METHODOLOGY

CHAPTER – III

METHODOLOGY

The methodology for the study "Impact of Vertical FDI Spillovers on the Productivity and Export Performance of Indian Manufacturing Firms" is discussed under the following headings:

I. Background of the Study
II. Sample Design of the Study
III. Database of the Study
IV. Construction of Variables
V. Tools of Analysis
VI. Limitations of the Study

I. Background of the Study

There has been a growing recognition in India that any credible attempt towards economic reforms must involve upgradation of technology, scale of production and linkages to the increasingly integrated globalised production system chiefly through the participation of transnational corporation. Neglected in India's development strategy before 1991, the government is now pursuing a pro-active policy to attract FDI. The Industrial Policy of 1991 provided a fairly liberalised policy framework to attract FDI into the country. FDI rose considerably in post-reform India fuelling high hopes that FDI may serve as a channel to advance economic growth. A number of research papers argued that an important potential engine of growth for developing countries is FDI. The most important reason why countries try to attract FDI is perhaps the prospect of acquiring modern technology, interpreted broadly to include product, process, and distribution technology, as well as management and marketing skills. FDI is believed to bring positive spillovers to domestic firms in the recipient country. The idea is that the presence of multinational corporations, which are among the most technologically advanced firms, can facilitate the transfer of technological and business know-how. This transfer may then spread over the entire economy leading to productivity gains in domestic firms. This kind of consideration has motivated authorities in many countries to ease restrictions on direct foreign investment and even to offer foreign investors more favourable conditions than those granted to domestic firms. The advanced technology adopted by the developed countries' affiliates may spread to local firms and yield technological benefits to

developing countries (spillover effects), which promotes the host country's economic growth. Besides, FDI also creates backward and forward linkages which spurt up economic growth and development in the host country. FDI contributes positively to income growth and productivity provided that the host country has attained a certain degree of absorptive capacity. With the economic reforms since 1991, the Indian government has been making consistent efforts to bring in more FDI into the country being well aware of its direct benefits. Though many studies have been done by researchers, the indirect benefits of FDI through its spillover effects are still being debated.

The present study is an attempt to examine the impact of FDI spillovers on the productivity and export performance of manufacturing firms in India. Studying the impact of FDI spillovers; both the horizontal and vertical spillovers on the Indian manufacturing firm was of particular interest for two reasons; first, the opening up of various sectors of the Indian economy to the multinational corporations raised the question on the potential of FDI in increasing the economic growth by increasing the productivity of Indian manufacturing firms. Second, the existing literature paid little attention to the impact of foreign presence on exporting firms of India. Usually multinational corporations build up an extensive international distribution network, and possess the knowledge and experience of international marketing. By simply imitating or collaborating with foreign enterprises, domestic exporters may learn how to improve their performance in foreign markets. Third, although this study relates to the Indian experience, it has a broader appeal. Most of the developing countries like India are opening up their economy to attract considerable amount of FDI to enjoy its benefits. Therefore, an empirical investigation of the spillover effects of FDI on the productivity and export performance of the Indian manufacturing firms could provide useful policy suggestions to those countries.

Further, most of the earlier studies looked mainly into the impact of horizontal and vertical spillovers on the productivity of domestic firms taken as a whole, without bifurcating them on the basis of their technology intensity i.e., low technology firms and high technology firms. Moreover, the extent of spillovers from the foreign-owned firms based on the structure of foreign ownership i.e., from the majority-owned foreign firms and the minority-owned foreign firms were largely neglected in the previous studies. In the light of the above discussions, the objective of the present study was focused on to analyse the impact of horizontal and vertical FDI spillovers

on the productivity and export performance of Indian manufacturing firms. This study contributes to the existing FDI spillovers literature in several ways. Firstly, the study used a longer time period when compared to other studies so that a better understanding of the transmission of FDI spillovers from the foreign firms to the domestic firms could be possible. Secondly, the study used a broad dynamic indicator, total factor productivity (TFP) to measure the productivity of Indian manufacturing firms. TFP was computed using the semi-parametric approach i.e., the Levinsohn-Petrin (2003) approach which is a better indicator of changes in productivity. Thirdly, fixed effects regression with Driscoll-Kraay standard errors was used to analyse the impact of horizontal and vertical FDI spillovers on the productivity of Indian manufacturing firms. Driscoll-Kraay standard errors produce heteroscedasticity and autocorrelation consistent standard errors.

II. Sample Design of the Study

The current study was bifurcated into two parts; the first part dealt with the trends in FDI inflows, determinants of FDI inflows and the causal nexus between FDI and economic growth for the period 1990-91 to 2012-13; and the second part dealt with the impact of vertical FDI spillovers on the productivity and export performance of Indian manufacturing firms for the period 2000-01 to 2012-13. The Indian economy was opened up with the economic reforms in the name of liberalisation, privatisation and globalisation (LPG) in 1991. With the economic reforms, the Indian government allowed FDI in many sectors of the economy which led to substantial FDI inflows into the economy. Hence this study analysed the trends and determinants of FDI inflows and its impact on India's economic growth after globalisation. The second part of the study investigated the impact of vertical FDI spillovers on the productivity and export performance of Indian manufacturing firms after 2000-01 because the foreign equity information for the Indian manufacturing firms was available only from 2000-01. Foreign equity information was essential to classify the Indian manufacturing firms into domestic-owned firms and foreign-owned firms and for classifying them as majority-owned and minority-owned foreign firms.

The micro-level data or the firm level data for the second part of the study was obtained in the following way. Initially all Indian manufacturing firms were selected for the study, and then firms for which information regarding equity, sales, raw materials, energy, salaries and wages and gross fixed assets were not available were dropped from the study. With the dropping of such firms, 525 manufacturing

firms from 18 industries remained in the study with 6825 observations. Firms having foreign equity greater than 10 per cent of total equity were classified as foreign firms or foreign affiliates (Malik, 2015). Of the 525 firms, 440 firms (84 per cent) were domestic firms and 85 firms (16 per cent) were foreign firms (Appendix-I). The motor vehicles, trailers and semi-trailers industry has the highest percentage (67 per cent) of foreign firms followed by the tobacco products industry with 50 per cent of the firms being foreign firms. The textiles industry and the food products industry had the least number of foreign firms with eight per cent and five per cent foreign firms respectively. The foreign firms were classified on the basis of structure of foreign ownership as majority-owned firms and minority-owned firms (Appendix-II). Majority-owned foreign firms were those firms with at least 50 per cent foreign equity participation and the minority-owned firms were those with above 10 per cent but below 50 per cent foreign equity participation. Manufacturing firms with less than 10 per cent foreign equity participation were considered as domestic or indigenous firms. Similar approach has been used by Demelis and Louri (2002) and Malik (2015) for defining majority and minority-owned foreign firms.

III. Database of the Study

The present study relied on secondary sources of data. Trends in FDI inflows, determinants of FDI inflows and the causal nexus between FDI and economic growth were examined for the period 1990-91 to 2012-13. The study period for the second part of the study analysing the impact of vertical FDI spillovers on the productivity and export performance of Indian manufacturing firms was 2000-01 to 2012-13. The secondary data for analysing the trends in FDI, determinants of FDI and the causal nexus between FDI and economic growth was obtained from various issues of SIA Newsletters and Factsheet on FDI, published by the Department of Industrial Policy & Promotion (DIPP), Ministry of Commerce and Industry, Government of India. The data from the Handbook of Statistics on the Indian Economy (2012-13) published by Reserve Bank of India and the Data Tables published by the Planning Commission, Government of India were also used. The impact of vertical FDI spillovers on the productivity and export performance were analysed using the data from Centre for Monitoring Indian Economy's (CMIEs) electronic database PROWESS. Firm level panel data for 18 Indian manufacturing industries in National Industrial Classification, 2008 (NIC-2008- two digit classification) obtained from PROWESS was used in the study. OECD (2007) classification was used for classifying the Indian manufacturing

firms on the basis of technology intensity viz., low technology firms and high technology firms (Appendix-III). In addition to PROWESS database, the data for Input-Output Transactions Table, 2007-08 (IOTT), National Industrial Classification, 2008 and price indices were obtained from the reports published by Central Statistical Organisation (CSO), Ministry of Statistics and Programme Implementation, Government of India.

IV. Construction of Variables

The details of the variables used in the study are outlined below.

FDI Inflows (lnf)

It is defined as the natural logarithm of total FDI inflows in the Indian economy measured in million dollars.

Gross Domestic Product (lng)

It is defined as the natural logarithm of gross domestic product measured in rupees crores. It was deflated using the GDP deflator (2004-05 prices).

Trade Openness (lnt)

It is the natural logarithm of trade openness (as a proxy to globalisation) measured as a percentage of total volume of trade (exports and imports) to GDP. Exports, imports and GDP were deflated using GDP deflator (2004-05 prices).

Exchange Rate (lner)

Exchange rate is the natural logarithm of average exchange rate and it is the dollar value of Indian rupee. It refers to the rate at which the domestic currency (rupee) is converted to the U.S. dollar.

Current Account Deficit (lncad)

It is the natural logarithm of current account deficit (CAD) in balance of payments, which is measured as total CAD as percentage of GDP. Since cad was measured as percentage it was not deflated using the GDP deflator.

Infrastructure (lnei)

Infrastructure is defined as the natural logarithm of economic infrastructure which is measured as government expenditure towards energy and transport. This variable was deflated using the GDP deflator (2004-05 prices).

Output (lny)

Output is defined as the sales of goods minus indirect taxes minus rebates and discount expenses plus closing stock of finished goods plus closing stock of

work-in-progress and semi-finished goods. Output series thus obtained was deflated using a more disaggregated level of industry price indices (2004-05 prices).

Raw materials (lnm)

Raw materials are the sum of raw materials consumed and opening stock of work-in-progress and semi-finished goods. Raw materials were deflated with the GDP deflator with 2004-05 being the bench mark year.

Energy (lne)

Energy is power and fuel deflated using wholesale indices for electricity industry (2004-05 prices).

Capital (lnk)

The capital variable was constructed using the gross fixed assets deflated using the wholesale price indices for machinery and machine tools. The deflated capital variable is in 2004-05 prices.

Labour (lnl)

Salaries, wages, bonus, ex-gratia provident fund and gratuities paid were deflated using the GDP deflator to construct the variable on labour and the bench mark year is 2004-05.

Total Factor Productivity (tfp)

It is well acknowledged that economic growth depends both on the use of factors of production such as labour and capital, the efficiency in resource use and technical progress. The efficiency in resource use is often referred to as productivity. Some researchers note that growth in productivity is the only plausible route to increase the standard of living (Balakrishnan and Pushpangadan, 1998) and is therefore a measure of welfare (Krugman, 1990). The relevance of economic growth is less meaningful if it has not affected productivity growth and hence the standard of living. This increase in productivity or productivity growth can be caused by several factors including investment in human capital, infrastructure, research and development apart from healthy business environment.

The current study analysed the influence of technology spillovers from FDI on the TFP of Indian manufacturing firms. The TFP measures the increase in total output which is not accounted for by increase in total inputs.TFP is deemed to be the broadest measure of productivity and efficiency in resource use. It aims at decomposing changes in production due to changes in quantity of inputs used and changes in all the residual factors such as change in technology, capacity utilisation,

quality of factors of production, learning by doing, etc. An increase in TFP, therefore, implies a decrease in unit cost of production (Kathuria et.al., 2013). Although the origins of TFP analysis can be traced back to the seminal paper by Solow (1957), recent years have seen a surge in both theoretical and empirical studies on TFP. However, several methodological issues emerge when TFP is estimated using traditional methods, i.e. by applying ordinary least squares (OLS) to a panel of firms.

TFP cannot be measured directly. Instead it is a residual, often called the Solow residual, which accounts for effects in total output not caused by inputs. To proxy TFP the study used firm-level residual from production function estimated at firm level. TFP was estimated using output and all production inputs such as capital, labour, raw materials and energy. It is a well known fact that estimation of production function using OLS gives inconsistent and biased estimates of explanatory variables. There are a host of firm, industry, time, and region-specific influences that are unobservable to the econometrician but are known to the firm. These unobservables might influence the usage of production inputs and usage of inputs thus determined endogenously. Since OLS technique assumes production inputs are uncorrelated with omitted unobservable variables, it fails to address this endogeneity issues and thereby results in inconsistent and biased estimates of production function, which is otherwise known as endogeneity bias (Malik, 2015). In a statistical model, a parameter or variable is said to be endogenous when there is a correlation between the parameter or variable and the error term.

At least as early as Marschak and Andrews (1944), applied researchers have worried about the potential correlation between input levels and the unobserved firm-specific productivity shocks in the estimation of production function parameter. One of the many alternatives to OLS proposed was Olley and Pakes (1996). They show the conditions under which an investment proxy controls for correlation between input levels and the unobserved productivity shock. But this advantage is strictly data-driven. It turns out that the investment proxy is only valid for plants reporting non-zero investment. Unfortunately, if plants report zero-investment these plants have to be truncated. Therefore, using intermediate input proxies as suggested by Levinsohn and Petrin (2003) instead of investment avoids this problem. This is because firms almost always report positive use of intermediate inputs like materials or electricity. Another theoretical benefit of the L-P approach is that it may be a better indicator of changes in productivity. Another method is Blundell and Bond’s (2000)

GMM estimator. This method uses lagged inputs for the endogeneity problem but it is not applicable with short time series data. This method cannot be employed to the present study owing to short time series data. Since the current study reports positive intermediate input viz., energy, the study used Levinsohn and Petrin (2003) methodology to estimate firm-level production function. The mechanics of the estimator is explained following Malik (2015). The production technology is assumed to be Cobb-Douglas

$$y_t = \beta_0 + \beta_k k_t + \beta_l l_t + \beta_m m_t + \beta_e e_t + \omega_t + \eta_t \quad (1)$$

where y_t, k_t, l_t, m_t and e_t are the logarithm of output, capital, labour input, raw materials, and energy of firm respectively, ω_t denotes productivity of the firm and η_t stands for measurement error in output, which is uncorrelated with input choices. Subscripts for firm and industry in the above equation are not used for notational convenience.

The study has taken energy as proxy to take care of the endogeneity bias. LP assume that firm's intermediate inputs (say energy) demand function, $e_t = e_t(\omega_t, k_t)$ is monotonically increasing in productivity given its capital stock. This allows inversion of energy demand function as $\omega_t = \omega_t(e_t, k_t)$. Thus the unobservable productivity term (ω_t) depends solely on two observed inputs, e_t and k_t. Rewriting equation (1) gives:

$$y_t = \beta_l l_t + \beta_m m_t + \phi(k_t, e_t) + \eta_t$$

where, $\phi(k_t, e_t) = \beta_0 + \beta_k k_t + \beta_e e_t + \omega_t(k_t, e_t)$ (2)

Here the error term (η_t) is not correlated with the inputs. The estimation of production function takes place at two stages. In the first stage, conditional moments $E(y_t \mid k_t, e_t)$, $E(m_t \mid k_t, e_t)$, and $E(l_t \mid k_t, e_t)$ are estimated. Conditional moment, say, $E(y_t \mid k_t, e_t)$, is approximated by a third order polynomial in k and e with full set of interactions. Conditional moments e.g., $E(m_t \mid k_t, e_t)$, and $E(l_t \mid k_t, e_t)$ are also approximated in the same way. Next the following equation is considered

$$y_t - E(y_t \mid k_t, e_t) = \beta_l (l_t - E(l_t \mid k_t, e_t)) + \beta_m (m_t - E(m_t \mid k_t, e_t)) \quad (3)$$

No-intercept OLS, is then used on this equation to estimate parameters, $\hat{\beta}_l$ and $\hat{\beta}_m$. In the second stage, LP assume that productivity is governed by a first-order Markov process, $\omega_t = E(\omega_t \mid \omega_{t-1}) + \xi_t$, where ξ_t is an innovation to productivity. Now compute

$\phi_t + \eta_t = y_t - \hat{\beta}_l l_t - \hat{\beta}_m m_t$ and find the estimate $\hat{\phi}_t(.)$ from the regression of $\phi_t + \eta_t$ on 3rd order polynomial of e_t and k_t with full sets of interaction terms. For the candidate value of β_k and β_e as $\beta^*{}_k$ and $\beta^*{}_e$ respectively (which can be got from OLS regression of (1), following can be computed.

$$\omega_t + \eta_t = y_t - \hat{\beta}_l l_t - \hat{\beta}_m m_t - \beta^*{}_k k_t - \beta^*{}_e e_t$$

$$\omega_{t-1} + \eta_{t-1} = \hat{\phi}_{t-1} - \beta^*{}_k k_{t-1} + \beta^*{}_e e_{t-1}$$

$E(\omega_t \mid \omega_{t-1})$ can be estimated by regressing of "$\omega_t + \eta_t$" on fourth order polynomial in "$\omega_{t-1} + \eta_{t-1}$". Given $\hat{\beta}_l$, $\hat{\beta}_m$, $\beta^*{}_k$, $\beta^*{}_e$ and $E(\omega_t + \omega_{t-1})$, the residual of the production function could be written as

$$\xi_t + \eta_t(\beta^*{}_k, \beta^*{}_e) = y_t - \hat{\beta}_l l_t - \hat{\beta}_m m_t - \beta^*{}_k k_t - \beta^*{}_e e_t - E(\omega_t + \omega_{t-1})$$

For the estimation of coefficients in the second stage, two moment conditions to identify β_e and β_k was used. First moment condition identifies β_k by assuming that capital does not respond to the innovation in productivity i.e., $E(\eta_t + \xi_t \mid k_t) = 0$; second moment condition identifies β_e by using the fact that last period's energy choice should be uncorrelated with innovation in productivity this period, i.e., $E(\eta_t + \xi_t \mid e_{t-1}) = E(\xi_t e_{t-1}) = 0$. Thus, only two population moment conditions are there given by the vector of expectations:

$$E[(\eta_t + \xi_t) Z_t]$$

where Z_t is the vector given by

$$Z_t = \{k_t, e_{t-1}\}$$

Finally, the estimators of (β_k, β_e) is got by minimising the GMM criterion function

$$Q(\beta^*) = \min \beta^* \sum_{h=1}^{2} \left\{ \sum_i \sum_t (\eta_{i,t} + \xi_{i,t}(\beta^*)) Z_{i,h,t} \right\}^2$$

where i indexes firms, h indexes two instruments and t indexes time.

However, as the estimation requires several steps and taking care of variances and covariances of estimates at each stage is quite tedious job, estimates have been bootstrapped to draw inferences. The bootstrap technique resamples the empirical distribution of the observed data to construct new "bootstrapped" samples. The value of the statistic is computed for each of these samples and the distribution

of estimates so generated provides the bootstrap approximation to the sampling distribution of the statistics.

Using the estimated coefficients of production function, $\hat{\beta}_l$, $\hat{\beta}_m$, $\hat{\beta}_k$, and $\hat{\beta}_e$ were estimated.

$$\ln tfp_{ijt} = y_{ijt} - \hat{\beta}_l \ln l_{ijt} - \hat{\beta}_m \ln m_{ijt} - \hat{\beta}_k \ln k_{ijt} - \hat{\beta}_e \ln e_{ijt}$$

where i, j, t denotes firm, industry and time respectively.

TABLE – 3.1.1

PRODUCTION FUNCTION ESTIMATION FOR TFP

Independent Variables	Coefficients	Standard Error
Raw Materials (lnm)	0.661**	0.018
Labour (lnl)	0.212**	0.017
Capital (lnk)	0.137	0.071
Energy (lne)	0.150**	0.047

NB: (i) Production Function estimated using Levinsohn-Petrin (2003) Methodology

(ii) ** denotes significance at 5 per cent level

Dependent variable: Output

Table 3.1.1 denote the coefficients of raw materials (lnm), labour (lnl), capital (lnk) and energy (lne) estimated using the Levinsohn-Petrin (2003) methodology. All the coefficients viz., lnm, lnl and lne were found to be statistically significant at 5 per cent level except capital (lnk). This implies that the inputs raw materials, labour, and energy contributed significantly to the production of output. But the results made it clear that capital as an input did not significantly contribute to the production of output. This may be because of the fact that India being a developing country it may be using labour intensive technology and the quality of capital may be poor. The coefficients were estimated using the domestic firms only.

Herfindahl Index (lnhin)

Herfindahl index is meant to capture the effect of competition in industry. It is the proxy for the level of industry concentration and it is the natural logarithm of sum of the squared market shares of firms in a given industry. It is computed as follows:

$$HIN_j = \sum_i \left(\frac{S_i}{\sum S_i} \right)^2$$

where S_i is the sale of i^{th} firm and j stands for industry. Higher value of HIN indicates a more concentrated industry. A more concentrated industry implies lower competition, which creates inefficiency and thereby lowers productivity of firms in the industry.

The study followed Malik (2015) in the construction of the spillover variables, viz., horizontal FDI, backward FDI, and forward FDI that capture the technology spillovers from FDI. The construction of these variables is explained as under:

Horizontal FDI (lnh)

Horizontal FDI is the natural logarithm of a measure of the share of output produced by the foreign firms in the total output of the industry**.** Horizontal FDI is defined as

$$\ln h_{jt} = \frac{\sum_{i=1}^{m} Y_{it}^{f}}{\sum_{i=1}^{n} Y_{it}}$$

where Y_{it} is the output of firms i, in year t and Y_{it}^{f} are output of foreign firms i in same year. n stands for total number of firms in an industry consisting of both domestic and foreign firms and m denotes number of foreign firms in an industry.

Backward FDI (lnb)

Backward FDI is the natural logarithm of the share of total output of an industry that is sold to foreign firms in downstream industries. To measure the share of a firm's output sold to foreign-owned firms, the study proxies the share of the firm's output sold to foreign firms by the share of an industry's output that is sold to foreign firms. Then how to measure the share of an industry output sold to foreign firms in other industries? "If we assume that a firm's share of an industry's use of a particular input is equal to its output share, then a measure of the share of an industry output sold to foreign firms is the sum of the output shares purchased by other industries multiplied by the share of foreign output in each purchasing industry". The backward FDI for industry j at time t is given as follows:

$$\ln b_{jt} = \sum_{k \neq j} \alpha_{jk} \ln h_{kt}$$

where, α_{jk} is the proportion of industry $j's$ output supplied to industry k, which is taken from the industry x industry coefficient matrix (2007-08) constructed at

two-digit level (NIC-2008). The industry x industry coefficient matrix was constructed using Input-Output Transaction Table (2007-08). The formula shows that inputs supplied within the sector are not included, since the horizontal FDI captures this effect. This variable states that higher presence of foreign firms in downstream industry generates higher backward linkages to firms in upstream or supplying industry in host country.

Forward FDI (lnf)

Forward FDI is defined as the natural logarithm of the proportion of an industry's intermediate consumption supplied by foreign-owned firms and it measures the degree of forward linkages from foreign firms to domestic firms in downstream industries. The share of an industry's intermediate consumption supplied by foreign firms is approximated as the sum of shares of intermediate input sourced from other industries multiplied by share of foreign firms' output in each supplying industry. While measuring share of foreign firms' output in upstream or supplying industry, the study has excluded goods produced by firms for export, since only intermediate sold in the domestic markets are relevant for construction of forward FDI. The forward FDI is given as:

$$\ln f_{jt} = \sum_{w \neq j} \sigma_{wj} \left[\frac{\sum_{i=1}^{m} \left(Y_{it}^{f} - X_{it}^{f} \right)}{\sum_{i=1}^{n} \left(Y_{it} - X_{it} \right)} \right]$$

where σ_{wj} is the share of inputs purchased by industry j from industry w in total inputs sourced by industry j and superscript f stands for foreign firm and the second term of right side of equation computes the share of foreign firms' output in upstream or supplying industry. Inputs purchased within the sector are excluded. The value of the variable increases with the increase in the share of foreign firms' output in upstream industries.

Majority - Horizontal FDI (lnmjh)

It is the natural logarithm of the share of output of majority-owned foreign firms in a given industry. It is given as:

$$\ln mjh_{jt} = \frac{\sum_{i=1}^{m}\left(Maj_{it} * Y_{it}^{f}\right)}{\sum_{i=1}^{n} Y_{it}}$$

where, the numerator is the total output of majority-owned foreign firms functioning in India in industry j and year t and denominator is the total output of the same industry in the same year. Maj_{it} is a dummy variable that takes the value one for majority-owned foreign firms and zero for other firms. The value of the above variable expresses the proportion of output produced by majority-owned foreign firms in the total output of a given industry during a given year.

Majority - Backward FDI (lnmjb)

It is the natural logarithm of the share of output of an industry that is supplied to majority-owned foreign firms in downstream industry. Majority-backward FDI is defined as follows:

$$\ln mjb_{jt} = \sum_{k \neq j} \alpha_{jk} \ln mjh_{kt}$$

This variable shows that higher presence of majority-owned foreign firms in downstream industry generates higher backward linkages to firms in upstream or supplying industry.

Majority - Forward FDI (lnmjf)

It is the natural logarithm of the proportion of output of an industry that is purchased from majority-owned foreign firms in upstream industry. Following the procedure applied for forward FDI, the study approximated the share of an industry's intermediate input supplied by majority-owned foreign firms as the sum of the shares of intermediate input bought from other industries multiplied by share of output of majority-owned foreign firms in each supplying industry. The study excluded the goods produced by firms for export while measuring share of foreign firms' output in upstream or supplying industry, since only intermediate inputs sold in the domestic market are relevant for construction of majority-forward FDI. Majority-forward FDI is given as:

$$\ln mjf_{jt} = \sum_{w \neq j} \sigma_{wj} \left[\frac{\sum_{i=1}^{m}\left(Maj_{it} * \left(Y_{it}^{f} - X_{it}^{f}\right)\right)}{\sum_{i=1}^{n}\left(Y_{it} - X_{it}\right)} \right]$$

where the second term in the right side of the equation is the share of output of a given industry produced by majority-owned foreign firms.

The measures of foreign presence such as the natural logarithm of minority-horizontal FDI (lnmnh), minority-backward FDI (lnmnb) and minority-forward FDI (lnmnf) were constructed in a similar manner.

The values of σ_{wj} and α_{jk} were taken from the industry x industry coefficient matrix.

The procedure for constructing an industry x industry matrix is explained in detail below.

Industry-Industry Matrix

The Input-Output Transactions Table (IOTT) pertaining to 2007-08 was used to construct the industry x industry coefficient matrix at two-digit level (NIC-2008). The input-output table consists of two matrices: absorption matrix (commodity x industry) and make matrix (industry x commodity). The absorption matrix (of order 130x130) consists of values of commodities supplied to different industries for final use as well as intermediate inputs. The make matrix (of order 130x130) represents the values of output produced by different industries. The first step to create the industry x industry matrix is to aggregate the input-output table (absorption matrix and make matrix) for the manufacturing sector to two digit level using the sector specification for the input-output transactions, 2003-04. Secondly, a matrix X has been created by dividing each row of the absorption matrix by the total output of the commodity. Another matrix Y has been created by dividing the each row of the make matrix by the total output produced by the respective industry. As a final step, the study created a new matrix Z=YX. The new matrix Z is the industry x industry matrix (Appendix – IV). Each row of the matrix Z represents the total industry output delivered to different industries in the economy. The spillover variables σ_{wj} and α_{jk} were obtained from the industry x industry matrix.

Technology Import Intensity (Intin)

Technology import intensity controls for how the expenditure on technology imports influence the productivity of the domestic firms. Modern and advanced technologies are always priced at higher rate, higher expenditures on technology import show the firm's interest in improvement and hence there is increase in productivity of firms. Technology import intensity is measured as the natural

logarithm of the ratio of firm's expenditure on technology import to its sales value in a year. The technology import expenditure includes the expenditure on the import of capital goods and foreign exchange spending on royalty/technical know-how. Foreign exchange spending on royalty/technical know-how is the expenditure on the import of disembodied technology.

Export Intensity (lneix)

Exporting facilitates the interaction with foreign buyers and foreign markets and the consequent learning from it which boosts up the productivity of domestic firms. The study used export intensity to see the effect of exports on firm's productivity. Export intensity is defined as the natural logarithm of the ratio of firm's export to its sales value.

R&D Intensity (lnrin)

Research and development expenditure generally signals a firm's in-house technology content and its endeavour to be on the frontier technology. So it affects the productivity of the firm. The study used R&D intensity to see the impact of R&D expenditure on the productivity of firms. R&D intensity is defined as the natural logarithm of the ratio of firm's R&D expenditure to its sales value.

Export Dummy (dexp)

This variable is the dummy variable taking the value one if the domestic firms engaged in exporting in the particular year; else it takes the value zero.

V. Tools of Analysis

The current study was based on time series and firm-level panel data which are secondary and quantitative in nature. The analysis used in the study are annual per cent growth rate, compound annual growth rate, Johansen's cointegration test, vector error correction model (VECM), fixed effects regression with Driscoll-Kraay standard errors, feasible generalised least squares (FGLS) regression and Heckman selection maximum likelihood methodology. Before doing the computations, the stationarity of each series was tested using the Augmented Dickey-Fuller test. The existence of multicollinearity, heteroscedasticity, and autocorrelation in the models were tested using the Variance Inflation Factor (VIF), Breusch-Pagan/Cook-Weisberg test and Wooldridge test respectively. Apart from this the Lagrange-Multiplier (LM) test was done to test for residual autocorrelation and Jarque-Bera test was done to test for normal distribution of disturbances. The conclusions were drawn on the basis of 5 percent level of significance.

Annual Percent Growth Rate

The annual per cent growth rate or annual per cent change in FDI inflows into India was calculated using the following formula:

$$PR = \frac{(V_{Present} - V_{Past})}{V_{Past}} * 100$$

where

PR = Per cent Rate

$V_{Present}$ = Present Value

V_{Past} = Past Value

Compound Annual Growth Rate (CAGR)

CAGR is a useful measure of the growth of investment over multiple time periods, especially if the value of investment has fluctuated widely during the time period in question. CAGR of FDI inflows into India during the period 1990-91 to 2012-13 was calculated using the following formula,

$$CAGR = \left(\frac{FV}{PV}\right)^{\frac{1}{n}} - 1$$

where

FV = Ending Value

PV = Beginning Value

N = Number of Years

Unit Root Test

The current study explored the causality between FDI inflows and economic growth of India. Since the data on both FDI and GDP variables were time series in nature there was a need to test for the stationarity of time series data. The stationarity of time series data was investigated by unit root test. The study employed the Augmented Dickey Fuller Unit Root Test (Dickey and Fuller, 1979 and 1981; and Dickey et.al., 1986) to investigate the same. The test involved the estimation of following equation:

$$\Delta Y_t = \beta_1 + \beta_2 Y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta Y_{t-i} + \varepsilon_t \qquad (4)$$

where, Δ is the first difference forward operator; β_i and δ_i are constant unknown parameters; Y_{t-1} is a variable of interest and ε_t is a stationary stochastic process.

The lag lengths were chosen by Schwarz (1978) Information Criterion. The task is to test the null hypothesis of non-stationarity [I(1)] against an alternative hypothesis of stationarity [I(0)]. To determine the order of integration, equation 4 has to be modified. That is by including the second differences on lagged first and k-lags of second differences, which is as follows:

$$\Delta^2 Y_t = \lambda_t \Delta Y_{t-1} + \sum_{i=1}^{k} \mu_t \Delta^2 Y_{t-i} + \varepsilon_t \qquad (5)$$

where, $\Delta^2 = \Delta Y_t - \Delta Y_{t-1}$; λ_1 and μ_t are constant unknown parameters and ε_t is a stochastic process. The k-lagged difference terms were included so that the error terms in both the equations are serially independent.

Cointegration Test

Johansen's (1988) cointegration and VECM were employed to examine the causal nexus between FDI and economic growth in India for the period 1990-91 to 2012-13. The cointegration test is to know, whether the time series variables have different unit roots (non-cointegrated) or same unit roots (cointegrated). It clarifies the existence of long run equilibrium relationship between two time series variables. In other words, cointegrated variables if disturbed will not drift apart from each other and thus, possess a long run equilibrium relationship. Testing of the existence of cointegration among economic variables has been widely used in the empirical literature to study economic interrelationships. The presence of cointegration indicates that two series would never drift too far apart. A non-stationary variable, by definition, tends to wander extensively over time, but a pair of non-stationary variables may have the property that a particular linear combination would keep them together, that is, they do not drift too far apart. It is important to note that equations estimated with stationary variables but without regard to the underlying cointegration are also inappropriate due to the model misspecification (i.e., an omitted variable bias). Theoretically, cointegration between two time series variables can be obtained, if the linear combination of two non-stationary variables is stationary. Technically, two time series variables are cointegrated, if the following three conditions are satisfied:

(i) The variables must be integrated of the same order i.e., the number of times each variable has to be differenced in order to turn the series stationary.

(ii) There should be a linear relationship between them. That is,

$$Y_t = \alpha + \beta X_t + \varepsilon_t \quad (6)$$

and β coefficient should be significant.

(iii) The residuals in the above equation, i.e., the extent by which the two variables deviate from the long run equilibrium relationship (given by the equilibrium error (ε_t)) should be stationary.

To test the third condition, the study has applied the ADF test and for this, the the following equation has to be regressed:

$$\Delta u_t = \rho u_{t-1} + \sum_{i=1}^{n} \alpha_i \Delta u_{t-i} + \varepsilon_t \quad (7)$$

where, Δ is the first difference operator; ρ and α_i are the constant unknown parameters; u_{t-1} is the variable of interest; and ε is a stochastic process. The task is to reject the null hypothesis $H_0 : \rho = 1$ against an alternative hypothesis $H_A : \rho \neq 1$.

Before implementing the cointegration and VECM, econometric methodology needs to verify the stationarity of each individual time series since most macro economic data are non-stationary, i.e., they tend to exhibit a deterministic and/or stochastic trend. Though the cointegration approach applies to non-stationary series, it requires that all variables in the system are integrated of the same order I(1). The first step in the analysis is to test for non-stationarity of the data series. Variables that are non-stationary can be made stationary by differencing i.e., the number of differencing (d) required to make the series stationary and that is the order of integration I(d). For the purpose, ADF test was employed to verify the stationarity of the data series and to determine the order of integration of each of the data series studied. Since the selected data series were found to be integrated in an identical order, Johansen's cointegration test was employed to examine long run (cointegrating) relationship among the selected variables.

Vector Error Correction Model (VECM)

Once a single cointegration vector among the selected variables is identified, VECM can be employed to establish the Granger causal direction. VECM allows the modelling of both the short run and long run dynamics for the variables involved in the model. Engle and Granger (1987) show that cointegration is implied by the existence of a corresponding error correction representation which implies that

changes in the dependent variable are a function of the level of the disequilibrium in the cointegrating relationships (captured by error correction term) and changes in other independent variables. According to Granger representation theorem, if variables are cointegrated then their relationships can be expressed as VECM. Since the variables FDI and GDP were cointegrated in the same order, VECM can be written as:

$$\Delta \ln FDI_t = c_1 + \sum_{k=1}^{n} \alpha_{1i} \Delta \ln FDI_{t-k} + \sum_{k=1}^{n} \beta_{2i} \Delta \ln GDP_{t-k} + \rho_1 ECT_{t-k} + \varepsilon_{fdit} \quad (8)$$

$$\Delta \ln GDP_t = c_2 + \sum_{k=1}^{n} \beta_{1i} \Delta \ln GDP_{t-k} + \sum_{k=1}^{n} \alpha_{2i} \Delta \ln FDI_{t-k} + \rho_2 ECT_{t-k} + \varepsilon_{gdpit} \quad (9)$$

where Δ is the first difference operator and ε_{fdit} and ε_{gdpit} are white noise disturbance terms. FDI_t and GDP_t are foreign direct investment and gross domestic product of India at time 't', respectively, and ECT_{t-k} is the lagged error correction term. In terms of the VECM of equations (8) and (9), GDP_t Granger causes FDI_t, if some of the β_{2i} coefficients, $i = 1,2,3,.....n-1$ are not equal to zero and the error coefficient ρ_1 in the equation of FDI flows is significant at conventional levels. Similarly, FDI_t Granger causes GDP_t, if some of the α_{2i} coefficients, $i = 1,2,3,.....n-1$ are not zero and the error coefficient ρ_2 in the equation of GDP is significant at convention levels. These hypotheses can be tested by using either t-tests or F-tests on the joint significance of the lagged estimated coefficients. If both FDI_t and GDP_t Granger cause each other, then there is a feedback relationship between FDI and GDP. The error correction coefficients, ρ_1 and ρ_2 serve two purposes. They are (1) to identify the direction of causality between FDI and GDP and (2) to measure the speed with which deviations from the long run relationship are corrected by changes in the FDI and GDP. On the other hand, if FDI and GDP are not cointegrated, the standard Granger (1969) bivariate causality must be performed without including the error correction term. Since FDI and GDP were cointegrated the study employed VECM.

Lagrange-Multiplier Test

The Lagrange-Multiplier test was done to test for the autocorrelation in the residuals of the vector error-correction model. Estimation, inference, and post-

estimation analysis of VECMs are predicated on the errors' not being autocorrelated. The stata command veclmar implements the LM test for autocorrelation in the residuals of a VECM. The test was performed at lags $j = 1,.....,mlag()$. For each j, the null hypothesis of the test is that there is no autocorrelation at lag j. Consider a VECM without any trend:

$$\Delta y_t = \alpha\beta y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \in_t$$

As long as the parameters in the cointegrating vectors, β, are exactly identified or overidentified, the estimates of these parameters are superconsistent. This implies that the $r*1$ vector of estimated cointegrating relations

$$\hat{E}_t = \hat{\beta} y_t \tag{10}$$

can be used as data with standard estimation and inference methods. When the parameters of the cointegrating equations are not identified, it does not provide consistent estimates of $\hat{E}_t$; in these cases, veclmar exits with an error message.

The VECM above can be rewritten as

$$\Delta y_t = \alpha\hat{E}_t + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \in_t$$

which is just a VAR with $p-1$ lags where the endogenous variables have been first-differenced and is augmented with the exogenous variables $\hat{E}$. The command veclmar fits this VAR and then calls varlmar to compute the LM test for autocorrelation. The above discussion assumes no trend and implicitly ignores constraints on the parameters in α. The other four trend specifications considered by Johansen (1995) complicate the estimation of the free parameters in β but do not alter the basic result that the $\hat{E}_t$ can be used as data in the subsequent VAR. Similarly, constraints on the parameters in α imply that the subsequent VAR must be estimated with these constraints applied, but $\hat{E}_t$ can still be used as data in the VAR.

The formula for the LM test statistic at lag j is

$$LM_s = (T - d - .5)\ln\left(\frac{|\hat{\Sigma}|}{|\tilde{\Sigma}_s|}\right)$$

where T is the number of observations in the VAR; d is the number of coefficients estimated in the augmented VAR; $\hat{\Sigma}$ is the maximum likelihood estimate of Σ, the

variance-covariance matrix of the disturbances from the VAR; and $\tilde{\Sigma}_s$ is the maximum likelihood estimate of Σ from the following augmented VAR.

If there are K equations in the VAR, e_t is a $K*1$ vector of residuals. After K new variables are created $e_1, e_2, ...e_k$ containing the residuals from the K equations, the original VAR can be augmented with the lags of these K new variables. For each lag s, an augmented regression is formed in which the new residual variables are lagged s times. As per the method of Davidson and MacKinnon (1993), the missing values from these s lags are replaced with zeros. $\tilde{\Sigma}_s$ is the maximum likelihood estimate of Σ from this augmented VAR, and d is the number of coefficients estimated in the augmented VAR. var for a discussion of the maximum likelihood estimate of Σin a VAR. The asymptotic distribution of LM_s is χ^2 with K^2 degrees of freedom. The LM test was done in Stata using the post-estimation command veclmar.

Jarque-Bera Test

The Jarque-Bera test is used to test if the disturbances in VECM are normally distributed. It computes and reports a series of statistics against the null hypothesis that the disturbances in a VECM are normally distributed. For the individual equations, the null hypothesis is that the disturbance term in that equation has a univariate normal distribution. For all equations jointly, the null hypothesis is that the K disturbances come from a K-dimensional normal distribution. The stata command vecnorm computes and reports a series of statistics against the null hypothesis that the disturbances in a VECM are normally distributed. As noted by Johansen (1995), the log likelihood for the VECM is derived assuming the errors are independently and identically distributed (i.i.d.) normal, though many of the asymptotic properties can be derived under the weaker assumption that the errors are merely i.i.d. Many researchers still prefer to test for normality.

Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) are a method of measuring the level of collinearity between the regressors in an equation. VIFs show how much of the variance of a coefficient estimate of a regressor has been inflated due to collinearity with the other regressors. They can be calculated by simply dividing the variance of a coefficient estimate by the variance of that coefficient had other regressors not

been included in the equation. There are two forms of the Variance Inflation Factor: centered and uncentered. The centered VIF is the ratio of the variance of the coefficient estimate from the original equation divided by the variance from a coefficient estimate from an equation with only that regressor and a constant. The uncentered VIF is the ratio of the variance of the coefficient estimate from the original equation divided by the variance from a coefficient estimate from an equation with only one regressor (and no constant). If the original equation did not have a constant only the uncentered VIF will be displayed. The centered VIF is numerically identical to $1/\left(1-R^2\right)$where R^2 is the R-squared from the regression of that regressor on all of the other regressors in the equation. Since the VIFs are calculated from the coefficient variance-covariance matrix, any robust standard error options will be present in the VIFs. VIF was computed using the stata command estat vif.

Breusch-Pagan / Cook-Weisberg Test for Heteroscedasticity

Recalling the OLS assumption that $V\left(\varepsilon_j\right)=\sigma^2$ for all j. That is, the variance of the error term is constant, which means there is homoscedasticity. If the error terms do not have constant variance, they are said to be heteroscedastic. If the OLS assumptions are violated it can produce biased and misleading parameter estimates. The Breusch-Pagan test is designed to detect any linear form of heteroscedasticity. Breusch-Pagan / Cook-Weisberg tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. A large chi-square would indicate the presence of heteroscedasticity. In the present study, the Breusch-Pagan test was performed using the stata command estat hettest.

Wooldridge Test for Autocorrelation

Because serial correlation in linear panel-data models biases the standard errors and causes the results to be less efficient, researchers need to identify serial correlation in the idiosyncratic error term in a panel-data model. While a number of tests for serial correlation in panel-data models have been proposed, a test discussed by Wooldridge (2002) is very attractive because it requires relatively few assumptions and is easy to implement. The null hypothesis that there is no first-order autocorrelation was tested using the stata command xtserial. The test is found to have good size and power properties with samples of moderate size. Baltagi (2001) extensively discussed testing for serial correlation in the presence of random and

fixed effects. Many of these tests make specific assumptions about the nature of the individual effects or test for the individual-level effects jointly. Some of these tests, such as the Baltagi–Wu test derived in Baltagi and Wu (1999), are optimal within a class of tests. In contrast, because the Wooldridge test is based on fewer assumptions, it should be less powerful than the more highly parameterised tests, but it should be more robust. While the robustness of the test makes it attractive, it is important to verify that it has good size and power properties under these weaker assumptions (Drukker, 2003).

Consider the linear one-way model,

$$y_{it} = \alpha + X_{it}\beta_1 + Z_i\beta_2 + \mu_i + \in_{it} \qquad i \in \{1,2.....,N\},\ t \in \{1,2,.....,T_i\} \tag{11}$$

where y_{it} is the dependent variable; X_{it} is a $(1 * K_1)$ vector of time-varying covariates; Z_i is a $(1 * K_2)$ vector of time-invariant covariates; α, β_1, and β_2 are $1 + K_1 + K_2$ parameters; μ_i is the individual-level effect; and $\in_{it}$ is the idiosyncratic error. If the μ_i are correlated with the X_{it} or the Z_i, the coefficients on the time-varying covariates X_{it} can be consistently estimated by a regression on the within-transformed data or the first-differenced data. If the μ_i are uncorrelated with the X_{it} and the Z_i, the coefficients on the time-varying and time-invariant covariates can be consistently and efficiently estimated using the feasible generalised least squares method known as random-effects regression. All of these estimators assume that $E[\in_{it} \in_{is}] = 0$ for all $s \neq t$; i.e., that there is no serial correlation in the idiosyncratic errors, which would cause the standard errors to be biased and the estimates to be less efficient.

Wooldridge's method uses the residuals from a regression in first-differences. First-differencing the data in the model removes the individual-level effect, the term based on the time-invariant covariates and the constant,

$$y_{it} - y_{it-1} = (X_{it} - X_{it-1})\beta_1 + \in_{it} - \in_{it-1}$$

$$\Delta y_{it} = \Delta X_{it}\beta_1 + \Delta \in_{it}$$

where Δ is the first-difference operator.

Wooldridge's procedure begins by estimating the parameters β_1 by regressing Δy_{it} on ΔX_{it} and obtaining the residuals $\hat{e}_{it}$. Central to this procedure is Wooldridge's observation that, if the $\in_{it}$ are not serially correlated, then $Corr(\Delta \in_{it}, \Delta \in_{it-1}) = -.5$.

Given this observation, the procedure regresses the residuals $\hat{e}_{it}$ from the regression with first-differenced variables on their lags and tests that the coefficient on the lagged residuals is equal to $-.5$. To account for the within-panel correlation in the regression of $\hat{e}_{it}$ on $\hat{e}_{it-1}$, the VCE is adjusted for clustering at the panel level. Since cluster () implies robust, this test is also robust to conditional heteroscedasticity. This study used the stata command xtserial, which implements the Wooldridge test for serial correlation in panel data.

Fixed Effects Regression with Driscoll-Kraay Standard Errors

Panel (or longitudinal) data are cross-sectional and time-series. There are multiple entities, each of which has repeated measurements at different time periods. Panel data may have group effects, time effects, or both, which are analysed by fixed effect and random effect models. A panel data set contains n entities or subjects (e.g., firms), each of which includes T observations measured at 1 through t time period. Thus, the total number of observations is nT. Ideally, panel data are measured at regular time intervals (e.g., year, quarter, and month). A short panel data set has many entities but few time periods (small T), while a long panel has many time periods (large T) but few entities (Cameron and Trivedi, 2009).

Panel data models examine group (individual-specific) effects, time effects, or both. These effects are either fixed effects or random effects. A fixed effect model examines if intercepts vary across groups or time periods, whereas a random effect model explores differences in error variances. A one-way model includes only one set of dummy variables (e.g., firm), while a two-way model considers two sets of dummy variables (e.g., firm and year). The current study made use of the two-way models considering both year and firm dummies.

In balanced panel data, all entities have measurements in all time periods. In a contingency table of cross-sectional and time-series variables, each cell should have only one frequency. When each entity in a data set has different numbers of observations due to missing values, the panel data are not balanced. Some cells in the contingency table have zero frequency. In unbalanced panel data, the total number of observations is not nT . Unbalanced panel data entail some computational and estimation issues although most software packages are able to handle both balanced and unbalanced data.

Fixed Effect versus Random Effect Models

Panel data models examine fixed and/or random effects of entity (individual or subject) or time. The core difference between fixed and random effect models lies in the role of dummy variables. If dummies are considered as a part of the intercept, this is a fixed effect model. In a random effect model, the dummies act as an error term. A fixed group effect model examines group differences in intercepts, assuming the same slopes and constant variance across entities or subjects. Since a group (individual specific) effect is time invariant and considered a part of the intercept, u_i is allowed to be correlated to other regressors. Fixed effect models use least squares dummy variable (LSDV) and within effect estimation models. Ordinary least squares regressions with dummies, in fact, are fixed effect models. A random effect model, by contrast, estimates variance components for groups (or times) and error, assuming the same intercepts and slopes. u_i is a part of the errors and thus should not be correlated to any regressor; otherwise, a core OLS assumption is violated. The difference among groups (or time periods) lies in their variance of the error term, not in their intercepts. A random effect model is estimated by generalised least squares (GLS) when the Ω matrix, a variance structure among groups, is known. The feasible generalised least squares (FGLS) method is used to estimate the variance structure when Ω is not known. A typical example is the groupwise heteroscedastic regression model (Greene, 2003). There are various estimation methods for FGLS including the maximum likelihood method and simulation (Baltagi and Cheng, 1994). Fixed effects are tested by the (incremental) F test, while random effects are examined by the Lagrange Multiplier (LM) test (Breusch and Pagan, 1980).

Two-way panels

The word panel is derived from Dutch and originally describes a rectangular board. In econometrics, it denotes data sets that have a time dimension as well as a non-time dimension. A genuine panel has the form

$$X_{it}, \quad i = 1,.....,N,\ t = 1,.....,T \tag{12}$$

Thus, it can be represented in a rectangular form, like a board. Dimension i is called the 'individual dimension', t is the time dimension. X can be a scalar (real) variable or also a vector-valued variable. Often, data sets do not correspond exactly to this pattern, even though they have similar dimensions i and t. For example, t may denote an individual time dimension rather than a common time. Such data sets are

sometimes called longitudinal data. It may also occur that there exists a common time index but the physical identity of individuals changes over time. Such data sets are called repeated cross sections (RCS) or pseudo panels. If lengths of time series depend on i, thus violating the rectangular shape, panels are called unbalanced. The dimensions of width and length of the 'board' are not exchangeable coordinates. The time index t is logically ordered, while the individual index is not. Thus, it may make sense to admit a correlation structure over time, for X_{it} and $X_{i,t-1}$, while still assuming independence across individuals X_{it} and $X_{i-1,t}$. The time index t has a quite similar meaning in different panels: it may denote years, days, hours etc. By contrast, the interpretation of the subscript i varies a lot across applications. It may refer to persons, countries, firms, municipalities, trees, lab animals. According to the positioning of the board (portrait or landscape), one may also distinguish time-series panel $(T > N)$ and cross-section panels $(N > T)$.

Two-way Panels - Fixed effects

At first, the following regression model is considered,

$$y_{it} = \alpha + \beta' X_{it} + u_{it}, \quad i = 1,.....,N,\ t = 1,.....,T$$

as the basic model. A third index may be added to the subscripts i and t, as the vector of regressors X_{it} and thus β have dimension K.

If the errors u_{it} are independent across time and across individuals with $Eu = 0$ and $\operatorname{var} u = \sigma^2$, then this is a traditional econometric regression model that can be estimated via OLS. In panel analysis, this usually is a quite restrictive and unrealistic model is called 'pooled regression'. The more common assumption is that regression constants vary across individuals (countries, firms). In this case, many texts on panels use the notation α_i in lieu of α. Alternatively, one may subtract the mean across individuals from α_i and subsume the deviations $\mu_i = \alpha_i - N^{-1}\sum_{i=1}^{N}\alpha_i$ in the disturbances u. Observe $\sum \mu_i = 0$. The individual-specific constants α_i are called effects. With these specifications, the model obtains the form

$$y_{it} = \alpha + \beta' X_{it} + u_{it},$$

$$u_{it} = \mu_i + v_{it}, \quad i = 1,.....,N, \quad t = 1,....., T, \tag{13}$$

where various assumptions can be used for the properties of μ_i and v_{it}. The simplest assumption is that all μ_i become model parameters, together with α and β, which however can only be estimated if T gets large, not for $N \to \infty$. In statistics, such parameters on which information does not increase as the sample size grows are called incidental parameters. An important assumption is $\sum_{i=1}^{N} \mu_i = 0$, otherwise the global intercept α is not identified. This assumption does not represent a restriction but it is necessary in order to make the model empirically valid. Because the μ_i are fixed values, the model is called the 'fixed-effects' model (FE).

In some cases, it may be attractive to extend the panel model by so-called 'time effects':

$$y_{it} = \alpha + \beta' X_{it} + u_{it},$$

$$u_{it} = \mu_i + \lambda_t + v_{it}, \qquad i = 1,.....N, \qquad t = 1,.....T. \tag{14}$$

If the $N+T$ constants μ_i and λ_t are interpreted as parameters, the resulting model is the two-way fixed-effects model. It can be analysed in complete analogy to the one-way fixed-effects model. Firstly, it can be rewritten in compact notation

$$y = \alpha + X\beta + Z_\mu \mu + Z_\lambda \lambda + v \tag{15}$$

Matrices Z_μ and Z_λ have dimensions $TN * N$ and $TN * T$. $\sum_{t=1}^{T} \lambda_t = \sum_{i=1}^{N} \mu_i = 0$ must be assumed, otherwise α is not identified.

For uncorrelated errors v, the OLS estimator is BLUE. This estimator, however, requires inverting a matrix of dimension $(K+N+T-1)*(K+N+T-1)$ [avoiding the dummy trap; alternatively a constrained regression with an inversion of a matrix of dimension $(K+N+T+1)*(K+N+T+1)$] which should be avoided. In analogy to the one-way model, the Frisch-Waugh theorem can be applied. One may first 'purge' all individual and temporal means from all variables y and X, and then execute an OLS regression that requires inversion of a $K * K$) – matrix only.

Formally, the purging matrix is given as

$$Q_2 = I_N \otimes I_T - I_N \otimes \bar{J}_T - \bar{J}_N \otimes I_T + \bar{J}_N \otimes \bar{J}_T \tag{16}$$

The matrix $I_N \otimes \bar{J}_T$ is a block-diagonal and subtracts time averages for each i. The matrix $\bar{J}_N \otimes I_T$ contains $N * N$ identical blocks and subtracts for each time point t

means across individuals. The last term is a matrix filled with $NT * NT$ times the value $(NT)^{-1}$.

With this two-way Q_2, the FE estimator can be expressed as

$$\hat{\beta} = \left(X' Q_2 X\right)^{-1} X' Q_2 y \tag{17}$$

Again, formally the FE estimator is reminiscent of a GLS estimator, although Q_2 is singular and cannot be the inverse of an errors variance matrix. Similarly, the variance matrix for the estimator is given as

$$\operatorname{var} \hat{\beta} = \sigma_v^2 \left(X' Q_2 X\right)^{-1}, \tag{18}$$

where σ_v^2 denotes the variance of the stochastic error v_{it}. The FE estimator is consistent for β and α, for $\mu(\lambda)$ only if $T \to \infty (N \to \infty)$.

Random Effects - The GLS estimator for the RE Model

If N becomes large, such as in the typical microeconometric cross-section panels, the number of estimated parameters of the FE model increases considerably. This motivates the idea of viewing μ_i not as parameters, but rather as unobserved variables with mean 0 and variance σ_μ^2. For small N, it is less plausible to assume that individual characteristics have been generated randomly.
The random effects (RE) model can be written as

$$y_{it} = \alpha + \beta' X_{it} + u_{it},$$

$$u_{it} = \mu_i + v_{it},$$

$$\mu_i \sim i.i.d.\left(0, \sigma_\mu^2\right),$$

$$v_{it} \sim i.i.d.\left(0, \sigma_v^2\right),$$

$$i = 1,....., N, \quad t = 1,.....,T \tag{19}$$

The two error components μ and v are assumed to be independent from each other. u_{it} follows a mean-zero probability distribution for all i and t, therefore the RE model is a 'regular' GLS model. If both variances σ_μ^2 and σ_v^2 are known, α and β can be estimated efficiently, with normal errors efficiently among all unbiased estimators, otherwise BLUE, via the GLS estimator

$$\left\{ X_\#' \left(Euu'\right)^{-1} X_\# \right\}^{-1} X_\#' \left(Euu'\right)^{-1} y$$

A drawback for the direct application of this method is the large matrix $\Omega = Euu'$, which has dimension $NT * NT$ and must be inverted. The dummy problem does not appear in the RE model by construction, as the effects are specified to be stochastic with zero mean. Therefore, all regression can be conducted on the basis of the $(NT * (K+1))$ - matrix of regressors $X_{\#}$ that has been extended by a column of ones for the overall intercept.

Two-way Panels - Random Effects

The two-way random-effects model assumes that the error components μ_i and λ_t are drawn from independent distributions with variances σ_μ^2 and σ_λ^2. The variance matrix of the total unobserved errors u results as the sum of its three components (using the explicit notation $\sigma_v^2 = Ev_{it}^2$:

$$\begin{aligned}\Omega &= E\left(uu'\right) \\ &= \sigma_\mu^2 (I_N \otimes J_T) + \sigma_\lambda^2 (J_N \otimes I_T) + \sigma_v^2 I_{NT}\end{aligned} \tag{20}$$

Only the middle term is new as compared to the one-way model. It contains $N * N$ blocks of I_N identity matrices. The three terms are not orthogonal to each other and Ω cannot be inverted on the basis of these components. The orthogonal decomposition is not as easily found as in the one-way model. It can be constructed by tentatively specifying the inverse Ω^{-1} as a weighted sum of four plausible matrices, i.e. the three components of Ω above and the $NT * NT$ matrix of ones J_{NT}:

$$\Omega^{-1} = a_1 (I_N \otimes J_T) + a_2 (J_N \otimes I_T) + a_3 I_{NT} + a_4 J_{NT}$$

Equating $\Omega\Omega^{-1} = I$ yields the coefficients a_j by a comparison of coefficients. The results are reported here:

$$a_1 = -\frac{\sigma_\mu^2}{\left(\sigma_v^2 + T\sigma_\mu^2\right)\sigma_v^2},$$

$$a_2 = -\frac{\sigma_\lambda^2}{\left(\sigma_v^2 + N\sigma_\lambda^2\right)\sigma_v^2},$$

$$a_3 = \frac{1}{\sigma_v^2},$$

$$a_4 = \frac{\sigma_\mu^2 \sigma_\lambda^2}{\sigma_v^2 \left(\sigma_v^2 + T\sigma_\mu^2\right)\left(\sigma_v^2 + N\sigma_\lambda^2\right)} \frac{2\sigma_v^2 + T\sigma_\mu^2 + N\sigma_\lambda^2}{\sigma_v^2 + T\sigma_\mu^2 + N\sigma_\lambda^2} \tag{21}$$

The expression for a_4 is a bit involved. For $N,T \to \infty$ the matrix Ω^{-1} converges to the matrix Q_2 that was found for the two-way FE model in (16). It is easily checked directly that $Ta_1(T) \to a_3, Na_2(N) \to a_3, NTa_4(N,T) \to a_3$, using some plausible notation. Thus, the GLS estimator approaches the FE estimator. The exact proof that for (a) $N \to \infty$, (b) $T \to \infty$, (c) $N/T \to c \in \Re \setminus \{0\}$ both estimators are asymptotically equivalent, including all distributional properties.

Using estimates for all variances $\sigma_\lambda^2, \sigma_\mu^2, \sigma_v^2$ yields an estimator $\hat{\Omega}^{-1}$ for the variance matrix Ω^{-1}. Then, the expression for the feasible GLS estimator for the two-way RE-models reads:

$$\hat{\beta}_{\#,RE} = \left(X_\#' \hat{\Omega}^{-1} X_\#\right)^{-1} X_\#' \hat{\Omega}^{-1} y$$

The required variance estimates for the error components can be computed from the (two-way) FE regression residuals. By iteration, one may again approximate the ML estimator (Kunst, 2011).

The stata command xtgls fits panel-data linear models by using feasible generalised least squares. This command allows estimation in the presence of AR(1) autocorrelation within panels and cross-sectional correlation and heteroscedasticity across panels. The command xtgls will estimate a model by feasible generalised least squares (FGLS) under the assumption that all aspects of the model are completely specified. Here that includes that the disturbances have different variances for each panel and are constant within panel. Under these assumptions, FGLS is asymptotically efficient and if iterated will produce maximum likelihood estimates of the parameters. FGLS is preferred over OLS under heteroscedasticity or serial correlation (Baltagi, 2008). FGLS might be inconsistent is if there are individual specific fixed effects. Hence the investigator used FGLS regression when the Hausman test rejected the use of fixed effects model.

Compared with purely cross-sectional data, panels are attractive since they often contain far more information than single cross-sections and thus allow for an increased precision in estimation. Unfortunately, however, actual information of micro-econometric panels is often overstated since micro-econometric data are likely to exhibit all sorts of cross-sectional and temporal dependencies. Erroneously

ignoring possible correlation of regression disturbances over time and between subjects can lead to biased statistical inference. To ensure validity of the statistical results, most studies that include a regression on panel data therefore adjust the standard errors of the coefficient estimates for possible dependence in the residuals. Furthermore, although most empirical studies now provide standard error estimates that are heteroscedasticity and autocorrelation consistent, cross-sectional or "spatial" dependence is still largely ignored. However, assuming that the disturbances of a panel model are cross-sectionally independent is often inappropriate. Provided that the unobservable common factors are uncorrelated with the explanatory variables, the coefficient estimates from standard panel estimators, e.g., fixed-effects (FE) estimator, random-effects (RE) estimator, or pooled ordinary least squares (OLS) estimation are still consistent but inefficient. However, standard error estimates of commonly applied covariance matrix estimation techniques, e.g., OLS, White, and Rogers or clustered standard errors, are biased, and hence statistical inference based on such standard errors is invalid.

Fortunately, Driscoll and Kraay (1998) propose a nonparametric covariance matrix estimator that produces heteroscedasticity and autocorrelation-consistent standard errors that are robust to general forms of spatial and temporal dependence. Hoechle (2007) aimed to provide a Stata implementation of Driscoll and Kraay's (1998) covariance matrix estimator for use with pooled OLS estimation and FE regression. In contrast to Driscoll and Kraay's original contribution that considers only balanced panels Hoechle (2007) used their estimator for use with unbalanced panels and use Monte Carlo simulations to investigate the adjusted estimator's finite sample performance in case of medium and large-scale (micro-econometric) panels. Consistent with Driscoll and Kraay's original finding for small balanced panels, the Monte Carlo experiments reveal that erroneously ignoring spatial correlation in panel regressions typically leads to overly optimistic (anticonservative) standard error estimates, irrespective of whether a panel is balanced. Although Driscoll and Kraay standard errors tend also to be slightly optimistic, their small-sample properties are considerably better than those of the alternative covariance estimators when cross-sectional dependence is present. The xtscc program in Stata produces Driscoll and Kraay standard errors for coefficients estimated by pooled OLS/weighted least-squares (WLS) regression and FE (within) regression.

Hausman Test

Hausman is a general implementation of Hausman's (1978) specification test, which compares an estimator $\hat{\theta}_1$ that is known to be consistent with an estimator $\hat{\theta}_2$ that is efficient under the assumption being tested. The null hypothesis is that the estimator $\hat{\theta}_2$ is indeed an efficient (and consistent) estimator of the true parameters. If this is the case, there should be no systematic difference between the two estimators. If there exist a systematic difference in the estimates, there are reasons to doubt the assumptions on which the efficient estimator is based.

The Hausman statistic is distributed as χ^2 and is computed as

$$H = (\beta_c - \beta_e)'(V_c - V_e)^{-1}(\beta_c - \beta_e)$$

where

β_c is the coefficient vector from the consistent estimator

β_e is the coefficient vector from the efficient estimator

V_c is the covariance matrix of the consistent estimator

V_e is the covariance matrix of the efficient estimator

When the difference in the variance matrices is not positive definite, a Moore-Penrose generalised inverse is used. The number of degrees of freedom for the statistic is the rank of the difference in the variance matrices. When the difference is positive definite, this is the number of common coefficients in the models being compared.

The Hausman specification test compares the fixed versus random effects under the null hypothesis that the individual effects are uncorrelated with the other regressors in the model (Hausman, 1978). If correlated (H_0 is rejected), a random effect model produces biased estimators, violating one of the Gauss-Markov assumptions; so a fixed effect model is preferred. Hausman's essential result is that the covariance of an efficient estimator with its difference from an inefficient estimator is zero (Greene, 2003).

$$m = \left(b_{Robust} - b_{Efficient}\right)'\hat{\Sigma}^{-1}\left(b_{Robust} - b_{Efficient}\right) \sim \chi^2(k),$$

where, $\hat{\Sigma} = Var\left[b_{Robust} - b_{Efficient}\right] = Var\left(b_{Robust}\right) - Var\left(b_{Efficient}\right)$ is the difference in the estimated covariance matrix of the parameter estimates between the LSDV model (robust) and the random effects model (efficient). It is notable that an intercept

and dummy variables should be excluded in computation. When comparing fixed effect and random effects models, the fixed effect estimates are considered as the robust estimates and random effect estimates as the efficient estimates (Park, 2009).

There are two statistical formulations of panel data model commonly used in the empirical literature. These are known as fixed effects (FE) or least squares dummy variables (LSDV) model and random effects (RE) model. Traditionally, the way to choose between these two models is to employ the Hausman specification (HS) test, which measures the distance between the estimated FE and RE coefficients. If the Hausman statistic is very large (that shows there are systematic differences between the coefficients of these two models), the random effects can be easily rejected in favour of the fixed effects specification (Joseph, 2007).

Heckman Selection Maximum Likelihood Method

The study has attempted to analyse the export spillovers from the foreign firms. It is considered that exporting activities involve a two stage decision process: (i) the firm decides whether to export or not and, (ii) then the amount that it is willing to export (export intensity). Therefore, to take into account the two stage process, the study adopted the standard Heckman selection model (Heckman 1979). The presence of foreign multinationals affects the export decision behaviour of all domestic firms, not only exporting domestic firms (Greenaway et al., 2004). Therefore, if functions of export participation decision and export intensity decision of a firm are separately estimated, the problem of selection bias rises. Thus, jointly estimating the export intensity and the export propensity functions can avoid the sample selection bias. The equations involving export participation decision of domestic firms and their export intensity are estimated using the Heckman selection maximum likelihood methodology instead of the Heckman selection two step procedure since the maximum likelihood method is more efficient (Kneller and Pisu, 2007). The econometric analysis involving a two-stage decision process, using Heckman's maximum likelihood estimation, control for selection bias.

The Heckman selection model assumes that there exists an underlying regression relationship,

$$y_j = x_j\beta + u_{1j} \text{ regression equation}$$

The dependent variable, however, is not always observed. Rather the dependent variable for observation j is observed if

$z_j\gamma + u_{2j} \rangle 0$ selection equation

where

$$u_1 \sim N(0,\sigma)$$

$$u_2 \sim N(0,1)$$

$$corr(u_1, u_2) = \rho$$

when $\rho \neq 0$, standard regression techniques applied to the first equation yield biased results. Heckman selection provides consistent, asymptotically efficient estimates for all the parameters in such models (Gronau, 1974; Lewis, 1974; Heckman,1976).

Regression estimates using the nonselection hazard (Heckman, 1979) provide starting values for maximum likelihood estimation.

The regression equation is

$$y_j = x_j\beta + u_{1j}$$

The selection equation is

$$z_{j\gamma} + u_{2j} \rangle 0$$

where

$$u_1 \sim N(0,\sigma)$$

$$u_2 \sim N(0,1)$$

$$corr(u_1, u_2) = \rho$$

The log likehood for observation j, $\ln L_j = l_j$, is

$$l_j = \begin{cases} w_j \ln \Phi\left\{ \dfrac{z_j\gamma + (y_j - x_j\beta)\rho/\sigma}{\sqrt{1-\rho^2}} \right\} \dfrac{w_j}{2}\left(\dfrac{y_j - x_j\beta}{\sigma}\right)^2 - w_j \ln\left(\sqrt{2\Pi\sigma}\right) \\ w_j \ln \Phi(-z_j\gamma) \end{cases}$$

Where $\Phi(\cdot)$ is the standard cumulative normal and w_j is an optional weight for observation j.

In the maximum likelihood estimation, σ and ρ are not directly estimated. Directly estimated are $\ln\sigma$ and atanh ρ:

$$a\tanh\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

The standard error of $\lambda = \rho\sigma$ is approximated through the propagation of error (delta) method; that is,

$$Var(\lambda) \approx DVar\{(a\tanh\rho \ln\sigma)\}D'$$

where D is the Jacobian of λ with respect to atanh ρ and $\ln\sigma$.

With maximum likelihood estimation, this command supports the Huber/White/sandwich estimator of the variance and its clustered version using vce (robust) and vce (cluster clustvar), respectively.

All the analysis in the study was done using the software packages Eviews version 7 and Stata version 12.

VI. Limitations of the Study

Trends, determinants and the causal nexus between FDI and economic growth were studied using the time series data for the period 1990-91 to 2012-13. The investigator was interested in the performance of FDI after globalisation because after the economic reforms many sectors were opened up for FDI and the quantum of FDI inflows surged since then. If the study had used FDI data prior to 1990-91 the results could have improved due to longer time series data. Also the firm level panel data for the Indian manufacturing data were used from 2000-01 to 2012-13 owing to the non-availability of firms' equity holding information prior to 2000-01. Also, the study is based on the secondary data which has its own limitations. The accuracy of the results presented in the study depends on the reliability of the data sources.

One important methodological issue is what threshold of foreign equity holding should be used to define a foreign firm or foreign acquisition. In studies undertaken on Indian manufacturing firms, the cut-off level has commonly been taken as 10 per cent. But, Arnold and Javorcik (2009), on contrast, have used the cut-off level 20 per cent. One interesting empirical question that arises here is that defining foreign acquisition or treatment at 10 per cent foreign equity may yield different results than the alternative option of defining foreign acquisition or treatment at 20 per cent foreign equity. This issue has not been investigated in this study and left for future research. For the analysis undertaken in this study, the threshold level of foreign equity participation has been taken as 10 per cent. Despite these limitations, the conclusions drawn from the findings of the study will add to the existing literature and provides scope for further research in future.