Chapter 6:

Tools and Models

This chapter deals in all the models and tools that have been employed in the study. The models and tools have been explained in a detailed manner making it easier to understand their application later in the proceeding chapters.

Arithmetic Mean or Mean:

The mean is one of the most important and frequently used tool of mathematics and statistics. It is widely used in other disciplines like Economics, Psychology, Anthropology, etc. It is the sum of the observations under consideration divided by the total number of observations. It gives a number which is most representative of the observations.

$$\text{Arithmetic mean or mean} = (x_1 + x_2 + \ldots + x_n)/n$$

However, arithmetic mean is affected by outliers. Therefore, median is used sometimes to find out the 'middle' value of the distribution.

Standard Deviation:

Arithmetic mean, by itself, cannot tell us a lot about any distribution. We also need to know how spread the distribution is. In such situations, we use the standard deviation. Standard deviation is the square root of the mean of the square of the deviations of the observations from the mean or the median. It shows how far away the observations are from the mean or the median. If the standard deviation is more, it means the observations are too far away from the mean and so the mean is not a true representative of the distribution. The reverse is true when the standard deviation is small.

$$\sqrt[2]{\sum_{i=1}^{n}(\mathrm{Xn} - mean)/n}$$

X_n = n^{th} value of a random variable X

n = number of observations

Skewness:

The asymmetry in a statistical distribution is called skewness. When a distribution is skewed, one of the tails of the distribution is longer than the other and the mean, median and mode are not the same.

It can be quantified using the formula:

$$\text{Skewness} = [\sum (X_i - \text{mean})^3] / (N - 1)*\dot{O}^3$$

X_i = i^{th} randon variable

N = number of variables in the distribution

$\dot{O}$ = standard distribution

If the skewness value is 0, then the distribution is perfectly symmetric. If the value is positive, then the distribution is skewed towards the right and if the value is negative, the distribution is skewed towards the left.

Kurtosis:

Kurtosis is a statistical measure that shows how much the tails of a distribution differ from the tails of a normal distribution. In other words, it shows if there is existence of any extreme values in the tails of the distribution.

It can be quantified with the following formula:

$S2 = \sum (X_i - \text{mean})^2$

$S4 = \sum (X_i - \text{mean})^4$

$M_2 = S2/N$

$M_4 = S4/N$

X_i = i^{th} randon variable

N = number of variables in the distribution

$$\text{Kurtosis} = (M_4/M_2^2) - 3$$

If the kurtosis value is greater than 3 then the distribution is called leptokurtic and it means that the distribution has heavier tails than a normal distribution (presence of outliers) and if its value is lower than 3 it is called platykurtic and its tails are lighter than a normal distribution (lack of outliers).

Correlation analysis:

Correlation is a statistical technique which helps the researcher to understand the proximity between two variables in order to comprehend the relationship between them. In other words, correlation is directed towards measuring the degree of association between two variables and it refers to any broad class of statistical relationships involving dependence between two variables. The higher is the value of correlation between the variables, the more is the probability of one variable explaining the changes in the other variable provided the significance value is less than 0.5. Usually, in simple correlation between two variables the researchers are looking for a linear relationship.

$$r = \frac{\mathbf{n}.\sum \mathbf{dx}.\mathbf{dy} - \sum \mathbf{dx}.\mathbf{dy}}{\mathbf{n}.\sum \mathbf{dx2} - (\sum \mathbf{dx})\mathbf{2}\ \mathbf{n}.\sum \mathbf{dy2} - (\sum \mathbf{dy})\mathbf{2}}$$

r = Coefficient of correlation.

dx = Deviation of x series from an assumed mean.

dy = Deviation of y series from an assumed mean.

Σdx.dy = Sum of the product of the deviation of x and y series from their assumed mean.

Σdx2 = Sum of squares of x series from an assumed mean.

Σdy2 = Sum of squares of y series from an assumed mean.

n = Number of observations

Correlation is always between -1.0 and +1.0. If the correlation is positive, the variables are directly related to each other whereas if the correlation is negative, the variables are inversely related to each other.

Outliers:

Outliers are certain extreme observations in a distribution that unduly affects any statistical or econometric operations done upon it. Therefore, one of the primary assumptions in any statistical operation is to identify and remove the outliers.

The values outside certain minimum and maximum threshold values are outliers and so they should be eliminated before doing any statistical operation.

Minimum threshold value = Observation – (1.5 x Interquartile range)

Maximum threshold value = Observation + (1.5 x Interquartile range)

Interquartile range = 3^{rd} Quartile – 1^{st} Quartile

1^{st} Quartile = $(N+1)/4^{th}$ value of the observation.

25% of the observations are below the 1^{st} quartile.

3^{rd} Quartile = $3(N+1)4^{th}$ value of the observation.

25% of the observations are above the 3^{rd} quartile.

Data Envelopment Analysis [Hadad, Y. et al. (2007)]**:**

The Data Envelopment Analysis or, in short, the DEA is a non-parametric methodology for evaluating the relative efficiency of Decision-Making Units (DMU) based on multiple inputs and multiple outputs. The efficiency score is measured as a ratio between weighted outputs and weighted inputs, even if the production function is unknown. The DEA provides a dichotomy classification into two groups; efficient and inefficient. Later, other writers have developed this method into other more refined versions. For example, Anderson and Peterson (1993) came up

with the super efficiency method and Silkman et al. (1986) have put forwarded the Cross Efficiency Matrix method.

DEA permits each DMU to select any desirable weight for each input and output, provided that they satisfy certain reasonable conditions: first those weights cannot be negative, and second that the weights must be universal, which means that the resulting ratio should not exceed 1. The BCC model, named after Banker, Charnes and Cooper (1984) allows the production function to exhibit non-constant return to scale (Banker and Chang (1995)) while the CCR model imposes the additional assumption of constant returns to scale on the production function.

- <u>The CCR model is as follows:</u>

  Let there be n number of Decision-making Units (DMUs) where each DMU m (m = 1,2,…,n) uses j inputs X_m ($X_{1m,}$ $X_{2m,}$ …. , X_{jm}) > 0 and produces k outputs Y_m ($Y_{1m,}$ Y_{2m}, …. , Y_{km}) > 0. In the CCR model we try to the best weights U_r = (1, 2,…, R) and V_t = (1, 2,…, T) that maximise the ratio between weighted inputs and weighted outputs.

  Objective function: Max $\sum_{s=1}^{T} V_t^T Y_{km}$

  Subject to: $\sum_{r=1}^{S} U_r^R X_{jm} = 1$

  $\sum_{s=1}^{T} V_t^T Y_{km} - \sum_{s=1}^{T} V_t^T Y_{km} \leq 0$

  $U_r^R \geq 0$; r = 0, 1, 2,…., R

  $V_t^T \geq 0$; t = 0, 1, 2,…., T

  The weights are all positive and the ratios are bounded by 1 (100%). Any DMU which reaches the highest possible value of 100% will be efficient; otherwise it would be inefficient.

- The BCC model is as follows:

  Let there be n number of Decision-making Units (DMUs) where each DMU m (m = 1,2,…,n) uses j inputs X_m ($X_{1m}, X_{2m}, \ldots, X_{jm}$) > 0 and produces k outputs Y_m ($Y_{1m}, Y_{2m}, \ldots, Y_{km}$) > 0. In the CCR model we try to the best weights U_r = (1, 2,…, R) and V_t = (1, 2,…, T) that maximise the ratio between weighted inputs and weighted outputs.

  Objective function: Max $\sum_{s=1}^{T} V_t^T Y_{km} - W_k$

  Subject to: $\sum_{r=1}^{S} U_r^R X_{jm} = 1$

  $\sum_{s=1}^{T} V_t^T Y_{km} - \sum_{s=1}^{T} V_t^T Y_{km} - W_k \leq 0$

  $U_r^R \geq 0$; r = 0, 1, 2,…., R

  $V_t^T \geq 0$; t = 0, 1, 2,…., T

  W_k is free

Muller's method of calculating restaurant efficiency (Muller, 1999)**:**

Maximum seating capacity = (No. of seats X Hours of service) / Service cycle time

Restaurant capacity ratio = Actual cover count / Maximum seating capacity

Here, service cycle time is defined as the time required to serve a customer which starts from greeting the customer when he/she enters to bidding the customer farewell and ending in preparing the table for the next customer.

Actual cover count is the total number of customers a restaurant is able to serve in a day/week.

Learning Curve (Wright, 1936):

A learning curve is a diagrammatic representation of the relationship between the cost and output over a definite period of time. It basically shows how due to repetitive work, a worker get so much of skill or experience that the cost of production comes down. But it is also subject

to law of variable proportions. It means as we increase the time of operation, initially the cost of production comes down easily, but with subsequent increase in the time of operation, it get increasingly difficult to reduce the cost of production.

The learning curve was first described by psychologist Hermann Ebbinghaus in 1885 and it has henceforth been used to measure productive efficiency and to forecast cost.

Regression Analysis:

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The earliest form of regression was the method of least squares which was published by Legendre in 1805 and by Gauss in 1809. A regression is a statistical analysis assessing the association between two variables. Regression analysis is widely used for prediction and forecasting. Moreover, it is also used to understand which among the independent variables are related to the dependent variable and it explores the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

In this study, we have used various regression functions like linear regression, censored regression, binomial LOGIT regression and multinomial LOGIT regression.

Linear Regression Model:

In linear regression, the relationships are modelled using the linear predictor functions whose model parameters are estimated from the data. Such models are called linear regression models.

$$Y_i = \alpha + \beta X_i + e_i$$

Here, Y_i is the i^{th} dependent variable and X_i is the i^{th} independent variable. e_i is the random disturbance term or the error term. α and β are the regression coefficients or the model parameters that are needed to be estimated. This is a model with only one independent variable.

However, a model with multiple independent variables can also be constructed in the same manner.

Censored Linear Regression Model or the TOBIT Model:

This is analogous to the linear regression model discussed above only with the additional restriction that the value of the dependent variable is bounded or censored within limits in the upper side, lower side or both. For example, in constructing a model of the size of landholdings (dependent variable) and the income of a person (independent variable), the value of the size of landholdings cannot go below 0. Therefore, in such cases, we construct a censored linear regression model with restriction on the value of the lower side of the dependent variable.

Binary Logit Model:

Binary Logit (Logistic) Regression model is a statistical model that uses a logistic function to fit a binary dependent variable. More simply, in this model, there are two binary categories of the dependent variable and the independent variable(s) can be continuous, categorical or ordinal. Therefore, the binary variable is also called dichotomous variable as it can take only two values. Examples of binary dependent variable can be 'male/female', 'pass/fail', 'efficient/inefficient', etc. Therefore, using this model, the existence (or lack of it) of any kind of relationship between the binary dependent variable and one or more independent variables can be found out.

The binary logit regression can be understood as finding the parameters β in the model given below:

$$y = 1 \text{ if } \beta_0 + \beta_1 x + € > 0$$

$$= 0 \text{ else}$$

A binary logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables.

$$\text{logit}(p) = \text{logit}\,[p/(1-p)] = \beta_0 + \beta_1 x_1 + \ldots.. + \beta_k x_k$$

This model can be extended to multiple dependent variable categories too in which case it is called multinomial LOGIT model.

Multinomial LOGIT model:

A similar model can be constructed as the binary LOGIT model discussed above with the additional freedom that the dependent variable can have more than two categories.

For example, if we want to run a regression between gender (dependent variable) and level of education (independent variable), the dependent variable has three (or more) categories, namely, male, female and transgender. Therefore, such kind of models will be called multinomial LOGIT models.