CHAPTER – IV
DATA BASE AND RESEARCH METHODOLOGY

Present study is an investigation of barriers to entry in Indian manufacturing sector. This chapter seeks to discuss the sources of data, structure of data, underlying variables, methodology used for the analysis and limitations of the study.

Sources of Data and Data Structure

To study barriers to entry in the Indian manufacturing sector there is a need to get data on entry of firms at the industry level. Structured data on entry of firms' for each industry is not readily available in India. Consequently, for our research purposes various reports on Industry Financial Aggregates & Ratios published by the Centre for Monitoring Indian Economy (CMIE) have been utilized to derive various measures of entry. These reports of CMIE have also been used for the measurement of the independent variables at the industry level. Data has been collected for the period 1990 to 2005 at the industry level. CMIE classifies industry into four major groups: manufacturing, services, mining and electricity. This research focuses on the analysis of entry barriers related to the manufacturing sector. So, we have taken the data on manufacturing sector as per the classification of manufacturing sector by CMIE. The manufacturing sector as per CMIE comprises of 17 major industry groups and further classified into 98 sub-groups. Detailed listing of the 98 industry sub-groups is available in Annexure – I. Accordingly, in this study of the barriers to entry in manufacturing sector, the sector classified into these 98 industry sub-groups of the manufacturing sector has been analyzed. Number of entrants compiled by us relates to the period 1993 to 2005. Data on industry characteristics like sales, capital employed, advertising expenditure, and profits etc. has also been taken from the CMIE. These industry characteristics have been used for the measurement of independent variables. The study was conducted for a period of twelve years (1993-2005) and across 98 industry groups in the manufacturing sector in India. So we have 1176 (98 industries x 12 years) data points. Panel data has been used for the econometric estimation.
Panel Data

This is a special type of pooled data in which the same cross-sectional unit (say, an industry, a firm or a country) is surveyed over time. There are other names for panel data, such as pooled data (pooling of time series and cross-sectional observations), combination of time series and cross-section data, micropanel data and longitudinal data.

Why Panel Data?

The advantages of panel data over cross-section or time series data are explained by Baltagi (1995). These are listed as follows:

- Since panel data relate to individuals, industries, firms, countries, etc., over time, there is bound to be heterogeneity in these units. The techniques of panel data estimation can take such heterogeneity explicitly into account by allowing for individual-specific variables. We use the term individual in a generic sense to include micro units such as individuals, industries, firms, and countries.
- By combining time series of cross-section observations, panel data give “more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency.”
- By studying the repeated cross section of observations, panel data are better suited to study the dynamics of change.
- Panel data can better detect and measure effects that simply cannot be observed in pure cross-section or pure time series data.
- Panel data enables us to study more complicated models. For example, phenomena such as economies of scale and technological change can be better handled by panel data than by pure cross-section or pure time series data.
- By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.
In brief, panel data can enrich empirical analysis in ways that may not be possible if we use only cross-section or time series data.

**Functional Relationship and Analytical Techniques**

For the identification of the functional relationship, research on the barriers to entry of firms has used two main approaches. One is the indirect approach, in which profitability as a dependent variable is explained by factors like concentration, economies of scale, advertising intensity etc. Prominent among these studies have been Comanor and Wilson (1967) and Miller (1969). Most econometric investigations of entry barriers using this approach have only provided indirect tests. In this method of study, profitability rather than entry is used as a dependent variable and is regressed on those structural factors considered to raise barriers to entry. The main reason for adopting this methodology has been the non-availability of data on 'entry', and hence profitability is used as a substitute with the assumption that high profitability attracts large entry. Critics have argued against this approach on several grounds. Firstly, the dependent variable is a proxy measure, secondly, there may be a gap between true and measured profit and lastly, related measurement errors of profit may enhance the unreliability of estimates. Given the availability of data, it makes sense to use 'entry' as the dependent variable instead of using profitability. This approach referred to as the direct approach, better serves the objective of analyzing the barriers to entry, as it does not assume direct relationship between profitability and entry. It was Orr (1974) who first pointed out the shortcomings of the indirect approach and since then given the availability of data on entry, direct approach has been the preferred methodology in the literature. In this approach 'entry' rather than rate of profit is regressed on factors influencing entry.

Most studies fail to recognize the possibility to estimate the barriers to entry separately for the 'incidence' and 'extent' of entry. In case firms may be able to enter if entry is profitable but entry barriers are high, as a result, the influence of factors affecting entry may be different on the 'incidence' of entry vis-a-vis its 'extent'. A comparison of these two influences can provide insights into entry. For instance, barriers related to
capital availability and its costs are seen as important barriers to entry. However, capital intensity might not act as a significant barrier to the incidence of entry, as the firm because of economies of scale might be able to pool in resources with better bargaining power and so on. But capital intensity may continue to be a significant barrier for most entrants, as they may not have adequate access to the capital markets. This will result in lower number of entrants or to say lower 'extent' of entry. Extending the same logic, barriers to "incidence" may differ from those of "extent" of entry.

**The Entry Model and Tools of Analysis**

Entry and its barriers in the form of explanatory variables have been discussed earlier in Chapter - II. Now we discuss various models of entry estimated by us and the analytical tools used in this study. The specification of the dependent variable "entry" is not complete. To address the identified gap in the literature we distinguish between the "extent" and "incidence" of entry. Extent of entry is simply defined as the total number of entrants across industries over a period of time. Most studies of entry use the 'extent' specification. To it we have added another specification, which is referred to as 'incidence' of entry in a category. It is operationalised as dummy dependent variable that takes the value of "1" if entry in a specific category has taken place; else it is equal to "0". Thus, the incidence of entry captures the influence of the explanatory variables on the occurrence of the event called "entry". As mentioned earlier, the main purpose of differentiating "incidence" of entry from "extent" of entry is to explore if the two are differently influenced by the barriers to entry. For example, in an extreme situation while high levels of product differentiation in an industry may deter large-scale entry due to high resource requirements, single players may use the product differentiation route to enter the industry. Because of differences in their basic endowments they have different capabilities, which may help them to enter an industry, but this might not be the case for a larger number of players. Following this argument, different variables capturing barriers to entry may have a more significant negative effect on the 'extent of entry' than on the 'incidence of entry'.
As discussed, the set of industry level determinants explored in our analysis include: concentration ratio (CR), size (SIZE), minimum efficient scale (MES), capital intensity (CI), advertising intensity (AI), marketing and distribution intensity (DMI), export intensity (EI), return on capital employed (ROCE), industry risk (IR), vertical integration (VI) and growth (GR). All these are measured at the industry level.

The model to explore barriers to entry builds on the broad factors identified by studies using the S-C-P framework. To facilitate discussion, the explanatory variables have been categorized into three groups: Structure, Conduct and Performance related factors.

Overall, therefore, we have two basic models as:

Number of entrants = F \{S (CR, SIZE, MES, CI, VI), C (AI, DMI, EI), P (ROCE, IR, GR)\}

Incidence of entry = F \{S (CR, SIZE, MES, CI, VI), C (AI, DMI, EI), P (ROCE, IR, GR)\}

Incidence value is 1 for a specific year if entry has occurred in an industry else it is 0. The number of entrants is the sum total of all of entries in an industry group in a specific year. The probit model is used to estimate the "incidence" version of the model while pooled regression and tobit models are used to estimate the model on the "extent" of entry.

All the explanatory variables are measured with lags. It is presumed that the decision by an entrant to enter an industry is based on the observed industry characteristics over a period of time prior to entry. In other words, entry by firms in a specific year is based on the industry characteristics for a few periods prior to the period of entry. In this study it is assumed that a period of three years is a reasonable time frame for observing the industry characteristics by a potential entrant for her decision-making. Hence, all the explanatory variables are measured as simple three year averages with a one year starting lag from the year of entry.
Methods of Estimation

This study analyzes two measures of entry - extent and incidence. Now, we discuss the econometric methodology used for estimating the empirical models.

Extent of Entry

We have used two different econometric techniques to measure extent of entry.

1. Pooled Regression or Panel Regression

There are several regression techniques using panel data. The two most prominent are (1) the fixed effects model (FEM) and (2) the random effects model (REM) or error components model (ECM).

In FEM the intercept in the regression model is allowed to differ among individuals in recognition of the fact that each individual, or cross-sectional, unit may have some special characteristics of its own. The FEM using panel data is known as pooled regression. To take into account the differing intercepts, one can use dummy variables. The FEM using dummy variables is known as the least-squares dummy variable model. FEM is appropriate in situations where the individual-specific intercept may be correlated with one or more regressors. A disadvantage of least-squares dummy variable model is that it consumes a lot of degrees of freedom when the number of cross-sectional units, $N$, is very large, in which case we will have to introduce $N$ dummies.

An alternative to FEM is ECM. In ECM it is assumed that the intercept of an individual unit is a random drawing from a larger population with a constant mean value. The individual intercept is then expressed as a deviation from this constant mean value. One advantage of ECM over FEM is that it is economical in degrees of freedom, as we do not have to estimate $N$ cross-sectional intercepts. We need only to estimate the mean value of the intercept and its variance. ECM is appropriate in situations where the (random) intercept of each cross-sectional unit is uncorrelated.
with the regressors. The Hausman test can be used to decide between FEM and ECM. The null hypothesis underlying the Hausman test is that the FEM and ECM estimators do not differ substantially. The test statistic developed by Hausman has an asymptotic $\chi^2$ distribution. If the null hypothesis is rejected, the conclusion is that ECM is not appropriate and that we may be better off using FEM. Keeping this fundamental difference in the two approaches in mind, what more can we say about the choice between FEM and ECM? Here the observations made by Judge et al. (1985) may be helpful:

- If $T$ (the number of time series data) is large and $N$ (the number of cross-sectional units) is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.
- When $N$ is large and $T$ is small, the estimates obtained by the two methods can differ significantly. In case of FEM, statistical inference is conditional on the observed cross-sectional units in the sample. This is appropriate if we strongly believe that the individual, or cross-sectional, units in our sample are not random drawings from a larger sample. In that case, FEM is appropriate. However, if the cross-sectional units in the sample are regarded as random drawings, then ECM is appropriate, for in that case statistical inference is unconditional.
- If the individual error component and one or more regressors are correlated, then the ECM estimators are biased, whereas those obtained from FEM are unbiased.
- If $N$ is large and $T$ is small, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.

Out of the various specifications of the entry equation, we choose to follow the semi-logarithmic specification similar to Khemani and Shapiro (1986) as it is found to be the most appropriate one after testing for different functional forms. Thus, our estimation model is
\[ Y_{it} = \beta_1 + \beta_2 X_{1it} + \beta_3 X_{2it} + u_{it} \]

\( Y_{it} \) is defined as the total number of entries in \( i^{th} \) industry in \( t^{th} \) time period. In the literature, this model is known as the fixed effects regression model. The term “fixed effects” is due to the fact that, although the intercept may differ across individuals (here 17 major groups of industries), each individual’s intercept does not vary over time; that is, it is time invariant.

Where, \( Y_{it} = \log (\text{ENT})_{it} \), in this specification \( \text{ENT} \) is the number of entrants plus one. One was added to the number of entrants in each industry and logarithms were then taken., \( X_{1it} = \text{industry size} = \log (\text{industry Sales}) \) and \( X_{2it} \) are the vector of other explanatory variables.

We will also analyze the results with a better econometric technique to estimate extent of barriers to entry. Because this method creates biases as entry may not have occurred in many industries. As a result, there are a large number of 'zeros' on the left-hand side of the estimated equation. A censored regression approach is required in such a situation to analyze the 'extent' of entry, which has been used in recent years (see Geroski, 1991, Saha, 2001). Similarly, analysis of the factors determining 'incidence' of entry will require a limited dependent variable approach. The present study has used appropriate econometric techniques to take care of the problems in the current literature.

2. **Tobit Model**

Extent of entry (\( Y_{it} \)) is defined as the total number of entries in \( i^{th} \) industry in \( t^{th} \) time period. Since, however, there were no entries in a large number of cases; the dependent variable is censored at zero. To avoid censoring bias, this model is estimated using the Tobit procedure. The Tobit model is usually presented in terms of an index or latent variable. More specifically,
\[ Y_{it} = \beta' X_{it} + u_{it} \]

\[ Y_{it} = 0 \text{ if } Y^*_{it} \leq 0 \]

\[ Y_{it} = Y^*_{it} \text{ if } Y^*_{it} \geq 0, \]

where \( Y^*_{it} \) is the index or latent variable, \( Y_{it} \) is the observed dependent variable, \( \beta \) is a \((K \times 1)\) vector of unknown parameters; \( X_{it} \) is a \((K \times 1)\) vector of explanatory variables; \( u_{it} \) are residuals that are independently and normally distributed, with mean \( \mu \) and variance \( \sigma^2 \). Given the nature of the model, it is possible to obtain three conditional mean functions in this case

\[
E(Y^*_{it} / X_{it}) = \beta' X_{it}
\]

\[
E(Y^*_{it} / X_{it}, Y_{it} > 0) = \beta' X_{it} + \sigma \lambda_{it},
\]

\[
E(Y_{it} / X_{it}) = \Phi(\beta' X_{it} / \sigma) (\beta' X_{it} + \sigma \lambda_{it})
\]

where \( \lambda_i = \varphi(\beta' X_{it} / \sigma) / \Phi(\beta' X_{it} / \sigma) \), and \( \varphi \) and \( \Phi \) are the pdf and cdf of standard normal distribution. The elasticity estimates presented in this study are estimated using the conditional mean function given by \( E(Y_{it} / X_{it}) \)

**Incidence of Entry – Probit Model**

In this model, the dependent variable takes the value 'one' in case the number of entries is greater than zero, else it takes the value zero. Thus the model is:

\[ Y_{it} = \beta' X_{it} + u_{it} \]

Where \( Y_{it} = 1 \) if number of entries is greater than zero

\[ = 0 \text{ otherwise} \]

where \( \beta \) is a \( k \times 1 \) vector of unknown parameters; \( X_{it} \) is a \( k \times 1 \) vector of known constants; \( u_{it} \) are residuals that are independently and normally distributed, with mean
zero and variance $\sigma^2$. This model is estimated using the 'Probit' method. In this model, the conditional expectation of the dependent variable is interpreted as the probability of occurrence of the event. The elasticity of the probability with respect to the independent variable $X_{it}$ is calculated as $(\partial P_{it}/\partial X_{it})^*(X_{it}/P_{it})$, where $P_{it}$ are the estimated probabilities.

The result of the pooled regression, tobit and probit models will provide us with the direction of relationship between 'entry' and the various explanatory variables.

**Measurement of Explanatory Variables**

In this study it is assumed that a period of three years is a reasonable time frame for observing the industry characteristics by a potential entrant for her decision-making. Hence, all the explanatory variables are measured as simple three year averages with a one year starting lag from the year of entry. The details are given below.

**Size (SIZE)**

$$\text{Size}_{KT} = \log \left[ \frac{\sum S_{i,T-1} + \sum S_{i,T-2} + \sum S_{i,T-3}}{3} \right]$$

where $K = 1$ to $121$; $T = Ti$ to $T_{12}$; $S_i$ is the sales of the 'i' th firm in the 'K' th industry; and $\text{Size}_{KT}$ is the measure for the 'Size' of the 'K' th industry for the 'T' th year.

**Minimum Efficient Scale (MES)**

$$\text{MES}_{KT} = \left[ \frac{\log (\sum S_i/N)_{T-1} + \log (\sum S_i/N)_{T-2} + \log (\sum S_i/N)_{T-3}}{3} \right]$$

where $K = Ki$ to $K_{12}$; $T = Ti$ to $T_{12}$; $S_i$ is the sales of the 'i' th firm in the 'K' th industry; $N$ is the number of firms in the 'K' th industry; and $\text{MES}_{KT}$ is the measure for the minimum efficient scale of operation of the 'K' th industry for the 'T' th year.
Concentration Ratio (CR)

\[ CR_{KT} = \frac{C_{T-1} + C_{T-2} + C_{T-3}}{3} \]

where 'C' is the Herfindahl Index = \( \sum (S_i / \sum S_i)^2 \)

where \( S_i \) is the sales of the 'i' th firm in the 'K' th industry; \( N \) is the number of firms in the 'K' th industry.

Capital Intensity (CI)

\[ CI_{KT} = \frac{\sum_i \sum_i \text{Capital Employed}_{it}}{\sum_i \sum_i \text{Sales}_{it}} / 3 \]

Where \( K = K_1 \) to \( K_{121} \); \( T = T_1 \) to \( T_{12} \); and 'i' stands for a firm in the industry.

Vertical Integration (VI)

\[ VI_{KT} = \frac{\sum_i \sum_i \text{Gross Value Added}_{it}}{\sum_i \sum_i \text{Sales}_{it}} / 3 \]

Where \( VI_{KT} \) is the vertical integration for the 'K' th industry for the 'T' th year; \( K = K_i \) to \( K_{121} \); \( T = T_1 \) to \( T_{12} \); and 'i' stands for a firm in the industry.

Advertising Intensity (AI)

\[ AI_{KT} = \frac{\sum_i \sum_i \text{Advertising Expenditures}_{it}}{\sum_i \sum_i \text{Sales}_{it}} / 3 \]
Where $\text{AI}_{KT}$ is the advertising intensity for the 'K' th industry for the 'T' th year;
Where $K = K_1$ to $K_{121}$; $T = T_1$ to $T_{12}$; and 'i' stands for a firm in the industry.

**Distribution & Marketing Intensity (DMI)**

$$
\text{DMI}_{KT} = \frac{\sum_t \sum_i \text{Distribution & Marketing Expenditure}_{it}}{\sum_t \sum_i \text{Sales}_{it}} / 3
$$

Where $\text{DMI}_{KT}$ is the distribution & marketing intensity for the 'K' th industry for the 'T' th year; Where $K = K_1$ to $K_{121}$; $T = T_1$ to $T_{12}$; and 'i' stands for a firm in the industry.

**Export Orientation (EI)**

$$
\text{EI}_{KT} = \frac{\sum_t \sum_i \text{Exports}_{it}}{\sum_t \sum_i \text{Sales}_{it}} / 3
$$

Where $\text{EI}_{KT}$ is the export intensity for the 'K' th industry for the 'T' th year; Where $K = K_1$ to $K_{121}$; $T = T_1$ to $T_{12}$; and 'i' stands for a firm in the industry.

**Return on Capital Employed (ROCE)**

$$
\text{ROCE}_{KT} = \frac{\sum_t \sum_i \text{PBIT}_{it}}{\sum_t \sum_i \text{Capital Employed}_{it}} / 3
$$

Where $\text{ROCE}_{KT}$ is the returns on the capital employed for the 'T' th year; Where $K = K_1$ to $K_{121}$; $T = T_1$ to $T_{12}$; and 'i' stands for a firm in the industry.
**Industry Risk (IR)**

Industry risk is measured as the standard deviation of the industries ROCE over a period of five years with a starting lag of one year.

\[
IR_{KT} = \text{Std Dev. (ROCE}_{T,j})
\]

Where ROCE is the industries returns on capital employed; \( j = 1 \) to 5; \( K = 1 \) to 121; \( T = T_1 \) to \( T_{12} \); and \( IR_{KT} \) is the industry risk for the 'K' th industry for the 'T' th year.

**Growth (GR)**

Growth of the industry is obtained by regressing the equation given below after taking log–

\[
Sales_t = Sales_0 (1+g)^t
\]

**Some Limitations of the Study**

In terms of methodology, the contribution of this study has been the classification of barriers to entry. However due to paucity of data entrants were not classified further on the basis of foreign or domestic ownership. Ideally one should have explored if barriers to entry vary for domestic and foreign entrants. This is particularly relevant during post 1991 period, when the Indian economy was opened to foreign participation and domestic firms have repeatedly complained about the existence of a non-level playing field.

Furthermore, the study considered only number of entrants for entry. Mergers, acquisitions, and takeovers have been important modes of entry in India especially during the post 1993 period when competition and FDI laws were liberalized. A
comparison of our results with those on the barrier to entry through mergers and acquisition activities can provide useful insights.

Another limitation arises from the non-availability of data on the R&D, foreign technology transfer, the degree of excessive productive capacity, industry demand elasticity etc. These factors can also be considered as barriers to entry. 'Entry', entails an investment decision. From an investment perspective, if one studies the barriers to entry, we can get more insights into the factors determining entry, specifically the role to be played by the capital markets, if one incorporates the size of investments involved.

The other set of limitations arises from the measurements of the explanatory variables. Many of them have been identified earlier. In addition, operating with excess capacity has been identified in the literature as a strategic move by firms. Theoretically excess capacity provides a credible threat to a potential entrant who fears that the incumbent may flush the market with large supplies as he enters the market resulting in excess supply and decline in prices. If the prices fall below average costs of operation, only a financially strong firm may sustain the losses resulting. In such a context, capacity utilization of the industry gives a good measure to capture such strategic moves. This has not been included in the study because of non-availability of data. The other major variable, which has not been captured because of non-availability of data, is that of import competition. We have been able to capture the export orientation of the industries and its impact on the barrier to entry as it provides larger opportunity for Indian players. At the same time one also needs to capture the effect of import competition on the incidence and extent of entry.

This study uses panel data on 98 industries for twelve years. In this framework, it is important to control for omitted industry specific variables that are constant across time. Assuming that the industry specific effects are fixed, these can be controlled by introducing industry dummies. Since the number of industries included in this study is
very large, this method would have resulted in a loss of significant degrees of freedom. So, these omissions should be accounted for when interpreting the results. Due to these computational difficulties, the industry specific effects could not be controlled for. That remains one of the limitations of the study.