5.1 Introduction

The present chapter describes the research purpose of the current study and the methodology adopted to attain these objectives. The main aim of the present study is to identify the various behavioral biases affecting the decision-making process of investors and to analyse investor’s attitude towards various equity market investment. There are a variety of methodologies, however not contending, in order to address investor’s behavioral issues. Noticeable among these are econometrics, Discriminant Analysis, Principal Component Analysis, Logit model, Probit model and Tobit model. Discriminant analysis and logit model is most suitable for the present research.

The methodological approach of the current research involves the extensive study of demographic traits and investment decisions patterns based on various behavioral and financial factors. After that, segregating the data in different groups on the basis of specific characteristics and examining which investor group is more affected or unaffected by behavioral biases.

The chapter is structured into two broad sections. Section 5.2 particularizes the data collection process required to implement discriminant analysis and logit model and includes the specifics such as time frame of the study, investor’s description, selection of behavioral biases and the data sources. Section 5.3 provides the methodological specification of the model. Sub-sections within section 5.3 provide the methodology used for analysing the impact of various behavioral biases on investor’s decision making process. Section 5.3 The chapter is recapitulated in the concluding section 5.4.
5.2 Data Implementation

5.2.1 Data Collection

The present study focuses on primary data collected with the help of questionnaires which is considered as the appropriate technique and an effective tool for gathering information from large number of respondents about their opinions, behavior and their perspective towards financial investment decision. Moreover, the main aim of the present study is to collect the data on investor’s behavior in order to analyse which investor group is more affected or unaffected by the behavioral biases as well as to examine the role of demographic variables in influencing an investor from a behavioral bias. The questionnaire was drafted trickily in order to collect the behavioral information of the investors without actually hurting their sentiments. The questionnaire is divided into three sections.

Section 1 consists of nineteen questions and focuses on the general information of the investors, e.g. gender, age, educational qualification, years of investment experience, investment period preference, reason for investing in stock market and how much loss their portfolio has suffered till now.

In section 2, respondents were provided with various investment avenues and were asked to rank them according to their choice of preference. Further, they were also asked to tick the appropriate risk and return level which they think that particular investment avenue possesses. The risk and return level questions were designed on five point “Likert Scale” ranges from very high (shown by 1) to very low (shown by 5); where very high indicates that investment avenue is highly risky or possesses high return and likewise, very low denotes the low level of risk/return.
In section 3, respondents were provided with questions related to behavioral aspect in order to understand whether their investment decisions are affected by any behavioral bias or not. All bias related questions are also designed on five Point “Likert Scale” in which responses ranges from strongly disagree to strongly agree. Respondents indicating strongly disagree means that their decisions are not affected by that bias and likewise, strongly agree denotes that the presence of that behavioral bias in their decision making process. Moreover, these five points are quantified by assigning values of five, four, three, two and one to strongly agree, agree, neutral, disagree and strongly disagree.

5.2.2 Sample Profile

The present study focuses on individual investors who invested in Indian stock market. The main reason for considering retail investors is that these people are more prone to behavioral biases and are likely to exhibit irrational behavior. The present research aims to collect data from diversified investors of distinct age, educational qualification, years of investment experience with differing attitudes in different market scenarios. The questionnaires were distributed in person to individual investors having investment experience in Indian stock market. For this purpose, the data of 5,000 investors was collected from reputed brokerage houses as well as from money control website where investors upload their portfolios for tracking purpose. Questionnaires which were completely filled in all respect were only taken for analysis purpose. A sample of 10% (500 respondents) investors was picked randomly to ensure that sample size truly represents the whole population. Further, the questionnaire was made accessible to respondents via online and offline mode from July 2015 to March 2016. The credible number of responses accumulated by the
questionnaire was 419. After removing incomplete questionnaires, the sample respondents were curtailed to 380 to be used for analysis.

5.2.3 Fragmentation of chosen sample

In order to study the impact of various behavioral biases on investor’s decision-making; the entire sample was segregated in two investor’s group each on the basis of specific demographic characteristic, namely investment experience, marital status, age and percentage of savings invested in stock market. Further, individual investor’s group is examined using discriminant analysis and chi-square test to check which investor group is more influenced or not influenced by a behavioral bias. Further, for logit model the entire sample was changed in binary variables (0, 1) and then analysed to check the role that demographic variables plays in influencing an investor from a behavioral bias. The following investor’s group were formed on the basis of chosen demographic characteristics:

<table>
<thead>
<tr>
<th>Basis</th>
<th>Investor’s Categorization in groups</th>
<th>No.of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Investment Experience</strong></td>
<td><strong>Group 1:</strong> Less experienced investors (investors having 5 or less than 5 years of investment experience in share market).</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td><strong>Group 2:</strong> Experienced investors (investors having more than 5 years of investment experience in share market).</td>
<td>205</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td><strong>Group 1:</strong> Young investors (investors below 40 years of age)</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td><strong>Group 2:</strong> Matured investors (investors above 40 years of age)</td>
<td>268</td>
</tr>
<tr>
<td><strong>Marital Status</strong></td>
<td><strong>Group 1:</strong> Unmarried Investors</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td><strong>Group 2:</strong> Married investors</td>
<td>215</td>
</tr>
<tr>
<td><strong>Percentage of savings invested in stock market</strong></td>
<td><strong>Group 1:</strong> Investors investing 20% or less than 20% of their savings in stock market.</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td><strong>Group 2:</strong> Investors investing more than 20% of their savings in stock market.</td>
<td>225</td>
</tr>
</tbody>
</table>
5.2.4 Research Hypotheses

After segregating the sample in various groups on the basis of demographic characteristics, next we have to check various assumptions forming the basis of present research. Following are the hypotheses formed for the present study:

(a) \( H_0 \): Behavioral biases do not exist among various investor groups.

(b) \( H_0 \): Behavioral biases vary amid the investor’s group formed on the basis of demographic characteristics.

(c) \( H_0 \): Demographic variables do not play any role in influencing an investor from a behavioral bias.

These research hypotheses were analysed and checked using different methodologies namely discriminant analysis, chi-squared test and logit analysis. The detailed specification of all the chosen methodology is given in section 5.3.

5.3 Methodological Specification

Statistical Package for Social Sciences (SPSS) version 21.0 and STATA 13.0 has been used for analysing the data. The process of data analysis starts with entering of data collected with the help of questionnaires, segregating it in its respective bias categories and then evaluating it to extract some relationship among them. To graphically present the demographic characteristics of the respondents like their age, gender, educational qualification, risk taking capacity, etc. various charts and tables were made using Microsoft Excel. Different statistical techniques described in below sub-sections were also used to bring out a convincing presentation and analysis of the data.
5.3.1 Discriminant Function Analysis

Discriminant function analysis is a statistical tool which is used to anticipate the dependent variable by one or more explanatory/independent variables. In other words, Discriminant analysis is an appropriate tool when the total sample is to be segregated into two or more mutually exclusive groups on the basis of some clearly defined independent variables and establishing a linear combination among them. It starts with the observations wherein both the group membership and interval variables values are known. A Discriminant function (also known as canonical root), is a latent variable which is formed as a linear combination of predictor variables.

In other words, it helps in determining which specific categorization or group a variable associates on the basis of its attributes or features and which variables are the perfect predictors that contribute majorly in discriminating the groups. Discriminant function is similar to multiple regression analysis but just opposite of Multivariate analysis of variance (MANOVA) wherein the independent variables are categorized as groups and the dependent variables are treated as its predictors. But in DA, dependent variables are categorized as groups and independent variables are treated as predictors.

Discriminant function analysis is done in 2-steps: the primary step includes preparation of a matrix of pooled within-group variances and co-variances to test the significance of discriminant functions. These two matrices are compared with the help of F-test to check the presence of any statistically significant difference between the groups. If any difference exists, then it moves forward to check which bias variable mean is statistically significantly different among the groups. After this step the process of classification of variables starts. The main purpose of discriminant analysis
is to remove the variables which are contributing very little in differentiating the groups.

**Assumptions of Discriminant Function Analysis:**

(a) Uneven sample sizes are admissible. It should have at most ‘n-2’ independent variables, where n represents the sample size. Generally, it is preferred that a group’s sample size should exceed the number of independent variables and it is best if it is 4 to 5 times of independent variables.

(b) There must be at least two groups.

(c) No independent variables should be linearly related to other independent variables.

(No multicollinearity among independent variables).

(d) The within pooled variance and covariance matrices for each categorized group must have statistically significant different means.

(e) Predictors (independent) variables should be normally distributed.

(f) If the sample size in a group is equal, then the mean score will represents its cut-off. But in case it is unequal, the cut-off is determined with the help of weighted means.

**The Linear equation of Discriminant Analysis can be written as:**

$$D = d_1X_1 + d_2X_2 + d_3X_3 + d_4X_4 + d_5X_5+d_6X_6 + d_7X_7 +d_8X_8+d_9X_9 + a$$

Where, $D =$ Discriminate Function.

$d_i =$ weight of the variable;  \hspace{0.5cm} X_i =$ Respondent’s score for the variable.

$a =$ constant, similar to residual in regression.
In the previous equation, $X_i$ represents the following biases:

$X_1 = \text{Loss Aversion Bias},$

$X_2 = \text{Regret Aversion Bias},$

$X_3 = \text{Herding Bias},$

$X_4 = \text{Overconfidence Bias},$

$X_5 = \text{Anchoring Bias},$

$X_6 = \text{Cognitive Dissonance Bias},$

$X_7 = \text{Representativeness Bias}.$

In the present study, investor type groups are chosen as dependent variables (discriminator) and are tested for their validity as the effective discriminator i.e. whether these two groups exhibit various behavioral biases in different manner or not. But before starting with the other statistical tests it is necessary to check for the presence of multicollinearity.

### 5.3.1.1 Multicollinearity Check

Multicollinearity (also known as collinearity) is an aspect wherein two or more explanatory variables in a multiple regression model are moderately or highly linearly related. In other words, multicollinearity is a linear combination of two or more explanatory variables, in which one variable can be linearly anticipated from other variables with a significant degree of certainty. As far as moderate multicollinearity is concerned it may not be a trouble for analysis. But, high multicollinearity can be an obstacle as it can inflate the coefficient estimates variance as a result of which estimates becomes very susceptible even to a small alteration in the model. Thus, it becomes very difficult to interpret the model results.
In this section, the responses collected on 5-point Likert scale were united to form 7 behavioral biases and these biases were further checked for the presence of multicollinearity. This is done with the help of a correlation matrix that represents the correlation coefficient value between the behavioral biases along with their significance level. Thereafter, level of significance was further checked to see if any bias is having values greater than 0.05 in majority. If any bias falls under this category, in that case their correlation coefficient values is to be checked to see if they are having high correlation values or not. In case, if any biases were found with high correlation value, they are further tested for the presence of multicollinearity using Variance Inflation Factor (VIF).

5.3.1.2 Variance Inflation Factor

Variance Inflation Factor (VIF) is used to check the multicollinearity among behavioral biases showing high correlation coefficient values. VIF shows the variability of estimated regression coefficients as compared to when no correlation between explanatory variables prevails. In other words, VIF depicts the degree to which the standard errors associated with respective beta values are inflated because of multicollinearity. Some data analysis software instead of calculating variance inflation factor, calculates tolerance level (1- R²) which is just the reverse of VIF. Moreover, square root of VIF shows the value by which standard error is greater, as compared to when variables are not correlated with other explanatory variables. VIF is calculated in following manner:

(i) Run an OLS regression in which X_1 is a function of various explanatory variables:

\[ X_1 = b_0 + a_2X_2 + a_3X_3+ a_4X_4+ \ldots \ldots +a_nX_n + e \]
Where $b_0$ is constant and $e$ is error term.

(ii) After step (i), calculate VIF by using following formula:

\[
VIF = \frac{1}{1 - R^2_i}
\]

Where, $R^2_i$ represents the coefficient of determination.

(iii) Check the value of Variance Inflation Factor: If VIF level is equal to 1, predictors (explanatory variables) are said to be not correlated; if its value ranges from 1 to 5, predictors are said to be moderately correlated and if its value exceeds 5, predictors are said to be highly correlated (which signifies the presence of multicollinearity). In case, if value of VIF exceeds 5, that variable should be removed from the study. Thus, it is always desired to have lower values of VIF, as high values tend to affect the results and make interpretation difficult in multiple regression analysis.

5.3.1.3 Group Statistics Table

Group statistics provides information about the data and also tells the number of cases falls in different dependent variable category. This table shows the means and standard deviations for each of the bias variables in each investor-type groups. In other words, it compares the means of various bias variables among the two groups to check if any significant difference exists between the groups on each of the predictor variables. If differences among the means do not prevail, then the chosen categorization of investor-type in different groups could not be treated as a significant discriminator and it would not be possible to carry on with further analysis. Moreover, it assigned a default weight of 1 to all the cases in the data, that’s why total weighted
number of cases falling in each group is equal to the total unweighted number of cases falling in each group.

### 5.3.1.4 Equality of Group Means Test

Equality of Group Means test examines each independent variable’s capability in differentiating the group before creation of the model. Equality of Group means provides the statistical confirmation for the difference in means estimated in Group Statistics table. In this, Wilk’s Lambda test statistic is used to check the null hypothesis of whether the group categorized on the basis of discriminator (investor-type) have identical means or not.

Further, Wilk’s Lambda test is used to identify which predictor (independent) variable is contributing significantly to the discriminant function. Its value ranges from 0 to 1, where lower value of Wilk’s Lambda confirms that both the groups have different means and the variables are contributing significantly in discriminating the groups. And high value of Wilk’s Lambda indicates the presence of identical means validating the fact that both the groups are exposed to these biases in a related manner.

Moreover, Wilk’s Lambda value is further confirmed with F-test for all independent bias variables. F-test is used to examine which predictor variable is contributing significantly to the discrimination function. In F-test low value signifies that means are identical and high value signifies the difference in means of the groups.

### 5.3.1.5 Stepwise Statistics

In stepwise statistics, a discrimination model is created step-by-step. More precisely, all variables are analysed and evaluated at each step to figure out which variable is contributing majorly to the discrimination between groups. Thereafter, that variable is
incorporated in the model, and process starts again with the remaining variables. In this, only those variables are entered at each step which reduces overall Wilk’s Lambda. In stepwise process, F statistic value and its probability is used for entering and removing the variables. The F value of a variable shows its statistical significance in discriminating the groups. Variables having high F-value and significant p-values are entered into the model and variables having low F-values and insignificant p-values are removed from the model.

5.3.1.6 Summary of Canonical Discriminant Functions

Summary of canonical discriminant functions provides the information on the variance reported by each of the discriminate functions produced. As there are only two groups in the analysis, that’s why only one discriminant function will be generated. If number of groups exceeds two, then several discriminant functions will be generated. This summary calculates eigenvalue which represents the percentage of variance explained by the discriminant function. The larger the eigenvalue, the greater amount of variance explained by the linear combination of independent variables. Generally, the eigenvalues are arranged in declining order of importance. So, the primary one always reveals the majority of the variance explained by the function. Further, it shows function and number of discriminating function that is calculated by number of groups less 1.

This summary also provides percentage of variance that explains the importance of the discriminant function followed by cumulative percentage of the variance as every discriminant function is added to the table. In other words, if there is more than one discriminant function, then only the first few ones having cumulative percentages higher than 90% will be considered for the analysis. At last, it provides canonical
correlation that shows correlation coefficient of each discriminant function. In other words, canonical correlation is the value between discriminant scores and the levels of dependent variable. High value of correlation shows that function is discriminating well.

In addition to the above, summary of canonical discriminant function calculates wilks’ lambda to check if the discriminant function is statistically significant. Its value ranges from 0 to 1. If the value of wilks’ lambda is close to 0, it indicates the increased contribution of that variable in discriminating function. On the other hand, if the value of lambda is close to 1, it shows that means of the observed groups are identical. In other words, small Wilks’ lambda value shows that means of the observed groups are not identical. Wilk’s Lambda coefficient presents the percentage of variability which is not elucidated by the discriminant function. In this wilk’s lambda table, at first it shows test of function (s) wherein discriminant functions are incorporated with the null hypothesis that the canonical correlations of the functions are equal to zero. Chi-square test significance value is used to test the null hypothesis that canonical correlations of the functions are equal to zero. If the p-value is less than 0.05, then null hypothesis is to be rejected. If it is greater than 0.05, then null hypothesis cannot be rejected. F-value and significant p-values are entered into the model and variables having low F-values and insignificant p-values are removed from the model.

5.3.1.7 StandardizedCanonical Discriminant Function Coefficients

The Standardized Canonical Discriminant Function Coefficients are similar to the beta coefficients of multiple regressions. These coefficients are used to point out the relative importance of the independent (predictor) variables in anticipating the
dependent variable (investor-type). The value of standardized coefficient shows the relative importance of each independent variable in discrimination function, where the highest coefficient value shows the increasing importance of the variable in explaining the differences among investor-type groups. In other words, highest value of a variable shows that the groups vary a lot on that particular bias independent variable. However, the sign of the coefficients signifies the direction of relationship between dependent and independent variable but it can be ignored for analysis purpose.

5.3.1.8 The Structure Matrix Table

Structure matrix is another technique used extensively for examining the relative importance of independent variables as structure matrix coefficients are considered more accurate and reliable than Standardized Canonical Discriminant Function Coefficients. Structure matrix table shows the correlation between each independent (predictor) variable and discriminant function. These coefficients are also known as structure coefficients or discriminant loadings. They serve the same purpose as factor loadings in factor analysis. Generally, a value of 0.30 is considered as cut-off point to decide between less important and important variables. And any variable having a value of 0.30 or more is taken up as important in predicting the differences between the groups.

5.3.1.9 Canonical Discriminant Function Coefficients

Canonical Discriminant Function Coefficients are unstandardized scores of independent variables and are interpreted in the same manner as the beta coefficients of multiple regressions. These coefficients are used to create discriminate function equation. These coefficients indicate the extent to which the independent (predictor)
variable helps in discriminant function and are also used to categorize the new cases in investor-type groups. In other words, both coefficients i.e. canonical discriminant function coefficients and standardized beta shows the importance of each independent variable in discrimination function controlling other independent variables in the equation.

5.3.1.10 Functions at Groups Centroids

Functions at Groups Centroids are another way of understanding the results of discriminant analysis. It shows the group centroids value (group means of independent variables) for each investor-type group. ‘Functions at Group Centroids’ indicate the discriminant score for respondents belonging to the investor-type group. Generally, if the groups are of identical size, the respondents can be segregated using the halfway between the values of the functions at group centroids. But if the groups are not of equal size, as is the case in present research, the best way of segregating the respondents is by using weighted average of the values of functions at group centroids. Moreover, cases with scores close to a respective centroid value are anticipated to be fitting to that group.

5.3.1.11 Classification Table

Classification table (also known as confusion table) displays the rate of success in predicting the respondent’s category in various investor-type groups using discriminant function. It shows the psychological pattern of both the investor groups. This table has two sections: original and cross-validated section. Original classification section shows how many investors of one group are responding like investors of other group and vice-versa while making their investment decisions. The second part of the table, i.e. cross-validated section, provides more accurate results than original classification section. In this method, one independent variable is kept
out and the discrimination function is constructed using other variables and then the
same process is replicated for other variables. Generally, a hit ratio of 75% and above
is considered satisfactory which shows the certainty with which the discriminant
function is capable of anticipating the behavioral patterns of investor groups.

5.3.2 Bias specific analysis

5.3.2.1 Contingency Table
After categorizing respondents in different investor-type groups, their responses
related to a specific bias are summed, averaged and checked to see how many
investors are falling within a particular response category. These results are arranged
in a rectangular table known as contingency table. The values in contingency table
show either frequencies or displayed in the form of percentages. Each column of
contingency table represents the responses of investors under different categories.

5.3.2.2 Chi-Square test for Independence
Chi-square test for independence is used to examine which investor-type group is
more influenced or not influenced by behavioral bias. Chi-square test is a
nonparametric test and in the present study it is used to test specific bias related
hypothesis. This test can be used only on discrete data (also on continuous data by
categorizing it in distinct discrete data). Chi-square test is used to check if two or
more categorization of the respondents is independent or not. Moreover, Chi-square
test is valid only when qualitative variables are segregated in various categories. Chi-
square value is computed so as to compare the observed and expected frequencies
with the help of following formula:

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]
Where, O = observed frequencies, and E = Expected frequencies.

After calculating chi-square value, it is checked against the chi-square table containing critical values along with distinct degrees of freedom and probability levels to check the following hypothesis:

**H₀**: There is no relationship between the variables (i.e. two variables are independent)

**H₁**: There is some relationship between the two variables.

If the chi-square value is less than the critical value, then null hypothesis (H₀) cannot be rejected. In the contrary case, null hypothesis (H₀) is to be accepted. In the present research, three chi-squared tests are executed, i.e. (a) Pearson chi-square (b) Likelihood Ratio test, and (c) Linear-by-linear association to test if any relationship exists between the variables or not.

### 5.3.2.3 Weighted Scoring Method

The weighted scoring method is a kind of multi-criterion analysis. In weighted scoring method, specific weights are ascribed to the related attributes. After this total score is computed on the basis of weights assigned to each of the attribute to check which alternative is performing well in comparison to each attribute. It includes the following steps:

(i) Classification of the attributes

(ii) Ascribe weights to the attributes to indicate their significance.

(iii) Estimate individual score to check how individual alternative is performing in contrast to each attribute.

(iv) Compute weighted scores for each alternative.
(v) Examine and interpret the results.

5.3.3 Logistic Regression

In a model where Y is quantitative, the main objective is to estimate its expected or mean value given the values of the regressors; whereas, in models where Y is qualitative, the objective is to find the probability of something happening. Hence, qualitative response regression models are often known as probability models. In these models, the regressand is generally a binary or dichotomous variable. There are three approaches of developing a probability model for a binary response variable: The Linear probability Model (LPM), The Logit Model and The Probit Model. But for the present research logit model is used for analysis.

For analysis, average of all the behavioral biases (or specific bias) was calculated for each respondent and then it was compared with the overall average of all the respondents. Respondents whose respective mean value is more than the overall mean value, they are said to be influenced by the respective bias. On the other hand, respondents having respective mean value less than the overall mean value are said to be not influenced by the bias. Further this data is used in Logit model, also called as Logistic regression, to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as success/failure. The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success P, or the value 0 with probability of failure 1-P. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic
regression; rather, the logistic regression function is used, which is the logit
transformation of p:

$$P = \frac{e^{z}}{1 + e^{z}} \quad \cdots \quad (1)$$

Where $z = \beta_1 + \beta_2X_i$

$\beta_1 = \text{constant of the equation}$

$\beta_2 = \text{the coefficient of the predictor variables.}$

Equation 1 above represents logistic distribution function. In this, value of Z ranges
from $-\infty$ to $+\infty$, P ranges between 0 and 1 and that P is nonlinearly related to Z (i.e.
$X_i$). It can be seen here that we cannot use OLS procedure as P is nonlinear not only
in $X_i$ and also in the $\beta$'s. Moreover, equation 1 shows the probability of the success of
the event P, then probability of failure (1-P) can be written as:

$$(1 - P) = \frac{1}{1 + e^{z}} \quad \cdots \quad (2)$$

Therefore, we can write:

$$\frac{p}{1-p} = \frac{1 + e^{z}}{1 + e^{-z}} = e^{z} \quad \cdots \quad (3)$$

Now, $\frac{p}{1-p}$ is the odds ratio in favour of success - the ratio of the probability of success
to the probability of failure. Further, if we take the natural log of equation 3, we
obtain the following equation:

$$L = \ln \left[ \frac{p}{1-p} \right] = z = \beta_1 + \beta_2X_i \quad \cdots \quad (4)$$

where L, the log of the odds ratio, is not only linear in X, but also linear in
parameters. Here, L is called as Logit.
Main features of the Logit Model are:

(a) As $P$ goes from 0 to 1 (i.e. as $Z$ varies from $-\infty$ to $+\infty$), the Logit $L$ goes from $-\infty$ to $+\infty$. In other words, even though the probabilities lie between 0 and 1, the logits are not so bounded.

(b) Even though $L$ is linear in $X_i$, the probabilities themselves are not.

(c) Although in the above equation only a single variable $X$ is included, but one can add as many regressors as may be stated in the underlying theory.

(d) If $L$ is positive, it means that when the value of regressor(s) increases, the odds that the regressand equals 1 increase. If $L$ is negative, the odds that the regressand equals 1 decrease as the value of $X$ increases.

(e) The Logit model equation given in equation 4 can be interpreted as: $\beta_2$, the slope, measures the change in $L$ for a unit change in $X_i$.

For estimation purposes, we can write equation 4 as follows:

$$L = \ln \left[ \frac{P}{1-P} \right] = \beta_1 + \beta_2 X_i + u_i \quad \text{......................... (5)}$$

To estimate equation 5, we need, apart from $X_i$, the values of the regressand or logit, $L$. And the values of Logit $L$ cannot be determined by standard OLS. In order to estimate the parameters, maximum likelihood (ML) will be used.

The Logit model here can be written as:

$$L = \frac{p}{1-p}$$

$$= \beta_1 + \beta_2 \text{Gender} + \beta_3 \text{Age} + \beta_4 \text{Qualification} + \beta_5 \text{marital status} + \beta_6 \text{Occupation} + \beta_7 \text{members in the family} + \beta_8 \text{type of investor} + \beta_9 \text{short term investment} + \beta_{10} \text{intraday}$$
trading + \beta_{11} \ Primary \ market \ preference + \beta_{12} \ investment \ experience + \beta_{13} \ number 
of \ companies \ for \ investment + \beta_{14} \ % \ of \ savings \ invested \ in \ stock \ market + \beta_{15} \ loss \ incurred \ on \ portfolio + \beta_{16} \ downfall \ experienced \ in \ share \ market + u_i

Now, here we cannot directly put the value P = 1 if an investor is not influenced by overall bias and zero if he is influenced by overall bias as we cannot use standard OLS here. Binary logistic regression is estimated using Maximum Likelihood Estimation (MLE), unlike linear regression which uses the Ordinary Least Squares (OLS) approach. MLE is an iterative procedure, meaning that it starts with a guess as to the best weight for each predictor variable and then adjusts these coefficients repeatedly until there is no additional improvement in the ability to predict the value of the outcome variable for each case. Before interpreting the results it is necessary to understand few things:

(a) The estimated standard errors are asymptotic in the method of maximum likelihood which is generally a large-sample method.

(b) Instead of using t-statistic to evaluate the statistical significance of a coefficient, z statistic is used.

(c) The conventional measure of goodness of fit, R^2, is not useful in binary regressand models. Measure similar to R^2, called Pseudo R^2 is used in the current analysis. But in binary regressand models, goodness of fit is of secondary importance. What is more important are the expected signs of the regression coefficients and their statistical significance.

(d) To test the null hypothesis that all the slope coefficients are simultaneously equal to zero, the likelihood ratio (LR) statistic this is equivalent to the F test in the linear
regression. Given the null hypothesis, the LR statistic follows the Chi-square distribution with df equal to the number of explanatory variables.

After knowing which demographic variable is playing an important role in influencing investors, it is checked if the selected model is fit for the data. And to test this hypothesis Hosmer and Lemeshow test is used. It is a goodness-of-fit statistics which helps in determining whether the model adequately describes the data. This statistics is the most reliable test of model fit for binary logistic regression because it aggregates the observations into groups of similar cases. The statistic is then computed based upon these groups. The Hosmer and Lemeshow statistic indicates a poor fit if the significance value is less than 0.05.

5.4 Summary

To sum up, the methodology makes use of discriminant analysis and logit model. Initially, discriminant analysis is used to verify which behavioral bias is contributing majorly in discriminating investor’s group. At the second stage, chi-square test for independence is used to check which investor group is more influenced or not influenced by a behavioral bias. Thirdly, logit model is used to examine the impact of demographic variables on decision-making process of investors.