels are characterized by having a relatively large  $T$  and a relatively small  $N$ .
-  Micro panels, instead, usually cover a large set of units  $N$  for a relatively short number of periods  $T$ .

Panel data helps us to controls heterogeneity of cross-section units such as individuals, states, firms, countries etc., over time. Panel data estimation considers all crosssection units as heterogeneous. It helps us to get unbiased estimation. There are time-invariant and state-invariant variables which we observe or not.

## **Modelling Panel Data :**

In this section we describe the various panel data models used in our analysis and the additional assumptions that are needed to get consistent estimates of the coefficients using standard regression techniques.

### **Pooled OLS Model :**

The pooled OLS estimation is simply an OLS technique run on Panel data. Panel data contain the information of time and cross-sectional dimensions but pooled OLS, however, disregards this information of panel data. Therefore, all individual-specific effects are completely ignored and we have a common intercept for all individuals.

In a panel data set, we have repeated observations for each entity over time, and these observations are not independent of each other because they come from the same entity. Pooled OLS takes advantage of this structure by pooling all the data together and treating them as if they are from a single large cross-sectional dataset.

Consider the multiple linear regression model for individual  $i = 1, \dots, N$  who is observed at several time periods  $t = 1, \dots, T$  then the equation for Panel data for pooled OLS will be:

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

Where,  $y_{it}$  is the dependent variable for entity  $i$  at time  $t$ ,  $x'_{it}$  represents the transpose of the vector of independent variables for individual  $i$  at time  $t$ ,  $\alpha$  is the intercept,  $\beta$  is the vector of coefficients for the independent variables, and  $u_{it}$  is an idiosyncratic error term, which captures the unobservable factors that affect  $y_{it}$  but are not accounted for in the model.

The primary advantage of this method is its simplicity and ease of implementation. If the relationship between the variables is consistent across entities and time, pooled OLS can provide efficient estimates of the parameters. However, there are some important assumptions that need to be met for the estimates to be unbiased and efficient:

- **No individual-specific effects:** Pooled OLS assumes that there are no individual-specific effects or time-invariant individual-specific characteristics that influence the dependent variable. In other words, the individual entities have the same intercept and slope.
- **No serial correlation:** The error term should not exhibit serial correlation, meaning that the errors from one time period are not correlated with the errors from the previous or subsequent periods.



- **Homoscedasticity:** The error term should have constant variance across all observations.
- **No endogeneity:** The independent variable should be exogenous, i.e., it should not be correlated with the error term.

If these assumptions are violated, the Pooled OLS estimates may be biased or inefficient. In such cases, more sophisticated panel data methods like Fixed Effects (FE) and Random Effects (RE) models are often preferred, as they can better account for individual-specific effects and heterogeneity in the data.

### **The Fixed Effect Model :**

A fixed-effect model examines individual differences in intercepts, assuming the same slopes and constant variance across individuals (group and entity) or heterogeneity is fixed across the same panel. The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics...[like culture, religion, gender, race, etc].

Whenever there is a clear idea that individual characteristics of each entity or group affect the regressors or independent variables, use fixed effects. For example, macroeconomic data collected for most countries overtime. There might be a good reason to believe that countries' economic performance may be affected by their own internal characteristics: type of government, political environment, cultural characteristics, type of public policies, etc. Fixed-effects models are designed to study the causes of changes within a state or person or entity. This fixed effect model is estimated:

#### **a. Least squares dummy variable (LSDV) regression (OLS with a set of dummies) :**

The least squares dummy variable model (LSDV) uses dummy variables to estimate the fixed effect model. The model is estimated using OLS (Ordinary Least Squares) after transforming the data by creating dummy variables.

LSDV is widely used because it is relatively easy to estimate and interpret. The LSDV model allows each entity and each time period to have its own intercept but assumes a common slope (relationship) for all entities and time periods.

The LSDV works best when the panel data has relatively fewer cases and more time periods, as each dummy variable removes one degree of freedom from the model.

But it becomes more complex when the number of parameters to be estimated increases, i.e. when it includes many dummy variables. In this case, LSDV also loses degrees of freedom but returns less efficient estimators, and here it is better to go with the "within" model.

The general LSDV model is:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \sum_{j=1}^{16} \gamma_j D_{ji} + e_{it}$$

where,

- $Y_{it}$  is the GVA (real value)
- $\beta_1, \beta_2, \beta_3 \rightarrow$  parameters to be estimated
- $X_{1it} \rightarrow$  working capital,  $X_{2it} \rightarrow$  no. of workers and  $X_{3it} \rightarrow$  material input
- $\gamma_j$  is the dummy variable coefficient, where  $j=1,2,\dots,16$
- $D_{ji}$  is dummy variables for each cross-sectional unit excluding one, where  $i=1,2,\dots,17$
- $e_{it} \rightarrow$  error term

If the null hypothesis is rejected, we may conclude that the fixed effect model is better than the pooled OLS model.

#### **b. Fixed Effect Within Model :**

The fixed effects within model, also known as the Within Estimator model is a transformation of the original data, where the individual-specific means (fixed effects) are removed from each observation, leaving only the deviations from the individual-specific means from the original data. In other words, the model subtracts the individual-specific means from each individual's observations, and then it estimates the relationship based on the deviations from these means over time. This approach eliminates the individual-specific heterogeneity and focuses on the time variation within each individual. The fixed effects within model is specified as follows:

$$Y_{it} - \text{mean}(Y_i) = \beta(X_{it} - \text{mean}(X_i)) + e_{it}$$

Where,

- $\text{mean}(Y_i)$  is the mean value of the dependent variable for individual  $i$  across all time periods.
- $\text{mean}(X_i)$  is the mean value of the independent variables for individual  $i$  across all time periods.
- The other variables have the same meaning as in the LSDV model.

It is important to note that the fixed effects within model eliminates the individual-specific effects, but it cannot estimate the effect of time-invariant variables (since they are cancelled out in the differencing process).

#### **The Random Effect Model :**

In a random effect model, individual-specific effects are treated as random variables with their own distribution. These random effects are assumed to be uncorrelated with

the independent variables, and their presence allows for the inclusion of time-invariant variables in the model.

In Random effect, the heterogeneity varies across the individual panel. A random-effect model reduces the number of parameters to be estimated but will produce inconsistent estimates when the individual-specific random effect is correlated with regressors. So, a random-effects model assumes that individual effect is not correlated to any of the regressors, and it is estimated as error variances specific to groups or times.

A random effect is generally represented by

$$Y_{it} = a + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_i + e_{it}$$

Where,

- $\beta_1, \beta_2, \beta_3 \rightarrow$  parameters to be estimated
- $X_{1it} \rightarrow$  working capital,  $X_{2it} \rightarrow$  no. of workers and  $X_{3it} \rightarrow$  material input, where  $i=1,2,\dots,17$  ;  $j = 1,2,\dots,16$  and  $t=1,2,\dots,8$
- $u_i \rightarrow$  white noise
- $e_{it} \rightarrow$  error term

there are K regressors in addition to the constant term. The component  $u_i$  is the random disturbance characterizing the  $i$ th observation and is constant through time.

***Some Assumptions :***

$$\begin{aligned} E(e_{it}) &= E(u_i) = 0 \\ E(e_{it}^2) &= \sigma_e^2 \\ E(e_{it}, u_j) &= 0, \text{ for all } i, j, t \\ E(u_i, u_j) &= 0, \text{ if } i \neq j \\ E(e_{it}, e_{js}) &= 0, \text{ if } t \neq s \text{ or } i \neq j \end{aligned}$$

The random effects model is appropriate when the cross-sectional units are randomly selected from a large population. If a variance structure among groups is known, the random effects model is estimated by the generalized least squares (GLS) method, which takes into account the correlation between the error terms for each entity. On the other hand, if the variance structure is not known, the feasible generalized least squares (FGLS) method is appropriate to estimate the variance structure.

The random effect model is appropriate when we are interested in analyzing the effects of both time-varying and time-invariant variables, while accounting for unobserved heterogeneity among individuals.

## **Multiple Regression :**

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modelled as a function of several independent variables with corresponding coefficients, along with the constant term. Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.

The multiple regression equation explained above takes the following form:

$$y = b_1x_1 + b_2x_2 + \dots + b_nx_n + c$$

Here,  $b_i$ 's ( $i=1,2,\dots,n$ ) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes.

- There should be proper specification of the model in multiple regression. This means that only relevant variables must be included in the model and the model should be reliable.
- Linearity must be assumed; the model should be linear in nature.
- Normality must be assumed in multiple regression. This means that in multiple regression, variables must have normal distribution.
- Homoscedasticity must be assumed; the variance is constant across all levels of the predicted variable.

### **Cobb-Douglas Production Function :**

Cobb and Douglas (1928) introduced the most famous and known production function in the form,

$$Q = f(K, L, M) = AK^\alpha L^\beta M^\gamma \quad (2.1)$$

where  $Q$  is a total production which means the value of all goods produced in a year,  $K$  is the capital input which represents by the total investment in fixed assets,  $L$  is the labour inputs which is the total number of person or hours worked in a year) and  $A$  is a positive constant which means total factor productivity (Cobb & Douglas, 1928). The parameters  $\alpha, \beta, \gamma$  show the output elasticities to capital and labour and materials respectively. Output elasticity measures the sensitivity of output to a change in the levels of labour, capital and materials used in production. Cobb – Douglas production function allows the change of the size of the inputs affected by factor price changes. One of the limitations of this production model is that it uses three input factors to explain production (Liao, Wu, & Xu, 2010).

The general form of production function is described by

$$f: D \rightarrow \mathbb{R}_+, D \in \mathbb{R}_+^n$$

and,

$$\mathbb{D} = (\mathbb{D}_1, \mathbb{D}_2, \dots, \mathbb{D}_n) \quad (2.2)$$

Where  $\mathbb{D}_1, \mathbb{D}_2, \dots, \mathbb{D}_n$  are inputs and  $Q$  is production level. A production function with input factors is called  $h$  –homogeneous,  $h > 0$ , if

$$(\mathbb{D} \mathbb{D}_1, \mathbb{D}_2, \dots, \mathbb{D} \mathbb{D}_n) = \mathbb{D}^h \mathbb{D}(\mathbb{D}_1, \mathbb{D}_2, \dots,$$

$\mathbb{D}_n)$  Where  $\mathbb{D} \in (0, \infty)$  (Onalan & Basegmez, 2018 ).

Now, in our Cobb Douglas production function with three input factors (Capital, Labour and Materials) we will encounter either of these three cases:

(1)  $\alpha + \beta + \gamma > 1$

Which shows increasing returns to scale. With increasing returns to scale, a proportional increase in all inputs will increase output by more than the proportional constant.

(2)  $\alpha + \beta + \gamma < 1$

Which shows diminishing returns to scale. With decreasing returns to scale, a proportional increase in all inputs will increase output by less than the proportional constant.

(3)  $\alpha + \beta + \gamma = 1$

Which shows constant returns to scale. With constant returns to scale, output increases in exactly the same proportion as the factors of production.

### 3. Tests / Diagnostics for Panel Data Analysis

#### Hausman Test :

The Hausman test is the standard procedure used in empirical panel data analysis to help choose between two different econometric models: the fixed effects model and the random effects model. The choice of the appropriate model is crucial as it can affect the validity and efficiency of the statistical inferences drawn from the analysis.

How do we decide whether to use the fixed effects model or the random effects model?

The Hausman test helps with this decision. The null hypothesis of the test is that the preferred model is the random effects model (i.e., the individual-specific effects are uncorrelated with the independent variables). The alternative hypothesis is that the preferred model is the fixed effects model (i.e., the individual-specific effects are correlated with the independent variables).

If the Hausman test rejects the null hypothesis, it suggests that the fixed effects model is preferred because there is evidence that the individual-specific effects are correlated with the independent variables. On the other hand, if the test fails to reject the null hypothesis, the random effects model is considered appropriate.

The test statistic can be calculated is given as follows:

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})' \left( \frac{V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE})}{V(\hat{\beta}_{FE})} \right) (\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

Where, RE – random effect and FE – fixed effect.

Here,  $\hat{\beta}_{RE}$  and  $\hat{\beta}_{FE}$  are the vector of parameter estimates of random effect and fixed effect, respectively. Under the null hypothesis, this statistic has asymptotically the chi-squared distribution with the number of degrees of freedom equal to the rank of the matrix :

$$V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE})$$

### **The Breush-Pagan Test :**

The Breusch-Pagan (BP) test, also known as the Breusch-Pagan-Godfrey test, is a statistical test used in econometrics to check for heteroskedasticity in panel data analysis. Heteroskedasticity occurs when the variance of the errors in a regression model is not constant across all levels of the independent variables, which violates one of the assumptions of ordinary least squares (OLS) regression.

The null hypothesis of the Breusch-Pagan test is that there is homoscedasticity (constant variance) in the errors, while the alternative hypothesis suggests the presence of heteroscedasticity.

In panel data analysis, we often use fixed-effects or random-effects models to account for unobserved heterogeneity across individuals or entities over time. However, it is essential to verify whether heteroskedasticity is present, as it can lead to biased and inefficient estimates.

The Breusch-Pagan test is based on the idea that if heteroskedasticity is present, the variance of the residuals will be related to certain explanatory variables included in the model. The test evaluates the significance of the relationship between the squared residuals (squared OLS residuals) and the explanatory variables.

If the test yields a statistically significant result (i.e., rejecting the null hypothesis), it indicates the presence of heteroskedasticity in the original regression model. In this case, we might need to consider using heteroskedasticity-robust standard errors or other appropriate methods to obtain reliable parameter estimates.

### **The Poolability Test :**

A Poolability test is conducted to determine whether the data collected from different entities (e.g., individuals, firms, countries) in a panel dataset can be pooled together to estimate a common set of parameters. The main idea behind the poolability test is to assess whether the fixed effects or individual-specific effects are constant across all entities in the dataset. If the pooled model is a good fit, it suggests that the individual-specific effects are not significantly different across entities, and a pooled analysis is appropriate.

Panel data models often include fixed effects or individual-specific effects to account for unobserved heterogeneity among the entities. These effects capture entity-specific characteristics that remain constant over time. However, in some cases, these effects might vary significantly across different entities, and in such situations, a pooled model might not be appropriate. The two common types of poolability tests are:

- (i) **F-test (Chow test):** The F-test can be used to determine whether the coefficients of the independent variables are the same across different groups or subsets of the data. The null hypothesis of the Chow test is that the coefficients are the same across entities, implying that the data is poolable.
- (ii) **Hausman test:** It compares the efficiency of the estimates from a fixed effects model and a random effects model. The random effects model assumes that the

individual-specific effects are uncorrelated with the independent variables, while the fixed effects model allows for correlation between the individual-specific effects and the independent variables. If the test statistic is significant, it suggests that the random effects model is not appropriate, and the fixed effects model should be used.

Here, we have used F-test (Chow test) for time specific and individual specific effects to determine whether the coefficients of the independent variables are the same across different groups or subsets of the data.

#### **4. Materials and Methodology used**

This section describes in detail the data sources and the methodology of construction of the panel data set and all the variables included in the study.

##### **Data Source :**

In this study annual data pertaining to 17 different states of India over the period 2008-2016 was taken. The values of the variables are taken with respect industries whose data was available for all the states we have considered in our dataset in order to make our panel balanced.

Dataset in the study was taken from “ KLEMS India database ” (www.rbi.org.in). For the explain of productivity, we have selected the 17 different states of India viz. Andhra Pradesh, Goa, Gujarat, Haryana, Himachal Pradesh, Jharkhand, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Odisha, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh, Uttarakhand and West Bengal.

From the original data only the variables workers capital (real value), number of workers employed, material input (real value) and GVA (real value) were included in the panel dataset, which we are going to analyse throughout the rest of the study.

##### **Dataset :**

<https://docs.google.com/spreadsheets/d/1NdbzL6BJQEb2Z25zT9ZJwsANpvXCPcD0/edit?usp=sharing&ouid=116451843067148312587&rtpof=true&sd=true>

##### **Empirical Research Model :**

The Cobb-Douglas production function based on three factors of production namely ‘Material Input’ (M), ‘Labour input’ (L) and ‘Capital Input’ (K) (in our case we consider the investment capital) is given as follows:

$$Q = AK^{Q_1}L^{Q_2}M^{Q_3}$$

Where, Q= GVA (real value), K= Investment Capital, M= Material Input (no. of workers), A=Total factor productivity of the states and  $\beta_1, \beta_2, \beta_3$  = parameters to be estimated.

This production function is expressed linearly by taking natural logarithm of both sides of equation:

$$Y_{it} = Q_0 + Q_1X_{1it} + Q_2X_{2it} + Q_3X_{3it} + Q_4Z_i + u_{it}$$

Where,  $Z_i$ =unobserved time invariant heterogeneities across entities,  $i = 1, 2, \dots, 17$ .

$\beta_0 = \ln A$ ;  $X_{1it} = \ln K$ ;  $X_{2it} = \ln L$ ;  $X_{3it} = \ln M$ ;  $u_{it}$  = white noise;

Since, we have panel data at our disposal the model obtained for the production function is given in linear form is as follows:

$$Y_{it} = \alpha_i + Q_1 X_{1it} + Q_2 X_{2it} + Q_3 X_{3it} + u_{it}$$

by letting,  $\alpha_i = \beta_0 + \beta_4 Z_i$ , we obtain the model.

**Independent Variable** : The independent variables considered in our study are Working Capital (WK), labour (L) and Materials (M). In order to adjust for the effects of inflation we take the real values of the independent variables to measure the actual change (and not just a change seen due to the effects of inflation). Data pertaining to Current and constant GFCF (Gross Fixed Capital Formation) values at the national level (for the period 2008-16) was taken since the figures of GFCF were not available at state level. This data was taken from Table 13 of the 'Handbook of Indian Economy' section of RBI's website. Current and constant GFCF values for the period 2008-12 were taken with respect to base year 2004-05 and the values for the period 2012-12 were taken with respect to base year 2011-12.

Now these values (current and constant) are used to construct the deflator using the following formula –

$$\text{Deflator} = \frac{\text{Current Value}}{\text{Constant value}} \quad (3.1a)$$

The deflators thus constructed are used to obtain the real values of the working capital and material respectively using the formula –

$$\text{Real Value} = \frac{\text{Nominal value}}{\text{Deflator}} \quad (3.1b)$$

**Dependent Variable** : In our dataset although we have both the gross output and GVA values of the 17 different states of India over the period 2008-16, for the purpose of our study we consider the value added measure of output rather than the gross output. The former measure excludes intermediate inputs (materials, energy and services used up in the process of production) while the latter measure includes those inputs.

However, the difference between the two concepts of productivity growth is less pronounced at the aggregate (or national) level than it is at the sectoral or industry level. At the aggregate level, gross output-based and value-added based measures are close, only differing to the extent that intermediate inputs are sourced from imports. In proportional terms, this tends to be low. At the industry or sector level, however, intermediate usage tends to be a



much higher proportion of gross output. This results in greater variation between the two measures.

Moreover, the value-added approach has considerable advantages because it is a simple measure that ignores the difficulties of dealing with inter-industry and intra-industry flows of goods and services. Intermediate inputs are simply excluded by the value added measure.

Nominal GVA values are converted to real values to show the actual changes in GVA over time rather than denote a rising trend in prices due to inflationary effect.

Process of conversion is the same as was explained in the section of Independent Variables.

## **Methodology :**

This study analyses whether there is a significant difference in the productivity of the states of India and which are factors responsible for the variation of the productivity among states.

The estimation of panel data is done in three main methods, (i) LSDV or Least Square Dummy Variable, (ii) Fixed Effect methods, (iii) Random Effect method and (iv) Pooled OLS. We have used these methods and presented the results. After performing necessary tests on the model, we have selected our final model.

Hausman Specification test was conducted to determine whether the model is suitable for random or fixed effects model.

In this context, the Hausman test can be considered a test between the null hypothesis  $H_0$ : There is a correlation between independent variables and the error term.

Versus the alternative hypothesis  $H_1$ : there is no correlation between the independent variables and unit effect. If the null hypothesis  $H_0$  is rejected the result is that the fixed effects model should be used, and if not, the random effects model should be used (Hausman, 1978).

In case Hausman test shows evidence in favour of random effects model we use the Breusch Pagan test to determine whether random effects are significant or not i.e, testing the null hypothesis  $H_0$ : variance (white noise) = 0

Against the alternative hypothesis  $H_1$ : variance of white noise is not equal to 0.

If Null hypothesis is not rejected it implies that a pooled cross-sectional model is sufficient to model the variation in data.

Now if Hausman test indicated the presence of fixed effects we perform the Poolability test to determine whether there is significant cross-sectional effect or significant cross temporal effect.

In the absence of these fixed effects (individual or time) we may run an OLS regression model with common intercept constant across entities and time.

If we detect a significant cross sectional (statewise) effect on the GVA fluctuation, from the analysis performed previously we might be interested in knowing which macro-economic factors (statewise) result in the variation.

For that purpose, we conduct an OLS Regression where fixed effect coefficients obtained from previous LSDV model serves as the productivity indicator (dependent variable). Capital index, skill and PCNSDP (per capita net state domestic product) become our independent variables.

## **5. Exploratory Data Analysis & Interpretation**

### **Preliminary analysis :**

#### **Data Description :**

<b>yearcode</b>	<b>stateid</b>	<b>real_fk</b>	<b>realIK</b>	<b>realmaterial</b>	<b>realgva</b>	<b>workers</b>	<b>pcnsdp</b>
2008-09	Andhra Pradesh	27559.924	158073.2	7165.19987	644.94	835322	33733
2009-10	Andhra Pradesh	29674.322	146949	1808.490006	495.76	607962	35677
2010-11	Andhra Pradesh	44689.783	229941.4	4383.610044	418.53	974251	37708
2011-12	Andhra Pradesh	68421.54	389313.7	7894.010085	424.42	1043077	38556
2012-13	Andhra Pradesh	32126.623	189984.1	15273.56958	378.94	362321	39645

Table-1 : Sample of the panel dataset

After 'loading' and "preprocessing" the data, I have seen that the dataset contains 137 rows and 8 columns.

## Summary Statistics :

Let us see some summary statistics of the data.

**Table-2 : Descriptive statistics results of the GVA values for the 17 States is presented in the result below :**

States	Sum	Mean	Max	Min	SD	Jarque Bera p values
Andhra Pradesh	3249.49	406.1862	644.94	280.59	113.6245	0.381110324
Goa	1030.27	128.7837	181.94	94.49	26.00143	0.57152841
Gujarat	2403.36	300.42	440	194.96	76.51526	0.613443985
Haryana	1856.18	232.0225	308.17	194.32	39.57514	0.400730749
Himachal Pradesh	1597	199.625	307.09	146.1	52.82287	0.314746714
Jharkhand	2422.55	302.8188	542.39	174.1	113.0933	0.350445803
Karnataka	2340.58	292.5725	385.92	217.99	63.0537	0.61953032
Kerala	4084.22	510.5275	800.82	322.14	133.9211	0.22040781
Madhya Pradesh	2561.27	320.1588	451.69	192.79	94.06034	0.536136945
Maharashtra	1914.79	239.3487	331.87	124.12	63.69129	0.969527075
Odisha	2144.08	268.01	397.54	111.49	81.01284	0.742149128
Punjab	2250.24	281.28	434.51	103.65	92.4266	0.808892487
Rajasthan	1812.91	226.6138	368.36	127.23	72.14505	0.73562973
Tamil Nadu	2702.63	337.8288	592.16	123.1	120.1868	0.170196594
Uttar Pradesh	2400.65	300.0813	427.39	154.63	78.07105	0.940637453
Uttarakhand	1891.87	236.4837	436.38	122.33	98.98846	0.306133261
West Bengal	2434.38	304.2975	433.35	179.86	76.67257	0.938722035

The descriptive statistics results for the 17 states are presented in Table-2 and depicted in Figure-1. The descriptive measures output shows that Goa has the minimum value which is 26.00143 among all the states. The p-value associated with the Jarque-Bera test statistic tells us the probability of observing the given test statistic (or an even more extreme one) under the assumption that the data is normally distributed. If the p-value is high (greater than 0.05), we fail to reject the null hypothesis and we may assume that the data follows a normal distribution for our analysis.

P-values from Jarque-Bera test shows that GVA values for all the states follow Normal Distribution.

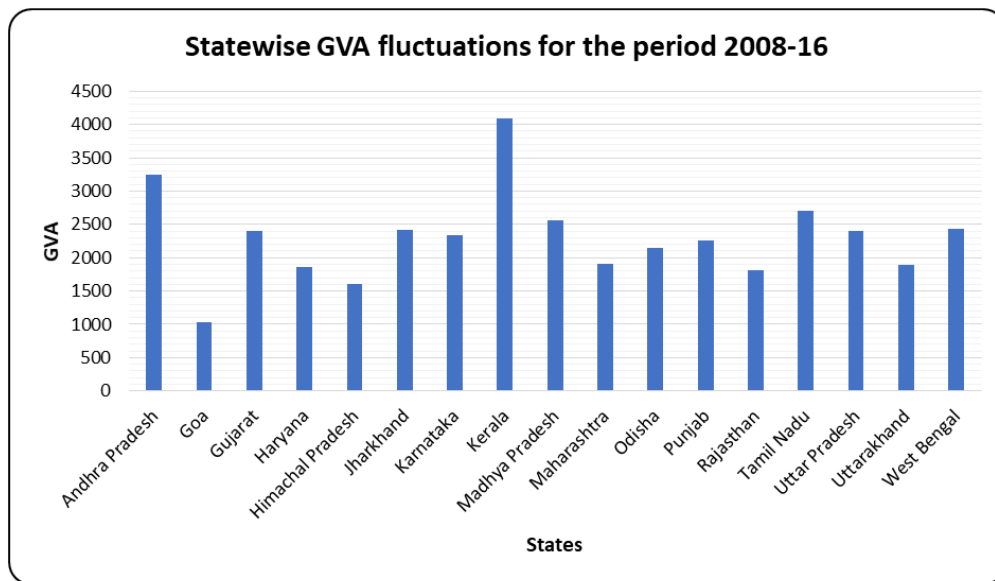


Figure 1 : Statewise GVA fluctuation

By seeing the figure-1 which is a column chart, we can say that productivity level is maximum for the state Kerala followed by Andhra Pradesh. Minimum productivity level is recorded in Goa.

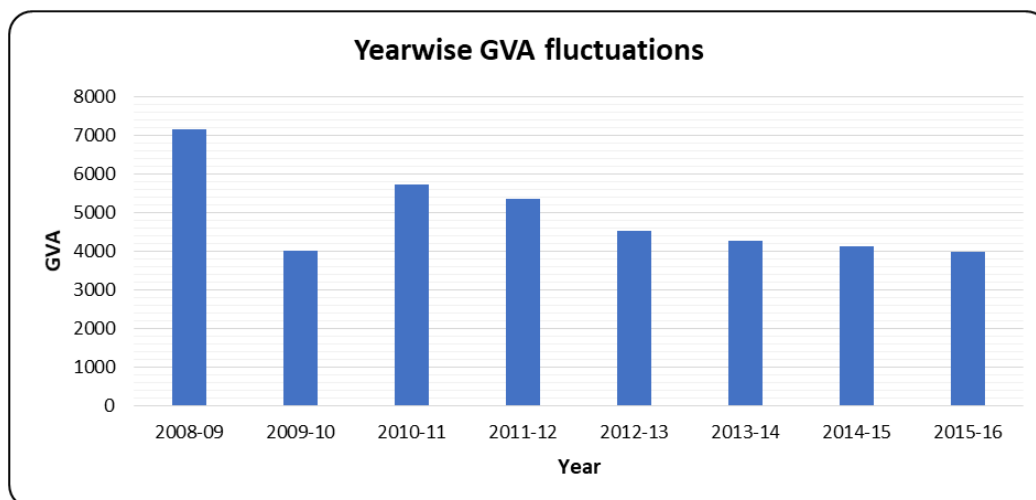


Figure 2 : Yearwise GVA fluctuation

In the column chart, we analyse the GVA fluctuations over the period 2008-16, i.e. by considering the GVA values of all states as a single entity we obtain a times series which shows us GVA variations over the years.

Figure-2 depicts that highest productivity level was recorded during year 2008-09 and after 2010-11, it is slightly decreasing upto 2015-16.

From the above two diagrams we concluded that heterogeneity across years is more than in the states.

### **Scatter Plots of GVA vs The factors of Productions :**

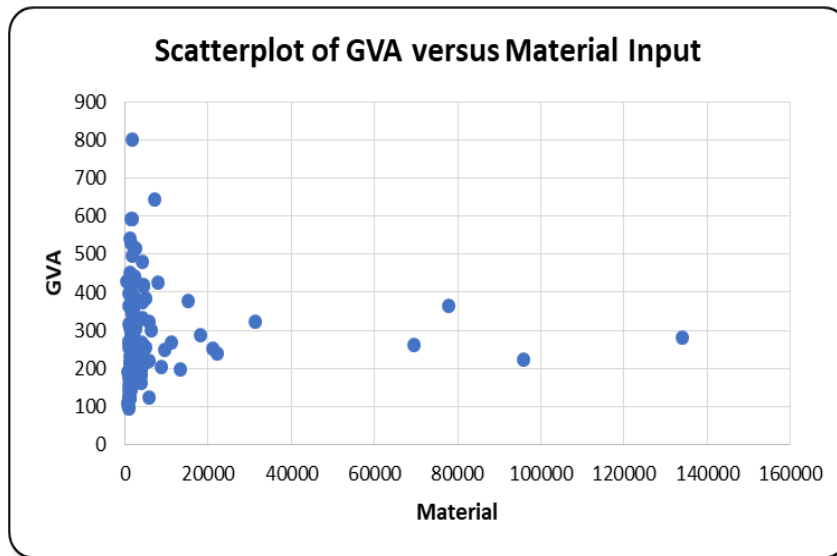


Figure 3 : Scatterplot of GVA vs Material Input

In figure-3, the scatter plot shows that even as the material input remains constant GVA keeps on increasing implying that changes in Material input does not cause a significant change in the productivity levels.

The presence of a few oscillating outliers in the right-hand side indicates that there might be some correlation between 'GVA' and 'Material' inputs. Whether the association is significant or not remains to be seen from further analysis.

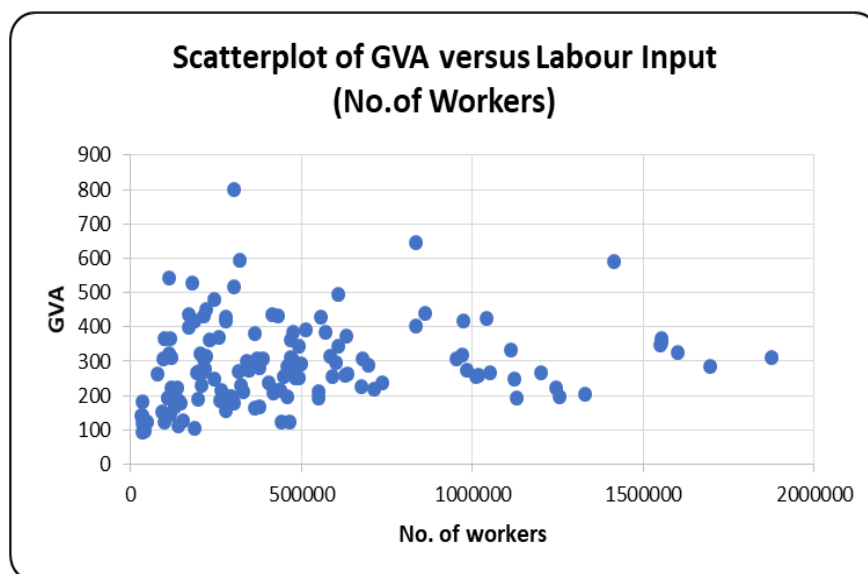


Figure 4 : Scatterplot of GVA vs Labour Input

In figure-4, the above scatterplot shows that number of workers employed (Labour Input) is positively correlated with the productivity level, i.e: as the Labour Input increases GVA or production level also increases. Although there are some points in the plot which suggests that as labour input increases GVA level remain constant. Further analysis will show us whether the changes in labour input are indeed significant or not.

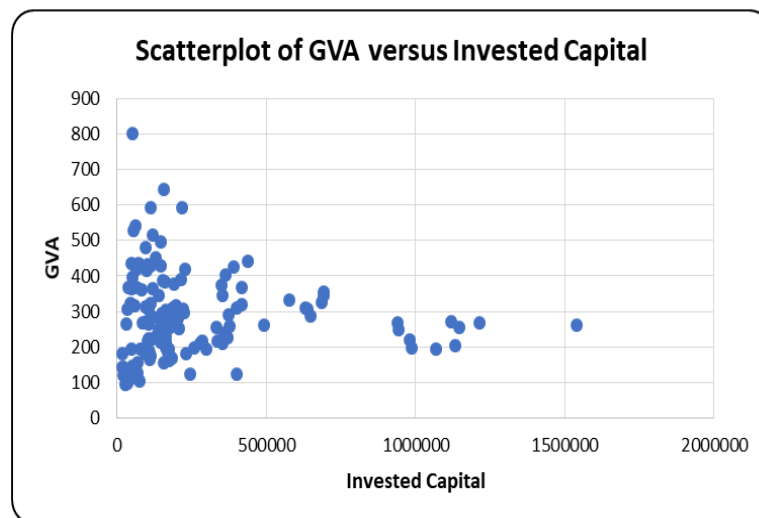


Figure 5 : Scatterplot of GVA vs Invested Capital

In figure-5, the above scatterplot roughly shows that GVA and invested capital are negatively correlated with each other, i.e. if the invested capital increase GVA decreases. However, even though the dominant trend reveals an inverse relationship between 'GVA' and 'invested capital', there are more than a few points in the bottom left corner of the plot which suggests that increase in invested capital can also cause an increase in the GVA.

## Model Fitting & Analysis

### LEAST SQUARE DUMMY VARIABLE (LSDV) REGRESSION MODEL : (with reference to Table No: 1)

<b>Table No. 1 : LSDV MODEL</b>				
	<u>Estimate</u>	<u>Std. Error</u>	<u>t-value</u>	<u>Pr(&gt; t )</u>
$X_1$ (Capital Input)	-0.5993***	0.061	-9.866	0.000
$X_2$ (Workers Input)	0.9746***	0.087	11.161	0.000
$X_3$ (Material Input)	0.0310	0.023	1.363	0.176
Andhra Pradesh	2.9087***	0.952	3.057	0.003
Goa	3.2064***	0.786	4.080	0.000
Gujarat	2.9122***	1.003	2.905	0.004
Haryana	2.6200***	0.939	2.789	0.006
Himachal Pradesh	3.1481***	0.851	3.698	0.000
Jharkhand	3.5297***	0.861	4.100	0.000
Karnataka	2.6781***	0.961	2.786	0.006
Kerala	3.4239***	0.893	3.835	0.000
Madhya Pradesh	3.2810***	0.886	3.701	0.000
Maharashtra	2.5259**	1.004	2.517	0.013
Odisha	2.9806***	0.881	3.384	0.001
Punjab	2.5797***	0.926	2.787	0.006
Rajasthan	2.6433***	0.910	2.904	0.004
Tamil Nadu	2.3579**	1.007	2.340	0.021
Uttar Pradesh	2.7954***	0.954	2.930	0.004
Uttarakhand	2.9057***	0.896	3.242	0.002
West Bengal	2.7250***	0.930	2.930	0.004
<b>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</b>				

<b>Table No. 1 : <u>LSDV Model</u></b>			
<b>Dep. Variable:</b>	Yit	<b>R-squared:</b>	0.999
<b>Model:</b>	LSDV	<b>Adj. R-squared:</b>	0.998
<b>No. Observations:</b>	136	<b>F-statistic:</b>	4428.095***
<b>Df Residuals:</b>	116	<b>Df Model:</b>	20
<b>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</b>			

## Interpretation:

1. The LSDV model accounts for heterogeneity by allowing different intercepts, one for each state in the data. It does this by using dummy variables. Differences in intercept capture the unique characteristic of the state, where the term varies across state but it is time invariant.  
Here each STATE is regarded as a dummy variable for estimation.
2. P-Value: Indicates the significance level of t-statistics. The variable which will be less than 0.05 will be considered statistically significant.  
And In the LSDV model we got all the coefficients of independent variables as statistically significant, then we can say that the variation of independent variables affects the variation of dependent variables, but IK coefficient is in negative which simply means investment capital is negatively correlated with the productivity of industries.  
Here the independent variable is  $\log(\text{gva})$  and the independent variables are also in  $\log$ .
3. Adjusted R-Squared: 0.9985  
In general, the larger the R-squared value of a regression model the better the predictor variables are able to predict the value of the response variable. In this case, we can say that the variation of independent variables explain 98% of the variation of dependent variable.
4. F-statistics also shows the overall goodness of fit.  
Each value of the coefficient of state (i.e, coefficient of dummy variables) indicates the TOTAL FACTOR PRODUCTIVITY of the Industries of a state.



### **FIXED EFFECT WITHIN MODEL** (with reference to table-2):

<b>TABLE No. 2 : FIXED EFFECT MODEL</b>				
	<b>Estimate</b>	<b>Std. Error</b>	<b>t-value</b>	<b>Pr(&gt; t )</b>
<b>Capital Input</b>	-0.599***	0.061	-9.866	0.000
<b>Workers Input</b>	0.975***	0.087	11.161	0.000
<b>Material Input</b>	0.031	0.023	1.363	0.176
<b>Note: * p&lt;0.1; ** p&lt;0.05; *** p&lt;0.01</b>				

<b>Table No. 2 : FIXED EFFECT MODEL</b>			
<b>Dep. Variable:</b>	Yit	<b>R-squared:</b>	0.544
<b>Model:</b>	Fixed Effect	<b>Adj. R-squared:</b>	0.469
<b>No. of Observations:</b>	136	<b>F-statistic:</b>	46.138***
<b>Df Residuals:</b>	116	<b>Df Model:</b>	3
<b>Note: * p&lt;0.1; ** p&lt;0.05; *** p&lt;0.01</b>			

This model of estimation uses variation within each panel or entity instead of many dummies.

1. The test results reveal the Cross-section F-statistic's value is significant at 1 % level of significance indicates that the presence of fixed effects and it is different from one state to another.
2. The individual effect of IK, Workers are statistically significant as the p-value is less than 5%, but for materials it is not shown as significant, and here also IK is showing negative correlation with the dependent variable.
3. Adjusted R2: 0.46937  
In this case, we can see that the variation of independent variables explains only 47% of the variation of dependent variables. which is very less compared to LSDV model.

### **ONE WAY (TIME FIXED) MODEL** (with reference of table-3):

<b>Table No. 3 : One Way (Time Fixed) Model</b>				
<b><u>Coefficients</u></b>	<b><u>Estimate</u></b>	<b><u>Std. Error</u></b>	<b><u>t-value</u></b>	<b><u>Pr(&gt; t )</u></b>
<b><math>X_1</math>(Capital Input)</b>	-0.4971139	0.1412050	-3.5205	0.0006299 ***
<b><math>X_2</math>(Workers Input)</b>	0.7881488	0.0987902	7.9780	1.61e-12 ***
<b><math>X_3</math>(Material Input)</b>	0.0511941	0.0204594	2.5022	0.0138263 *
<b>2009-10</b>	-0.3013688	0.0800866	-3.7630	0.0002721 ***
<b>2010-11</b>	-0.1133705	0.0800869	-1.4156	0.1597456
<b>2011-12</b>	0.0038983	0.1373477	0.0284	0.9774091
<b>2012-13</b>	-0.1078621	0.1347090	-0.8007	0.4250446
<b>2013-14</b>	-0.1849435	0.1274459	-1.4512	0.1496092
<b>2014-15</b>	-0.2618244	0.1349513	-1.9401	0.0549461 ‘
<b>2015-16</b>	-0.2986733	0.1395753	-2.1399	0.0345956 *
<b>Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1</b>				

<b>Table No. 3 : One Way (Time Fixed) Model</b>			
<b>TSS (Total Sum of Squares):</b>	12.205	<b>R-squared:</b>	0.66995
<b>Residual Sum of Squares:</b>	4.0281	<b>Adj. R-squared:</b>	0.59123
<b>p-value:</b>	< 2.22e-16	<b>F-statistic:</b>	22.1258
<b>Df Residuals:</b>	109	<b>Df Model:</b>	10

#### **Interpretation:**

The individual-specific error component, captures any unobserved effects that are different across individuals but fixed across time. Variable of interest which measures an intercept that is constant across all individuals and time periods.

## **RANDOM EFFECT MODEL** (with reference of table-4):

<b>TABLE No. 4 : RANDOM EFFECT MODEL</b>				
	<b>Estimate</b>	<b>Std. Error</b>	<b>t-value</b>	<b>Pr(&gt; t )</b>
<b>Intercept</b>	4.699***	0.660	7.117	0.000
<b>Capital Input</b>	-0.590***	0.060	-9.807	0.000
<b>Workers Input</b>	0.816***	0.072	11.374	0.000
<b>Material Input</b>	0.037	0.023	1.608	0.108
<i>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</i>				

<b>Table No. 4 : RANDOM EFFECT MODEL</b>			
<b>Dep. Variable:</b>	Yit	<b>R-squared:</b>	0.502
<b>No. of Observations:</b>	136	<b>Adj. R-squared:</b>	0.490
<b>Df Model:</b>	3	<b>F-statistic:</b>	132.948***
<i>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</i>			

In Random effect model we have only one intercept.

### **1. R-squared : 0.502**

The variation of independent variables explains only 50% of the variation of dependent variables, very less but better than Fixed effect within model.

- In this model also capital input is negativity significant and workers output is positively significant. Whereas material input is showing no significance.
- The test results reveal the Cross-section F -statistic's value is significant at 1 % level of significance indicating the presence of random effects and it is different from one variable to another.

### **POOLED OLS MODEL** (with reference of table: 5):

TABLE NO 5: POOLED OLS MODEL				
	Estimate	Std. Error	t-value	Pr(> t )
<b>(Intercept)</b>	5.549***	0.466	11.898	0.000
<b>Capital Input</b>	-0.553***	0.065	-8.566	0.000
<b>Workers Input</b>	0.677***	0.065	10.430	0.000
<b>Material Input</b>	0.078***	0.028	2.779	0.007
<i>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</i>				

Table No. 5: POOLED OLS MODEL			
<b>Dep. Variable:</b>	Yit	<b>R-squared:</b>	0.471
<b>Model:</b>	Pooled OLS	<b>Adj. R-squared:</b>	0.459
<b>No. of Observations:</b>	136	<b>F-statistic:</b>	39.177***
<b>Df Residuals:</b>	132	<b>Df Model:</b>	3
<i>Note: *p&lt;0.1; **p&lt;0.05; ***p&lt;0.01</i>			

#### **Interpretation of Pooled OLS model :**

1. The test results reveal the Cross-section F-statistic's value is significant at 1 % level of significance indicates that the presence significant difference in the variables.
2. The individual effect of IK, Workers and materials is also statistically significant as the p-value is less than 5%, here also IK is showing negative correlation with the dependent variable.
3. R-squared: 0.471

In this case, we can see that the variation of independent variables explains only 47% of the variation of dependent variables. which is very less compared to LSDV model.

## OLS REGRESSION MODEL (with reference to table-6):

TABLE No. 6: OLS REGRESSION MODEL				
<u>Coefficients</u>	<u>Estimate</u>	<u>Std. Error</u>	<u>t-value</u>	<u>Pr(&gt; t )</u>
<b>labshare</b>	-0.0919681	0.0199059	-4.62	0.0000
<b>pcnsdp</b>	0.0017357	0.003295	0.53	0.607
<b>skill</b>	2.815204	1.933094	1.46	0.169
<b>_cons</b>	3.2249	0.402667	8.01	0.000
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table No. 6: OLS REGRESSION MODEL			
<b>Model:</b>	OLS	<b>R-squared:</b>	0.6247
<b>Prob &gt; F:</b>	0.0043	<b>Adj. R-squared:</b>	0.5380
<b>Df Residuals:</b>	13	<b>F-statistic:</b>	7.21
<b>Df Model:</b>	3	<b>Root MSE:</b>	0.29026

### Interpretation:

The ordinary least squares (OLS) method is a linear regression technique that is used to estimate the unknown parameters in a model.

We obtain productivities using LSDV and Fixed model. And now, we got that productivity is different for different states i.e. there are variations among the states in productivity.

We could like to understand the causes of variations. There may be some independent variables. So, we performed OLS regression model. We have considered some of the independent variables as '**labshare**', '**pcnsdp**' and '**skill**'. And we could see that '**skill**' plays a major role in showing variations in productivity i.e. 2.82 whereas effect of '**labshare**' and '**pcnsdp**' are showing negligible variations, where all the significant at 10%.

**R-squared : 0.6247**

Which means that the variation of these independent variables explains only 62% of the variations of the dependent variables, which means there may be other unknown variables affecting variations in productivity across states through years.

## **Test Results & Interpretation**

### **Hausman Test:**

data:  $Y_{it} \sim X1 + X2 + X3$

chi-square = 9.1191, d.f. = 3, p-value = 0.02775

alternative hypothesis: one model is inconsistent.

***Interpretation of Hausman test:*** By doing Hausman test we got alternative hypothesis stating that one model is inconsistent.

**H1: significant difference is found=> Fixed Effect model is to be selected**

### **Lagrange Multiplier Test - (Breusch-Pagan):**

data:  $Y_{it} \sim X1 + X2 + X3$

chi-square = 77.803, df = 1, p-value < 2.2e-16

alternative hypothesis: significant effects.

***Interpretation of Lagrange Multiplier test:*** By doing Breusch Pagan test we got alternative hypothesis, which simply means there are significant effects.

The Breusch Pagan test for heteroscedasticity is sometimes referred to as the BPG or Breusch Pagan Godfrey test. It is one of the most widely known tests for detecting heteroscedasticity in a regression model.

Therefore, our data is significantly heteroscedastic.

### **F test for individual effects (poolability test for time specific effects):**

data:  $Y_{it} \sim X1 + X2 + X3 + \text{factor}(\text{yearcode})$

F = 9.8299, df1 = 23, df2 = 109, p-value < 2.2e-16

alternative hypothesis: significant effects

***Interpretation:*** The F test for individual effects (Poolability test for time specific effects) gave us the alternative hypothesis i.e to go with the panel model instead of pooled model.

The result clearly shows that there are significant effects through time.

**F test for individual effects (poolability test for individual specific effects):**

data:  $Y_{it} \sim X_1 + X_2 + X_3$

F-statistic = 8.8833, df1 = 16, df2 = 116, p-value = 8.391e-14

alternative hypothesis: significant effects

***Interpretation:*** The F test for individual effects ( Poolability test for individual specific effects ) gave us the alternative hypothesis i.e to go with the panel model instead of pooled model.

The result clearly shows that there are significant effects through individual effects.

## **CONCLUSION**

India's strong economic growth in the last two decades has generated much optimism about its long run growth potential. Interestingly, the economy's growth trajectory has been very different from that of China and other emerging countries of Asia. In contrast to major Asian economies, the sources of growth in India have not been based in the manufacturing sector.

Instead, the country has reflected a rapid expansion of the service sector. Economists suggest that, at this point, growth in Indian manufacturing is plausible via two channels. One is capital accumulation (or investment growth), traditionally important for growth in emerging economies, while the other is productivity growth. Capital accumulation because of diminishing returns may stall growth in the long term and is therefore questioned over its sustainability. Productivity growth, on the other hand, can play a critical role in sustaining long-term growth and is, most importantly, essential for improving manufacturing competitiveness in the country (Das et al., 2016).

The theoretical literature on growth and development has highlighted industrial productivity as a major source of growth in developing economies (Bosworth et al. (2006). Resultantly, supporting productivity growth in industries is the focus of much government policy around the world. The development literature in India is replete with studies on estimation of TFP for various policy purposes (see, Goldar, 2014 for a recent review). Total factor productivity is the growth in output that is not explained by a growth in inputs.

In our study, fitting of the LSDV model and TFP estimation ( fixed effect coefficients for the states ) showed that the unobserved variation in GVA for almost all the states is significant. 'Capital intensity', 'skill' and 'labour\_share' were some of the state wise factors with which we tried to explain the productivity differential. The conclusion was that although these factors managed to explain some of the productivity variation, it was unable to explain all. Thus, it counts as a limitation in our study. Inclusion of a few more productivity factors such as human resources, technology management or highest education achievement of workers might have explained the productivity differential to a greater extent.

## **Source Code**

**Github Link** : <https://github.com/Parim7/paneldataanalysis>



## **Challenges and Limitations**

During the course of working on this project, we ran into a number of challenges. Several of these issues include the following:

- 1) **Topic:** Regarding the subject at hand, we have no prior experience with it. Regarding this project, we have no prior knowledge whatsoever. We are not familiar with the process of panel data analysis. Therefore, in order to comprehend panel data analysis, we have a lot of books and research papers that we need to read.
- 2) **Gathering information:** Once more, to acquire data, we had to deal with many challenges. The collection of the dataset has evolved into a difficult challenge. In most cases, the collection of these data requires a significant amount of people in addition to a significant sum of money. A researcher needs to go to a challenging location, where there is a possibility that they will put their lives in danger, in order to acquire this kind of data. As a result of all of these factors, the internet does not make this kind of material easily accessible to the public. A few government websites releases information on the studies carried out on industrial productivity analyses. And additionally, publish the dataset that was utilized in website, finding raw data, as well as data that has been combined from a variety of variables, can be challenging.
- 3) **Preprocessing:** The purpose for which we have decided to start this project, and the accomplishment of my objective, it has become quite tough to acquire the data. In the vast majority of instances, the quantity of the dataset was huge. There were a few instances in which we obtained sufficient data; however, there was a deficiency in comparable information. According to what we've read on the internet about how industrial productivity as a major source of growth in developing economies. Each industry has applied a uniquely individualized approach to scaling. Due to the fact that we have employed a variety of variables, we face a significant challenge in aggregating all of the data because each variable uses a unique scale mechanism.
- 4) **Statistical analysis:** We have got no similar objective in any of the previous study papers. If we had any studies on this topic, then it could be possible for us to make significant progress toward achieving the goal we have set for ourself. Handling missing data appropriately is crucial to avoid bias and loss of statistical power. Choosing the right model specification, such as deciding between fixed effects and random effects models, is essential for us because the wrong model can lead to incorrect conclusions. So, various model selection tests, like the Hausman test, help us to guide this decision.
- 5) **Model Interpretation:** Interpreting panel data models can be complex, especially when dealing with fixed or random effects. So, It's important for us understand the implications of the chosen model for the interpretation of coefficients.
- 6) **Time Constraints:** Because our available time is restricted, we have fitted some of the models. However, this data can be utilized in a variety of analytical directions using a wide variety of models, statistical tests and approaches. By doing so, we are able to do further and more in-depth studies by employing these data.

# UNIVERSITY OF KALYANI

Kalyani, Nadia, West Bengal- 741235, India

## Acknowledgement

I would like to express my special thanks of gratitude to our HOD and Professor **Chandranath Paul** Sir and other teachers of our statistics department of KU for giving this golden opportunity to do this wonderful project. I express my profound gratitude to my Project Guide Professor **Chiranjib Neogi** Sir for his continuous guidance and constant supervision as well as for providing me with all the necessary documents and information regarding the project. He showed me the path to achieve our targets by explaining all the tasks to be done and explained me the importance of this project as well as its industrial relevance. We shall be forever indebted to prof. **Chiranjib Neogi** Sir for his constant support and help without which this project would not have been successful.

Beside this, I express my gratitude to my fellow project mates –

- ❖ Shreyaa Kar
- ❖ Nirmal Kuiry
- ❖ Alekhya Guha

for their continuous company and assistance in doing this project. I express my humble thanks to our friends and family for their constant support throughout this journey.

Place – Kalyani

Date – /08/2023

## **REFERENCES**

1. Rajarathinam Arunachalam, Subh S S, Ramji Madhaiyan. (2021). Panel Data modelling for Indian food grain production.
2. Rupika Khanna, Chandan Sharma. (2018). Manufacturing productivity in Indian states. The Singapore Economic Review. 66.
3. Subhash C. Ray. (2002). Did India's economic reforms improve efficiency and productivity? A nonparametric analysis of the initial evidence from manufacturing. Indian Economic Review. 37(January):23-57
4. Chiranjib Neogi, Buddhadeb Ghosh. (1994). Inter-Temporal efficiency variations in Indian manufacturing Industries. Journal of productivity analysis. 5(3):301-324
5. Coelli, T.J. (1996). A guide to FRONTIER Version 4.1: A computer program for stochastic frontier production and cost function estimation. CEPA working paper no. 7/96.

## **DECLARATION OF STUDENT**

I, U Parimala, a M.Sc. (2<sup>nd</sup> Year) student of Department of Statistics, University of Kalyani, hereby declare that my project entitled as “ ***Explaining Productivity Differential Among Major States of India : A Panel Data Analysis***” has been completed by me as a part of M.Sc. (4<sup>th</sup> Semester) educational curriculum, 2023 under the supervision of Prof. Chiranjib Neogi, Department of Statistics of the *University of Kalyani*.

*I further declare that this report, either in full or in part, has not been published anywhere.*

Date :    /08/2023

---

*Signature*