Productivity And Production Functions: An Indian Case Study

London School of Economics and Political Science

May 2015

Course: EC 427 - Economics of Industry

Candidate Number: 74676

Word Count: 6081

(excluding Abstract, Appendix and tables; formulae counted as word equivalent)

Contents

List of Tables

Abstract. *Estimating productivity in the Indian Manufacturing Industries has received a lot of attention in the past and the production function estimation technique has been used most often for the same. However endogeneity concerns due to the existence of simultaneity and selection bias dictate that an inconsistent estimate of the parameters will be obtained unless these biases are addressed. This paper discusses the pros and cons of the traditional approaches such as the Instrumental Variables mathod and the Fixed Effects method as well as the proxy variable approaches such as those introduced by Levinsohn and Petrin (2003). This includes a discussion on the direction of these biases and the extent of their correction. Since the values of productivity are affected by government policies and by trade barriers, a short analysis of the potential determinants of productivity is also conducted.*

1 Introduction

Economic growth has long been acknowledged to depend on not just increases in inputs such as labour and capital but also on technical progress. This unobservable input has been termed 'Total Factor Productivity' and is often associated with technological innovation, sophisticated managerial practices etc. To better understand market structure, cost regulations and scale effects, estimating TFP accurately is essential. Furthermore understanding what affects TFP is also necessary, especially for government agencies and competitive authorities who need to choose optimal trade policies that maximise welfare.

In simple terms, productivity (TFP), is the efficiency with which inputs are converted into outputs. The steam locomotive and the telegraph are the most famous causes of productivity increase of the 19th century. The advent of electricity and Information and Communications Technology (ICT) since then, have been most defining in increasing the role of TFP in output production. Therefore, economists have been trying to determine the most optimal way of estimating this parameter for centuries now. However, this varies depending on the nature of the data, industry and market conditions.

One of the first unique approaches to estimating productivity was introduced by Tinbergen (1942) and Solow (1957) and is known as the Growth Accounting Approach. This method of estimating productivity attempts to separate the change in production due to input changes from residual effects (TFP). Indices such as the Kendrick Index, Solow Index and the Theil-Tornquist or Translog-Divisia Index are used in this approach to estimate productivity. The parametric equivalent method known as growth regressions can also be used to identify structural equations and thus TFP levels from aggregate data (Caselli et al., 1996). The frontier methods operate under the assumption that firms do not utilise their technology efficiently. For instance, the Stochastic Frontier Analysis (SFA) is a parametric econometric method that explains away the shortfalls to be due to random shocks (Meeusen and van den Broeck, 1977; Aigner, Lovell, and Schmidt, 1977). The estimation technique most commonly used by econometricians due to data availability and its ability to answer key questions happens to be the semi-parametric non-frontier approach commonly known as the production function estimation. [1]

Production functions map inputs to the maximum output they can produce. However there are several econometric concerns that need to be addressed while estimating the parameters of the production function. The functional form determines how TFP increases output and is thus an essential assumption. Simultaneous bias, selection bias, collinearity concerns and measurement errors are all valid problems that are widely discussed in the literature. Thus running a simple Ordinary Least Squares (OLS) will not give us consistent estimates of the parameters. Econometricians often turn to other traditional methods such as Instrumental Variables (IV) method and Fixed Effects (FE) estimation. This has further evolved into two different branches; one that exploits the dynamic panel data and the other that uses proxies for productivity. In the next section, appropriate solutions are provided for each of the issues mentioned above and the estimation techniques are described in detail. In this paper, the proxies approach along with the traditional approaches is tested empirically and an evaluation of whether the directions of the biases are in line with economic theory is conducted.

The measurement of productivity, in the second most highly populated country, India has a rich literary history. The first-generation studies estimated the TFP of the Indian manufacturing industry to be zero or negative in the period 1959-1979 (Goldar, 1986; Ahluwalia, 1985). However, the second-generation studies drew attention to the biases involved in production function estimations and began a debate regarding measurement biases that still lies unresolved. Some economists (Ahluwalia, 1991) claimed an increase in productivity due to the 1991 liberalisation, while others claimed the opposite (Mohan Rao, 1996; Balakrishnan and Pushpangadan, 1994). The third-generation studies have focused on the impact of trade reforms and industry policies on productivity in the post-reform period. Trivedi et al (2011) shows that productivity has decreased since the post-reform period while Bollard. A et al. (2013) show a miraculous increase.

The main reasons for such wide-ranging discord in the estimated values of productivity lies in the assumptions made. For instance, choosing a value-added or a gross output production function results in completely different values for productivity that have different interpretations. In this paper,

[1] Figure 1 in Appendix A maps out all the methods of estimating productivity.

the post-reform period is analysed and a set of assumptions that only aim to answer simple economic questions such as how productivity evolved over the years, the differences across the various states that have different wage rates and tariff rates, the effect of FDI on TFP etc. is considered. These questions are not largely affected by measurement issues and can be expected to give consistent estimates as long as other econometric biases are controlled for.

While several papers in the past have measured productivity in the Indian manufacturing industry using the production function method, none, to my knowledge, have compared the various estimation procedures and discussed their limitations. This aids one in understanding the trade-off involved in choosing one over the other and the direction of the biases that occurs. In addition to conducting robustness checks, this paper also attempts to analyse how productivity differs across the country and industries. Thus the objectives of this paper are to
(i) Analyse the benefits and limitations of different methods of production function and productivity estimation.
(ii) Compute the above for the Indian Manufacturing sectors and do a robustness check by using different models and manipulating the data.
(iii) Examine the relationships that might exist between productivity and some of its potential determinants.

In the next section, various econometric methods will be used to estimate the production function that is described in detail. Section 3 discusses the Indian Manufacturing Industry dataset and the variables used in the empirical estimation. Section 4 discusses the results and conducts the analysis while section 5 lists the limitations and concludes.

2 Econometric Model

In this section different models of estimating firm-level productivity are compared while controlling for any endogeneity that might originate from simultaneity bias or selection bias.

A number of standard assumptions are usually made in both these approaches regarding the functional form of a firm's production function. Suppose that a firm i's output is given at time t by the following Cobb Douglas production function:

$$Y_{\mathrm{it}} = A_{\mathrm{it}} K_{\mathrm{it}}^{\beta_k} L_{\mathrm{it}}^{\beta_l} \tag{1}$$

In the above formulation, A_{it} is the productivity for the firm. In log form this can be rewritten in the following way:

$$y_{\mathrm{it}} = \beta_k k_{\mathrm{it}} + \beta_l l_{\mathrm{it}} + \mu_{it} + \varepsilon_{it} \tag{2}$$

where $log(Y_{\mathrm{it}}) = y_{\mathrm{it}}, log(K_{\mathrm{it}}) = k_{\mathrm{it}}, log(L_{\mathrm{it}}) = l_{\mathrm{it}}$ and $log(A_{\mathrm{it}}) = \mu_{it} + \varepsilon_{it}$. Here, μ_{it} is considered to be productivity arising from factors unobservable to the econometrician but known to the firm such as managerial ability, higher firm-specific technology etc. that the firm can use to make input decisions in every period. However, ε_{it} is treated as a random variable or shock that doesn't impact the firm's decisions in choosing inputs. This includes weather conditions, unexpected union strikes that occurs after the choice of inputs are made or could be a measurement error in Y.

Two additional simple but not innocuous assumptions of perfect competition and constant returns to scales (CRS) are further made such that $\beta_k + \beta_l = 1$. Thus, the input coefficients are elasticities for that input and also equal the share of revenues paid to each input. Note that the production function can easily be extended to include intermediate inputs such as fuels or electricity to the model.

2.1 Ordinary Least Square (OLS) and Endogeneity

The most obvious way of estimating the production function and productivity is to regress equation (2), but not including the firm observed μ_{it} while running a simple OLS for each firm will result in inconsistent estimates due to the existence of endogeneity between μ_{it} and the inputs. This endogeneity can occur in two ways. More efficient firms with everything else constant will hire more inputs due to a positive productivity shock (since μ is known to firms) which leads to the simultaneity bias, while less efficient firms might exit the industry due to low productivity. This second effect is referred to as selection bias in the literature. The first effect can be seen mathematically if it is assumed that all firms are profit maximising. In the short run capital can be taken as fixed, then the firms problem is given by:

$$\max_{L} K^{\beta_k} L^{\beta_l} e^{\mu} E e^{\varepsilon} - wL \tag{3}$$

where w is the price of labour or wages and the output good is assumed to be a numeraire good. The first order condition gives an expression for the labour input demand:

$$L = \left(\frac{\beta_l K^{\beta_k} e^{\varepsilon} E e^{\mu}}{w} \right)^{\frac{1}{1-\beta_l}} \tag{4}$$

Since the firms are not aware of the disturbance ε but do have information about μ, their choice of input depends only on μ such that $E(\varepsilon|logK, logL) = 0$ and $E(\mu|logK, logL) \neq 0$

$$logL = \frac{1}{1-\beta_l}log\beta_l + \frac{\beta_k}{1-\beta_l}logK + \frac{1}{1-\beta_l}\mu \tag{5}$$

Notice that if capital is considered to be a variable and the profit maximization in the long run is calculated, K will also be found to depend on μ positively, although in line with economic theory labour is likely to be more affected by the bias than capital. Hence, an upward bias can be expected in the inputs due to the simultaneity bias. However, since higher capital stocks lead to higher profitability, inefficient firms with higher capital stocks will be less likely to exit the market, so the direction of bias due to selection will be downward. Thus, there exists an endogeneity problem in the model where the estimators are inconsistent and the directions of the inconsistency depend on the cause and magnitude of the biases.

Another big econometric issue arises due to the endogeneity caused by mis-measurement of inputs and output. Measuring capital stock has always been widely debated and the quality of the inputs are not observed either. But past literature has found that more efficient firms will have higher productivity irrespective of the measurement method. This is discussed further in section 3.

The issue of endogeneity due to μ_{it} and the selection bias has given rise to several alternate models. The most obvious and the simplest solution to this problem is to use Instrumental Variables for the input variables. However, it is not easy in practice to find appropriate IVs that are highly correlated to inputs, but not to μ_{it}. Mundlak(1961) suggested using covariance analysis while Hoch (1962) building on it suggested using a fixed effects estimation approach by imposing the condition $\mu_{it} = \mu_i$ i.e the problematic productivity factors known to firms remain constant over time for each firm.

2.2 Instrument Variable Estimation (IV)

To circumvent the endogeneity and selection bias that exists in the basic OLS estimation model, it is necessary to find good instrument variables that provide consistent estimators for the regressors. For every endogenous regressor, at least one instrumental variable z is needed that is highly correlated with the 'k' endogenous regressors but does not belong in the production function satisfying $rank(E[Z'X]) = k$. This is known as the rank condition and implies that each instrument used is related to the endogenous regressors in a unique way. In the production function context, at least 2 instrumental variables is needed if both labour and capital are assumed to be endogenous. The other condition known as the orthogonality condition says that the instruments should be uncorrelated with the error term of the production function equation $E[Z'(\mu + \varepsilon)] = 0$.

In the past, economic intuition has often been used to uncover appropriate external instruments. Some of the popular choices for instruments include input prices such as wages 'w_{it}' and 'price' of capital 'r_{it}' which definitely impact input selection by firms but are not correlated with the error terms if input markets are assumed to be perfectly competitive so that firms are input price takers. Output price 'p' can also be considered an instrument under our perfect competition assumptions, but in reality, the market structure might be more competitive. Any other demand shifters 'x' such as buyer's income, prices of substitutes and complements, market size, tax rates etc. can also be used as instruments, however in reality, data on these are not easily available. Thus $\mathbf{Z} = (w,\ r,\ p,\ x)$ can be used in the instrumental variable estimation to obtain the production function consistently and subsequently estimate productivity accurately.

Due to the availability of data on wages, it is often used as an IV. However, the danger with this is that input quality that is part of the error term could be correlated with the input prices. i.e. if low quality labour is given lower wages, then the orthogonality condition will be violated. Violation can also occur if there is little variation across the cross section for instance if the wages are the same for all the firms in an industry then it cannot be used. Even if they are not the same but the variation in the input prices is correlated with error due to market power in the input markets etc. then the estimates will become more biased.

2.3 Fixed Effects Estimation (FE)

So far, both OLS and IV can be applied to any dataset that satisfies the above-mentioned assumptions. However in order to apply the fixed effects estimation technique, a panel dataset is required to manipulate the endogenous part of the error term. For FE to work a new condition is imposed on the error term that is known to the firm but not to the econometrician:

$$\mu_{it} = \mu_i \tag{6}$$

Then the assumption is made that this μ, which could be managerial ability or better technology does not change over time. Then the variables can be demeaned so that the new model as seen in equation

(7) drops the endogenous error term due to the above assumption.

$$y_{\mathrm{it}} - \overline{y_{\mathrm{i}}} = \beta_k(k_{\mathrm{it}} - \overline{k_{\mathrm{i}}}) + \beta_l(l_{\mathrm{it}} - \overline{l_{\mathrm{i}}}) + \varepsilon_{it} - \overline{\varepsilon_{\mathrm{i}}}, \qquad n = 1, ..., N; t = 1, ..., T \tag{7}$$

Now to run a simple OLS on equation (7), a strict exogeneity condition of ε needs to be imposed in order to get consistent estimates- $E[\varepsilon_{it}|k_{\mathrm{it}}, l_{\mathrm{it}} \forall t] = 0$. This is because the $\overline{\varepsilon_{\mathrm{i}}}$ term contains the errors ε_{it} from all periods.[2]

While FE successfully resolves endogeneity problems due to simultaneity bias and selection bias, this is only true if the exogeneity assumptions hold. However for long time periods equation (6) might not be very realistic and this results in inconsistent estimators. Dynamic panels however explore relaxing this assumption. Another key issue is concerned with the measurement bias. While removing the endogeneity bias, FE exacerbates the measurement bias in the regressors. If the correlation between the regressors over time is greater than the serial correlation in the error, this bias is even bigger. So even with FE, inconsistent estimates that are biased downwards (attenuation bias from mis-measurement) can be obtained. The interaction of all the biases could very well make the OLS estimator less biased than the FE estimator.

The necessity of panel data is crucial here, however in this paper, only a repeated cross section is available so then the method suggested by Deaton (1985) is used where cohorts are constructed to get consistent estimators. These cohorts should be constructed using time invariant variables such as region, industry, starting year of firm etc. [3]. As specified in equation (15) from Appendix B, the FE estimator with cohorts is mathematically equal to an IV estimator obtained by choosing as instruments a vector of the cohort dummies interacted over time.

2.4 Levinsohn Petrin Model (LP)

With the availability of panel data, many more approaches began to be applied. For instance, one of the most influential papers Olley and Pakes (1996)[4] used investment as a proxy by capturing its relationship with capital. Levinsohn and Petrin (2003) extended this research by using intermediate inputs as proxies instead of investment. Moreover, data driven problems dictate that the LP model is more useful. For instance, the investment proxy only works for non-zero investments and empirically due to non-convex adjustment costs, investments are often zero so many observations will not be

[2]See Appendix B for the theory behind using First Differences (FD) Estimation including the underlying assumptions. The benefits and drawbacks to choosing FD over FE is also highlighted.

[3]Derivation of equivalent FE model for the cohort is in Appendix B

[4]Olley and Pakes model henceforth called OP

used. On the other hand, firms are more likely to report intermediate inputs that are less costly to adjust. But while the OP method resolved both biases, the LP method uses the information from input choice equations to correct for the endogeneity problem from only the simultaneity bias. Using an unbalanced panel, conditioning expectation of μ on survival are some ways of tackling the selection bias. But if only a repeated cross section is available and not a panel data, then firm exit and entry data is not available anyway so using the LP model is a good idea.

The first stage of the estimation introduces the demand function of the proxy- intermediate inputs as a non-dynamic function of capital and the productivity shock known to the firm, that varies across time to accommodate for changing market structures and competition.

$$\iota_{it} = \phi_t(k_{it}, \mu_{it}) \tag{8}$$

To be a valid proxy, ι_{it} to required to be monotonic in μ_{it} for all relevant k_{it} i.e. $\frac{d\iota}{d\mu} \geqslant 0$. This can be done subject to some strict assumptions as outlined in LP. Then the function can be inverted to get an expression for μ_{it} and substituting that into equation (2) gives:

$$\mu_{it} = {\phi_t}^{-1}(k_{it}, \iota_{it}) \tag{9}$$

$$\begin{aligned} y_{\text{it}} &= \beta_k k_{\text{it}} + \beta_l l_{\text{it}} + {\phi_t}^{-1}(k_{it}, \iota_{it}) + \varepsilon_{it} \\ &= \beta_l l_{\text{it}} + \Phi_t(k_{it}, \iota_{it}) + \varepsilon_{it} \end{aligned} \tag{10}$$

This is the first stage of the estimation which runs a regression on the model (10) while treating Φ_{it} as a non-parametric function of capital and intermediate inputs (for example a second order polynomial series) since $\Phi_t(k_{it}, \iota_{it}) = \beta_k k_{\text{it}} + {\phi_t}^{-1}(k_{it}, \iota_{it})$. This provides a consistent estimate of β_l since the model no longer has any endogeneity.

The second stage supposes that the known productivity shock evolves according to a stochastic first-order markov process- $\mu_{it} = f(\mu_{it-1}) + \omega_{it}$ where ω_{it} is a white noise process and so $\mu_{it} - E[\mu_{it}|\mu_{it-1}] = \omega_{it}]$ Using equation (9), μ_{it-1} of the previous period estimated from (10) and the estimated $\widehat{\beta}_l$, the production function becomes:

$$\begin{aligned} y_{\text{it}} - \widehat{\beta}_l l_{\text{it}} &= \beta_k k_{\text{it}} + f(\Phi_{t-1}(k_{it-1}, \iota_{it-1}) - \beta_k k_{\text{it-1}}) + \omega_{it} + \varepsilon_{it} \\ &= \beta_k k_{\text{it}} + g(k_{it-1}, \iota_{it-1}) + \omega_{it} + \varepsilon_{it} \end{aligned} \tag{11}$$

Since capital k_{it} is chosen in the previous period $t-1$, it is not correlated with the errors in equation (11) i.e. $E[k_{it}.(\mu_{it} - E[\mu_{it}|\mu_{it-1}])] = 0$. However ι_t can be expected to be correlated with ω_{it}, but since ι_{t-1} is chosen before either of the errors occur it can be assumed that $\omega_{it} + \varepsilon_{it}$ does not produce any endogeneity in the model. Once again if $g(.)$ is taken as a non-parametric equation, β_l can be estimated consistently by running the regression on equation (11).

The above method holds true for a panel data, but in case of a repeated cross section dataset the second stage estimation of β_k becomes impossible since there are no firm identifiers and thus no way of using previous period productivity. To circumvent this issue, Sivadasan(2009) modified the above by constructing cohorts based on time invariant variables and then used the average productivity of the corresponding cohort in the previous period instead of the prior period productivity of the firm. [5] Thus without a dynamic model, the data can be manipulated to obtain consistent estimates. Empirically LP has often estimated a lower coefficient for labour than the OLS. However, LP never provides efficient estimates due to the existence of obvious serial correlation and therefore a bootstrapping procedure needs to be employed to estimate correct standard errors.

2.5 Other Estimation Techniques

Ackerberg, Caves and Frazer (2006) argued that the above model will not produce consistent estimators in the first stage due to the collinearity between l_{it} and $\Phi(.)$ in equation (13). This is because of the non-dynamic assumption of labour that will also depend on capital and productivity. If this is to be avoided, an external variable that moves around labour independently of ${\phi_t}^{-1}(k_{it}, \iota_{it})$ and k_{it} while maintaining above mentioned assumptions needs to be used in the data generating process (DGP) of labour. For instance, the assumption that firms make an optimization error while choosing the optimal level of labour might hold. Further making simple sensible assumptions regarding the timing of input selection and productivity shocks can also help resolve the collinearity issues. Two possible scenarios are- (i) labour l_{it} is chosen at time $t-b$ where $b \in (0,1)$ while intermediate inputs are chosen at time t so that there are i.i.d shocks that affect the price of labour and thus the optimal value of labour dispelling the collinearity, and (ii) assume labour is not perfectly variable and has to be chosen at $t-b$ where $b \in (0,1)$ while ι_{it} is chosen at t but μ_{it} evolves between $t-b$ and t and affects the choice of intermediate inputs after labour is chosen. Collinearity issues are not encountered often in practice, however awareness regarding the DGP is essential.

[5] Proof in Appendix B

The other branch of estimating production functions that is not explored in this paper deals with dynamic panel data. Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (2000), took a different approach to finding instruments by exploiting the dynamic panel data model. This was done by choosing past values of regressors as instruments as this increases efficiency in the model. Overall, the optimal choice of these different models varies depending on the nature of the data that is to be used to estimate productivity.

3 Data

The data on large, formal firms from all industrial sectors has been obtained from the Annual Survey of Industries (ASI), which also happens to be the only annual survey on Indian Manufacturing plants. It is conducted by the Central Statistical Organization, a department of the Ministry of Statistics and Programme Implementation, Government of India and includes all the industrial units registered under the Factory Act, 1948. The firm-level data spans across 9 years from 1998-1999 to 2007-2008 and contains detailed information for 93 two-digit industry codes (grouped together according to the NIC- National Industrial Classification which is based on the UNISIC- United Nations International Standard Industrial Classification). The data collection process separates the firms into two categories - (i) Census sector units, and (ii) Sample sector units. The firms in the census sector need to have more than 200 employees or must be public sector undertakings. Information regarding these firms is collecting annually. All other firms are part of the sample sector surveys that is conducted every 3 years.

Concerning the choice of production function, most studies prefer the value added production function estimation method to the gross output measure. The obvious difference between them is that the gross output measure of output includes intermediate inputs. Cobbold (2003) shows that using either method results in a different interpretation and that at a national level, the two methods do not produce very different results for TFP and so the Value Added estimation method is used. The variables required to estimate productivity include (i) Net value added, (ii) Number of workers and employees, (iii) Net value of fixed assets, (iv) Payment of wages and emoluments and (v) Net value of intermediate inputs.

The ASI framework provides Gross Value Added Output measured at current prices. To convert this nominal measure into real terms, the Wholesale Price Index (WPI) for the whole industry released by the Office of the Economic Advisor, Ministry of Commerce and Industry of the Government of

India is used. In the data however, taking a log transformation results in negative values for some small firms when the value added measured at constant prices is less than unity, so one is added to the Value Added Output.

Traditionally labour is measured as the total number of workers and employees where workers refers to contract labourers while employees refers to managerial staff and others. Other papers also use total number of hours worked or segregate blue collared workers from white collared workers. Labour is constructed as average number of workers multiplied by the number of working days to get a good estimate of the amount of work supplied by workers and employees to each firm.

The measurement of capital is widely debated and highly controversial even today since in practice, it is the most complex to measure. Conventionally, the book value of total net fixed assets is considered to constitute capital. Other studies have also employed the perpetual inventory method to create the capital stock series from annual investment data. Due to lack of investment data, this paper uses the book value method but since it does not truly represent the physical stock of machinery used and does not consider capacity utilization, it may not be very representative of capital expenditure and might be biased.[6] The ASI capital inputs is the sum of the net fixed assets that include the net value of land, buildings and other construction, plants and machinery, tools and other fixed assets. Capital is also deflated using the WPI at the 1998-99 values.

The value of wages provided refers to the payment made to workers while emoluments are the payment made to all employees and includes wages, salaries, bonuses and the surpluses. The variable wage used is total emoluments divided by total number of workers and employees. Note that different states have different minimum wages and laws thus ensuring variability. The firms' intermediate input costs are also reported in current prices, so they are deflated with the WPI too[7]. For the LP estimation method, the intermediate inputs used are fuel, electricity and raw materials. Since in the gross output production function method, estimates of all the intermediate inputs need to be obtained, this might result in collinearity issues, Bond and Söderbom (2005) argue that the value added production function is thus better, further justifying our choice.

Data on Foreign Direct Investment and wages across the states were retrieved from the IndiaStat

[6] As mentioned in Kathuria et al., 2012

[7] Since both output and input are deflated by the same price index this is known as the Single Deflation(SD) method and is usually less favored to the Double Deflation method where the inputs are assumed to move differently from the output and so a different index is used to transform them into real variables (Balakrishnan and Pushpangadan, 1994). However the difficulty involved in finding and compiling an appropriate index for inputs to use in the DD method implies that most papers use the SD method. In this paper, the SD method is used with the assumption that unlike the liberalisation time period, during 1998-99 to 2007-08, the input and output prices moved together.

website that has compiled data from the Indian Government and the Reserve Bank of India (RBI). Data on Tariff rates on goods and services was collected from World Integrated Trade Solution (WITS), World Bank for the corresponding industries.

There are exactly 528,215 firm-per-year observations recorded across all industries for the 9 years in the panel which includes information on variables such as total output, labour, capital, cost of labour, intermediate inputs such as fuel, electricity and raw materials. However, firms with negative output values even after the log transformation described above were dropped. Similarly firms with missing information on inputs, raw materials and industry SIC code were also dropped. Next only manufacturing industries are analysed and the agriculture firms, business firms etc. are dropped. Moreover, some observations were recorded with errors such as the starting year of the firm being recorded as 3001 etc. Firms with starting years <1800 have also been dropped and only firms that are in operation during the chosen interval are used. Thus after this cleaning process there are 297,133 firms to work with.

The biggest concern with the ASI dataset is that there are no establishment identifiers for every unique firm, so instead of a panel data, only a repeated cross section is available. So, in order to be able to use the FE and LP methods of estimation, it is necessary to match the firms across the years using some constant variables such as location, rural/urban, SIC5 code, starting year etc. Since there are 36 states in India and on average 14 districts per state with the largest state having 88 districts, to a large extent the firms can be identified. Information regarding ownership, type of organization, scheme, units etc. is also used for identification.

However there are some duplicate observations in the data and in some cases, there exist very similar firms that match in all the above-mentioned aspects. So the duplicates are dropped and very similar firms are averaged so as to get an unbalanced panel. The loss in terms of firms from averaging is 4.6%, which is small enough for us to accept the tradeoff of being able to use more sophisticated estimation techniques. Another major issue encountered is if the sampler entered the data incorrectly, then the same firms will not be matched across the years resulting in a highly unbalanced panel with data for several firms available for only a year. Therefore, those firms that have data for only one year are dropped and thus a balanced panel dataset with 118,524 firms is observed.

Finally, the repeated cross section structure is manipulated by creating cohorts that are built by using industry, starting year and location as the time invariant variables that unite these firms so as to apply the modified FE and LP described in section 2. Creating cohorts based on state, district, SIC2 industry information and the starting year gives 37,301 observations.

Thus for a robust analysis, the production function is estimated using the OLS, IV, FE and LP models with 3 different datasets- (i) Running OLS, IV on the full dataset, (ii) running OLS, IV, FE and LP on the collapsed, strongly balanced panel data, (iii) running OLS, IV and FE on the dataset with cohorts where the FE estimation is equivalent to estimating the original dataset with the time interacted cohort dummies as instruments and (iv) running modified LP using the repeated cross section and creating cohorts for the second stage. For the IV approach, wages are chosen as an instrument for labour and it is assumed that capital is chosen before the productivity shock is observed and is thus exogenous. For the LP estimation technique, electricity, fuel as well as total intermediate inputs are used as proxies for μ_{it} [8].

For OLS and IV, the TFP can be retrieved from the estimated residuals of the regression. For FE, the time invariant μ_i is obtained and exponentially transformed to obtain TFP, while a Stata command in LP directly gives us the estimated productivity.

4 Analysis

This section summarizes the estimated coefficients of the production function and thus productivity and discusses the robustness of the results. Then some analysis of the estimated productivity with policy-influenced variables such as FDI is made.

4.1 Estimation Results

Tables 1, 2 and 3 provide the summary statistics of the variables used in the three datasets described earlier. Table 4 maps the different estimates obtained by using all the methods on the different datasets. On all the datasets, the OLS and IV estimates are quite similar across samples, which justifies the use of the FE and LP on the pseudo panel and cohort sample. Since using wages as an IV results in a higher labour estimate and lower capital estimate than from the OLS, it can be concluded that there had existed a downward bias on labour and an upward bias on capital earlier due to simultaneity and selection problems. The fixed effects gives estimates that are very similar to that of the IV and consistent with past literature the coefficient of capital is much smaller in the FE

[8] These are represented as LP-E, LP-F and LP-M

than from other methods. This could be due to exacerbated measurement error problems. The LP test however, gives varying results for the manufacturing industry as a whole.

The standard errors are represented in brackets. The OLS errors are inconsistent and cannot be used for hypothesis testing. Therefore, under fixed effects the errors are clustered at the firm level while the standard errors for LP is bootstrapped using 250 replications.

Using the IV estimation method, table 5, provides the production estimates for the key states in India. During this estimation, logged wages were once again taken as instruments for labour while capital was assumed to be exogenously determined since it was chosen before the productivity shock. Due the existence of variability in wages across states, the instrument appears to be strong.

For the whole manufacturing industry, the OLS indicates constant returns to scales (CRS), while the FE and LP indicate increasing returns to scale at a 95% confidence interval. However, to interpret the Returns to Scale factor, it might be better to analyse it for every SIC2 industry separately. As can be seen from table 6, using the whole repeated cross section dataset, the coefficients for labour and capital add up to less than 1 implying a decreasing returns to scales. Nevertheless, care must be taken since the instrument might not be very strong and the assumption of exogenous capital might be misleading. So then the FE or LP for the cohorts might provide better insight.

Table 7 shows that there is very little variation across the years. The estimate for capital appears to be increasing implying a change from DRS to CRS, but if the instrument chosen is weak then these estimates are not consistent either.

The TFP level estimates obtained from taking the exponential of the residuals from the production function method are similar for OLS, IV and FE (Table 8). However, the values are much higher for the LP method. The LP-E is much closer in value to the OLS method, but the percentage of non-zero observations is also lower for the electricity proxy than for fuel or materials. However while looking at which industry is more productive than the other, all the methods indicate that the Food industry and Basic Metals show the highest TFP levels while Machinery and Minerals show the least. Similarly, among states Delhi and Rajasthan are seen to have the highest productivity in this time period while Kerala and Tamil Nadu have the least TFP levels (Table 5). Across the years, the TFP level estimates have been steadily increasing (Table 7).

4.2 TFP Implications

Fast growing industries tend to be better funded, pay better and enjoy several benefits over other industries. The relationship between FDI and TFP has been widely studied in the past since fast growing industries attract more investment but better invested industries tend to have higher technological advantages and grow faster. Using the one year lagged FDI data on the Food Industry from 2000-2007, the regression with the TFP as the dependent variable yields a small and positive coefficient on FDI equal to .0000647 (2.05E-06) significant at a 1% level, thus confirming the results obtained in other studies. The literature attempting to establish this correlation as causation is constantly evolving and studies that are trying to determine the optimal amount of FDI required to obtain maximum productivity are complex and controversial (Pessoa, 2005). For now, this paper just wishes to establish this relationship for the Indian manufacturing industry since this is relevant and opportune for the Indian government.

Similarly consistent with current literature (Dovis and Milgram-Baleix, 2007) a negative and significant (at 1% level) coefficient equal to -0.161 (.0019888) can be found when Indian tariff rates are regressed on the overall TFP across all the years. To further this analysis, the system-GMM can be used to see how the tariff rates affect the TFP rates obtained from the LP method instead of the OLS residuals that was used instead.

The results show that there exists a substantial difference between the OLS, IV, FE and proxy estimation methods implying that the simultaneous bias has been eliminated. Using a better dataset with firm establishers could help in controlling for the endogenous exit (selection bias) problems. The TFP measures provided by the LP method is too high which might stem from measurement errors. So robust analysis with a different measure for the proxies and capital can be used to cross check this. Collinearity issues as discussed earlier can be eradicated by using the ACF method. The entire branch of dynamic panel analysis has not been explored in this paper. Exploiting lagged regressors as IVs can be an excellent method of determining the validity of our current external IVs.

5 Conclusion

Estimating the parameters of the production function is an arduous process and different models require various strong assumptions that the data may or may not fulfill. In this Indian Case Study, from 1998-2008, the Food Industry and Basic Metals show the highest productivity irrespective of the model used.

The differences in the estimates across models raise the seriousness of the endogeneity bias and measurement problem. Special care should be taken to allow for expected directions of biases. In this paper, most of the empirical results are in line with theory, although access to firm identifiers could have made the results more powerful. While conducting further analysis with TFP, correlation between policy measures should not be mistaken for causation and more effort needs to be spent in exploring this.

6 Appendix

6.1 Appendix A: Statistical Tables and Figures

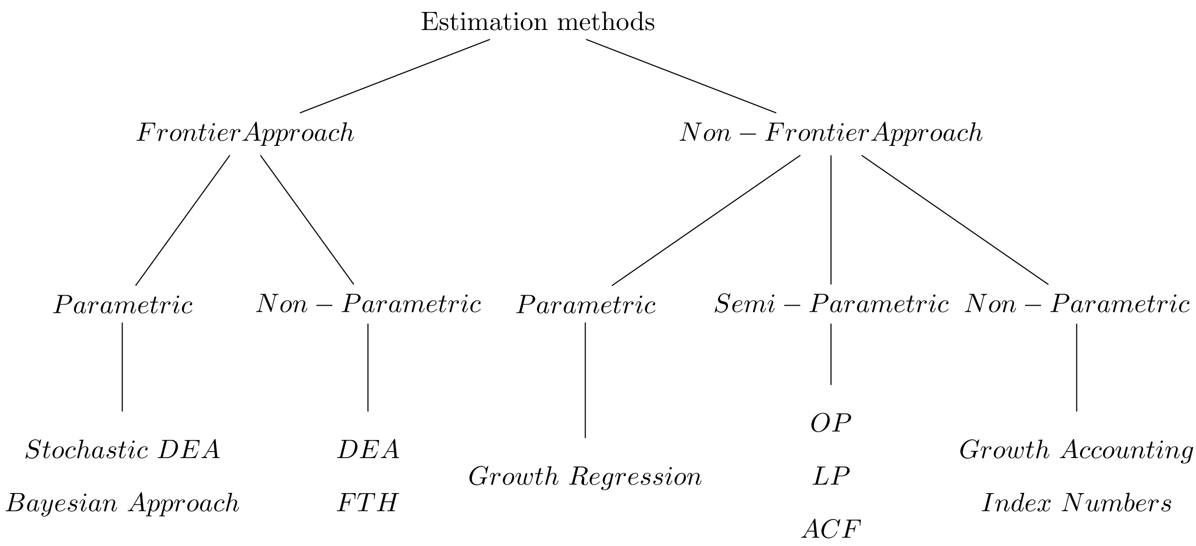

Figure 1: Estimation methods

	Mean	Std Dev	Min	Max	# Obs.
Value Added	3.58E+06	5.07E+07	0	9.29E+09	297133
Capital	1.54E+06	2.04E+07	0.01	3.08E+09	297133
Labour	185	800	1	61869	297133
Wage	1.68E+07	1.24E+10	0	1.57E+10	297133
Electricity	7.36E+05	3.18E+07	0	5.23E+09	265964
Fuel	2.67E+07	3.52E+08	1	7.65E+10	198424
Materials	2.80E+07	3.21E+08	1	6.81E+10	297133

Table 1: Summary Statistics of Entire Sample

	Mean	Std Dev	Min	Max	# Obs.
Value Added	4.07E+06	4.50E+07	0	6.54E+09	88072
Capital	1.94E+06	2.12E+07	0.01	2.29E+09	88072
Labour	248	1037	1	58972	88072
Wage	2.26E+07	1.66E+08	0	1.57E+19	88072
Electricity	1.26E+06	4.36E+07	0	5.23E+09	82348
Fuel	3.08E+07	2.73E+08	1	2.88E+10	64957
Materials	2.85E+07	2.62E+08	1	3.40E+10	88072

Table 2: Summary Statistics of Cohort Sample

	Mean	Std Dev	Min	Max	# Obs.
Value Added	4.88E+06	6.19E+07	0	9.29E+09	118524
Capital	2.06E+06	2.60E+07	0.01	3.08E+09	118524
Labour	258	1005	1	45901	118524
Wage	2.36E+07	1.53E+08	0	1.57E+10	118524
Electricity	1.13E+06	4.15E+07	0	5.23E+09	108397
Fuel	4.41E+07	4.84E+08	1	7.65E+10	85647
Materials	3.74E+07	3.83E+08	1	6.81E+10	118524

Table 3: Summary Statistics of Pseudo Panel Sample

Sample	OLS	IV	FE	LP-M	LP-F	LP-E
Repeated Cross Section						
Capital	0.5450	0.1330				
	(0.002)	(0.002)				
Labour	0.4566	0.7973				
	(0.004)	(0.002)				
Returns to Scale	1.0017	0.9304				
Cohort Sample						
Capital	0.5228	0.1492	0.1689	0.3882	0.3110	0.3058
	(0.004)	(0.005)	(0.007)	(0.007)	(0.010)	(0.009)
Labour	0.4664	0.7653	0.7730	0.4156	0.4741	0.5566
	(0.003)	(0.004)	(0.009)	(0.136)	(0.013)	(0.004)
Returns to Scale	0.9893	0.9145	0.9420	0.8039	0.7851	0.8624
Pseudo Panel Sample						
Capital	0.5776	0.1718	0.1497	0.3118	0.6188	0.2787
	(0.003)	(0.005)	(0.009)	(0.026)	(0.003)	(0.008)
Labour	0.4184	0.7473	0.9193	0.4161	0.4812	0.5728
	(0.003)	(0.004)	(0.125)	(0.006)	(0.005)	(0.003)
Returns to Scale	0.9960	0.9190	1.0690	0.7280	1.0999	0.8515

Table 4: Estmation of Production Function Parameters

State-wise	Capital	Labour	Returns to Scale	TFP Residual
Punjab	0.3610	0.6318	0.9927	0.2945
Haryana	0.4348	0.5401	0.9750	0.3214
Delhi	0.4050	0.5914	0.9964	0.8002
Rajastan	0.4046	0.5830	0.9875	0.5357
Uttar Pradesh	0.4532	0.5373	0.9904	0.1143
West Bengal	0.4383	0.5411	0.9794	-0.2606
Madhya Pradesh	0.3281	0.6381	0.9662	0.2233
Gujarat	0.3945	0.5816	0.9761	0.2851
Maharashtra	0.3324	0.6080	0.9403	0.0427
Andhra Pradesh	0.2541	0.7072	0.9613	0.0470
Karnataka	0.4147	0.5389	0.9536	-0.1368
Kerala	0.3302	0.6226	0.9528	-0.2786
Tamil Nadu	0.3328	0.6193	0.9521	-0.3148

Table 5: State-wise Estimates of Inputs and TFP Residual

Industry	Food	Textiles	Chemicals	Rubber	Minerals	Basic Metal	Metal Products	Machinery
Capital	0.1517	0.3272	0.2712	0.2561	0.3493	0.3013	0.2954	0.2626
(SE)	(0.005)	(0.007)	(0.007)	(0.011)	(0.007)	(0.009)	(0.010)	(0.009)
Labour	0.7605	0.5833	0.6419	0.6571	0.5508	0.6585	0.6170	0.6334
(SE)	(0.004)	(0.006)	(0.005)	(0.008)	(0.005)	(0.007)	(0.007)	(0.006)
Returns to scale	0.9122	0.9105	0.9132	0.9131	0.9000	0.9599	0.9124	0.8960

Table 6: Industry-wise Estimates of Inputs

OLS	1998-99	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08
Capital	0.2805	0.2744	0.2880	0.1962	0.2967	0.28201	0.2866	0.3121	0.3212	0.3095
(SE)	(0.006)	(0.006)	(0.005)	(0.016)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.003)
Labour	0.6493	0.6492	0.6386	0.7532	0.6328	0.6482	0.6484	0.6301	0.6260	0.6359
(SE)	(0.006)	(0.004)	(0.004)	(0.017)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)
RTS	0.9297	0.9236	0.9266	0.9494	0.9295	0.9302	0.9350	0.9422	0.9472	0.9454

Table 7: Year-wise Estimates of Inputs

TFP	Food	Textiles	Chemicals	Rubber	Minerals	Basic Metal	Metal Products	Machinery
OLS	1.223431848	1.193273654	1.171573129	1.152307092	1.115236065	1.233424689	1.176799894	1.159472778
IV	1.139166324	1.137315317	1.094565302	1.112819183	1.060940947	1.167730708	1.120847505	1.111853564
FE	2.066772	0.9101223	1.367847	1.462589	0.7568173	2.403407	43.55821	0.9878745
LP-M	20.31976	10.78822	10.99872	9.561249	6.28719	18.20376	10.65572	7.982181
LP-F	20.31976	12.57981	15.06659	12.80612	8.052952	25.37797	16.89289	9.948047
LP-E	5.363509	2.611725	3.260628	2.997946	1.859374	5.504287	7.035049	2.323058

Table 8: Estimated values of TFP (using cohort sample for FE and LP)

6.2 Appendix B: Theory and Mathematical Proofs

I. First Differences Estimation

A slightly weaker assumption of exogeneity can be made if we use First Differences(FD) method of estimation where we require the exogeneity of the error term with the past and current terms of the regressors- $E[\varepsilon_{i\tau}|k_{\text{i}\tau}, l_{\text{i}\tau} \text{ for } \tau = t-1, t, t+1] = 0$. This estimation technique requires equation (6) and estimates the following model:

$$y_{\text{it}} - y_{\text{it-1}} = \beta_k(k_{\text{it}} - k_{\text{it-1}}) + \beta_l(l_{\text{it}} - l_{\text{it-1}}) + \varepsilon_{it} - \varepsilon_{it-1}, \qquad n = 1, ..., N; t = 1, ..., T \tag{12}$$

If strict exogeneity holds we will have consistent estimates in both models and we can decide between FD and FE based on efficiency. This depends on the serial correlation nature of the errors. But within a time period, the same shock such as a national disaster or a global recession might affect all the firms equally so there exists serial correlation. To get efficient estimators, we must use a clustered estimator of the variance covariance matrix. However since FE has the property that as $T \to \infty$, the serial correlation in the FE error term goes to zero while the serial correlation in FD doesn't depend on T, in practice FE is preferred for production function estimation.

II. Fixed Effects for Repeated Cross Section

With a repeated cross section, we can build cohorts based on time invariant variables such as location, industry, starting year of firm etc. Then equation (7) becomes:

$$y_{\text{ct}} - \overline{y_{\text{c}}} = \beta_k(k_{\text{ct}} - \overline{k_{\text{c}}}) + \beta_l(l_{\text{ct}} - \overline{l_{\text{c}}}) + \varepsilon_{ct} - \overline{\varepsilon_{\text{c}}}, \qquad c = 1, ..., C; t = 1, ..., T \tag{13}$$

where y_{ct} is the average of all y_{it}'s in cohort c at time t and $\overline{y_{\text{c}}}$ is the time average of the observed mean for cohort c. Note that similar to equation (6) using the same lines of argument we assume that $\mu_{ct} = \mu_c$. Moffit(1993) assumes that N tends to infinity while C is constant and tends to infinity only asymptotically and shows that grouping works in a similar manner to an instrumental variable. The within estimator of each regressor on this pseudo panel can be written as:

$$\widehat{\beta}_x = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (x_{\text{ct}} - \overline{x_{\text{c}}})(x_{\text{ct}} - \overline{x_{\text{c}}})' \right)^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (x_{\text{ct}} - \overline{x_{\text{c}}})(y_{\text{ct}} - \overline{y_{\text{c}}})' \tag{14}$$

where $x = (k \quad l)$ is the vector of regressors. If we decompose each firm's μ_i into a cohort effect μ_c and the firm's deviation from this effect ν_i such that $\mu_i = \sum_{c=1}^{C} \mu_c \delta_{ci} + \nu_i$ by making use of a dummy variable δ for each cohort, we can substitute this into equation (2) to get an endogenous model where the deviation ν_i is correlated with the regressors. So we can use as instruments in this model the vector $\delta_i = (\delta_{1i}, \delta_{2i}, ..., \delta_{Ci})'$ interacted with time. We then get the predicted values $\widehat{x_{\text{it}}} = x_{\text{ct}}$ so that the IV estimator is the exact same expression as equation (15). Verbeek(2008) lists out the underlying assumptions and the asymptotic behaviour of the pseudo panel estimators.

III. Levinsohn Petrin for Repeated Cross Section

Since previous period productivity cannot be identified in a repeated cross section, cohorts based on time invariant variables are constructed and the average productivity of this cohort in the prior period is used instead. Then we require $E[k_{it}.\{\mu_{it} - E[\mu_{it}|\overline{\mu}_{it-1}]\}] = 0$ where $\overline{\mu}_{it-1} = \frac{1}{n_c}\sum_{x=1}^{n_c} \mu_{xt-1}$ and n_c is the total number of firms in each cohort $'c'$. Now we consider the following modified second stage regression equation:

$$y_{\text{it}} - \widehat{\beta_l} l_{\text{it}} = \beta_k k_{\text{it}} + E[\mu_{it}|\overline{\mu}_{it-1}] + \eta_{it} \tag{15}$$

where $\eta_{it} = \mu_{it} - E[\mu_{it}|\overline{\mu}_{it-1}] + \omega_{it} + \varepsilon_{it}$ and the predicted value of the expected productivity shock $E[\widehat{\mu_{it}|\overline{\mu}_{it-1}}]$ can be obtained from the data. [9] Minimizing the sum of the squared residuals of equation (15), gives a consistent estimate of β_k. However standard errors will be wrong and need to be calculated by using a bootstrapping procedure.

[9] Full details of estimating this term is available in the supplementary appendix of Sivadasan, 2009

References

[1] Ackerberg, D., Caves, K. and Frazer, G. (2008). Structural Identification of Production Functions. Working paper, UCLA.

[2] Ahluwalia, I.J. (1985). Industrial Growth in India. Stagnation since the Mid-Sixties. *Oxford University Press, Delhi.*

[3] Aigner, D.J., Lovell, C.A.K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6, 21-37.

[4] Bollard, A., Klenow, P. and Sharma, G. (2013). India's Mysterious Manufacturing Miracle. *Review of Economic Dynamics, Elsevier for the Society for Economic Dynamics*, vol. 16(1), pages 59-85, January.

[5] Arellano, M. and Bond, S. (1991). Some Test of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58, 277–297.

[6] Balakrishnan, P. and Pushpangadan, K. (1994). Total Factor Productivity Growth in Manufacturing Industry: A Fresh Look. *Economic and Political Weekly*, 29, 2028-35.

[7] Blundell, R. and Bond, S. (2000). GMM Estimation With Persistent Panel Data: an application to production functions. *Econometric Reviews*, 19: 321-340.

[8] Caselli, F., Esquivel, G. and Lefort, F. (1996): Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics. *Journal of Economic Growth*, 1, 363–389.

[9] Cobbold, T. (2003). A comparison of gross output and value-added methods of productivity estimation. *Camberra: Productivity Commission.*

[10] Deaton, A. (1985). Panel Data from Time Series of Cross Sections. *Journal of Econometrics*, 30, 109–126.

[11] Dovis, M., Milgram Baleix, J. (2009). Trade, Tariffs and Total Factor Productivity: The Case of Spanish Firms. *The World Economy*, 32 4 (04) 575-605.

[12] Goldar, B.N. (1986). Productivity Growth in Indian Industry. *Allied Publishers, New Delhi.*

[13] Hoch, I. (1962). Estimation of Production Parameters Combining Time-Series and Cross-Section Data. *Econometrica*, 30:34-53.

[14] Kathuria, V., Raj S.N.R. and Sen, K. (2012). Productivity measurement in Indian manufacturing: A comparison of alternative methods. *Manchester : Institute for Development Policy and Management.*

[15] J Mohan Rao (1996). Manufacturing Productivity Growth - Method and Measurement. *Economic and Political Weekly*, November.

[16] Levinsohn, J. and Petrin, A. (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies*, 70(2): 317-342.

[17] Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *Internation Economic Review*, 18, 435-444.

[18] Moffitt, R. (1993). Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections. *Journal of Econometrics*, 59, 99–123.

[19] Mundlak, Y. (1961). Empirical Production Function Free of Management Bias. *Journal of Farm Economics*, 44-56.

[20] Olley, S. and Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64:1263-1295.

[21] Pessoa, A. (2005). Foreign direct investment and total factor productivity in OECD countries: evidence from aggregate data. *FEP Working Papers 188, Universidade do Porto, Faculdade de Economia do Porto.*

[22] Petrin, A. and Poi, B. (2004). Production function estimation in Stata using inputs to control for unobservables. *The Stata Journal 4*, Number 2, pp. 113-123.

[23] Sivadasan, J. (2009). Barriers to Competition and Productivity: Evidence from India. *The BE Journal of Economic Analysis and Policy*, 9(1): Article 42.

[24] Solow, R. M. (1957). Technical Change and the Aggregate Production Function. *Review of Economics and Statistics*, 39(3), 312–320.

[25] Tinbergen, J. Z. T. (1942). Theorie der Langfristigen Wirtschaftsentwicklung. Weltwirtsschaftliches Archiv, 55(1), 511-549 (translated as On the theory of trend movements.) In: Klassen, L.H., Koych, L.M., and Witteveen H.J., eds. Jan Tinbergen Selected Papers. Amsterdam: North Holland; 1959: 1982-1221.

[26] Trivedi, P., Lakshmanan, L., Jain, R., and Gupta, Y.K. (2011). Productivity, Efficiency and Competitiveness of The Indian Manufacturing Sector. *Development Research Group Study (Mumbai: Department of Economic Analysis and Policy*, Reserve Bank of India).

[27] Verbeek, M. (2008). Pseudo Panels and Repeated Cross Sections. In The Econometrics of Panel Data: Fundamental and Recent Development *Theory and Practice Third Edition*, edited by L. Matyas and P.