Chapter 3
Research Methodology

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

The present chapter of this study deals with the research methodology which is adopted to attain the objectives of the current study. The main aim of the current study is to identify what all behavioral biases affect the decision-making process of individuals for the COVID-19 vaccination uptake or refusal. The study also deals with the attitude and perception of individuals in the case of vaccination and how this attitude leads to behavioral change among individuals. The methodology used to deal with the present research is the Logit model.

It becomes quite essential to investigate the behavior and emotional pattern of the respondents to garner meaningful responses and answers to the research question. Hence, firstly the methodology is framed to identify various behavioral biases, attitudes and demographic traits towards the decision- making process of individuals in the case of vaccination. After that, which specific factor is most significant to affect the decision of individuals to vaccinate is identified.

This chapter is divided into various sections and subsections underneath. The first section deals with the data collection process to run the logistic regression, data sources, tools used for collection of data, determination of sample and time frame of the study.The next section in this chapter provides the methodological specification of the model, methodology used to analyze the impact of demographic traits, behavioral factors, and geographical factors on the vaccination decision of individuals.

3.2 Data Implementation

3.2.1 Data Collection

The present empirical research study is based on primary data. The self-structured questionnaire is used for gathering information from respondents about their opinions, behavior and their perspective towards vaccination decisions. The main motive to undertake this research study is to identify all the various factors namely demographic traits, attitudes, perception towards COVID-19 vaccines and all the behavioral biases associated with the acceptance or refusal of vaccination. The questionnaire was trickily drafted to collect all the required information in the study without hurting the sentiments of individuals. The decision is also made by supplementing this study with personal interviews through video calls to gather a sufficient amount of information from the respondents related to the qualitative aspect. The self-structured questionnaire is divided into four different sections.

Section 1 consists of questions regarding general information of individuals which include name, age, gender, educational qualification, area of dwelling, occupation, whether they suffered from COVID-19 in past and whether they are vaccinated with the COVID-19 vaccine.

In **Section 2,** respondents were asked regarding their perception about COVID-19 vaccines, trust in the vaccines, whether they are worried about the side effects, all these questions were framed on a five-point "Likert scale" ranging from 1 to 5. Further, they were also asked some questions taking into account the geographical barriers associated with the uptake of vaccines and their confidence in the government and healthcare sector. These questions were also designed on five-point "Likert scale" ranges.

Section 3 consists of all the questions related to behavioral factors that are risk aversion, loss aversion, present bias, and impatient behavior to understand whether their vaccination decision is affected by any of the behavioral biases. All the behavioral related questions were designed

in the form of choices given to take part in a lottery and judgment is made accordingly.

Section 4 consists of the choices if respondents are interested in taking the vaccines in different cases. The questions related to these aspects were also designed on a five-point "Likert scale" ranging from highly interested (shown by 5) to not interested (shown by 1);where highly interested indicates that the respondent will take the vaccine in the given case and not interested means respondent is not interested to take the vaccine in the given case.

3.2.2 Sample Profile

The main aim of the present research study is to collect data from diversified individuals related to vaccines from distinct age groups, educational qualifications, occupations and areas of the dwelling. Since the population size is unknown, a sample of 150 respondents was picked randomly from the Delhi NCR region. Questionnaires were distributed to the respondents through online mode in June 2021. The credible number of responses from the questionnaires distributed was 125. Only those questionnaires were taken into consideration for further analysis which was duly filled in all the respects. After removing the incomplete questionnaires, the sample was reduced to 121 for further analysis. The sample is segregated based on specific demographic characteristics namely gender, age, area of dwelling, occupation and educational qualification.

3.2.3 Propositions of the Study

To meet the set objectives of the present study, the following propositions are built to verify empirically:

i) Demographic characteristics have a significant impact on thevaccination decision of an individual.

ii) Attitude, beliefs and opinions play a significant role in the vaccination decision of an individual.

iii) Confidence in the government and healthcare sector has a significant impact on the vaccination decisionof an individual.

iv) Geographical barriers have a significant impact on the vaccination decision of an individual.

v) Behavioral biases have a significant impact on the vaccination decision of an individual.

To test these propositions and to analyze them accordingly, various methodologies are used namely, chi-squared test and logit analysis. The level of significance is considered as 10%. The detailed specification of the given methodology is given in section 3.3.

3.3 Methodological Specification

For the detailed data analysis, Statistical Package for Social Sciences (SPSS) version 22.0 has been used. Data analysis is done by entering the collected data from theresponses received with the help of questionnaires, segregating the data accordingly and then evaluating it to extract some relationship among them.To graphically present the demographical characteristics of the respondents namely age, gender, educational qualification, occupation and area of dwelling, charts and tables are made using SPSS version 22.0. The various statistical techniques, tools used to bring out convincing presentation and analysis of the data has been discussed in the following sub-sections.

3.3.1 Logistic Regression

In various models, the dependent variable Y is quantitative; in such models, the main objective of running the regression model is to estimate its expected value for the various given values of the regressors. There are some cases when the dependent variable Y is qualitative; in such models, finding the probability of something happening is the main motive.

Therefore, these models are sometimes called qualitative response regression models or probability models. The dependent variable or regressand is generally binary or dichotomous

in such models. Three approaches are generally used in the case of binary response variables to develop a probability model: They are linear probability model (LPM), logit model, and probit Model. Among these, the logit model is used for the present research study.

The logit model is also known as logistic regression. The data used for the research purposein the present study is used to form a logit model to predict the discrete outcome from the various set of variables which can be discrete, continuous, and dichotomous or can be a combination of any of these. In the logit model, the dependent variable is generally dichotomous and can take two forms such as yes/no. It means that the dependent variable can take the value 1 or 0 where 1 means the probability of success p and 0 meansthe probability of failure 1- p. In the case of logistic regression, the relationship between predictor/dependent variable and the response variables is not a linear function rather logistic regression function is used. They also do not need to be normally distributed, linearly related or of equal variance within each group.

$$p = e^z / 1+ e^z \cdots\cdots (1)$$

Where, $z = \beta_1 + \beta_2 X_i$

β_1 = Constant term of the equation

β_2 = Coefficient of the predictor variables in the equation

The logistic distribution function is represented by Equation (1) where the value of z in the equation, ranges from $-\infty$ to $+\infty$ and the value of p range between 0 and 1. Moreover, pis non-linearly related to z. It can also be inferred that estimation cannot be done through the use of OLS because p is non-linear in X_i as well as in the β's.

Since Equation (1) above shows the probability of success p, therefore it means that the probability of failure i.e. 1 – p can be written as:

$$(1 – p) = 1/ 1+ e^z \cdots\cdots (2)$$

So, from Equation (1) and Equation (2),

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

Equation (3) indicates the odds ratio in the favor of probability of success p i.e. ratio of the probability of success to the probability of failure. Now, by taking a log in Equation (3), the following resultant equation is obtained:

$$L = \ln \left[\frac{p}{(1-p)} \right] = z = \beta_1 + \beta_2 X_i \ldots\ldots (4)$$

In the above Equation (4), L represents the log of odds ratio which is linear in parameters as well as linear in X. Therefore, it is called the logit model. The features of the logit model include; (i) Probabilities lie between 0 and 1 and the value of logit L goes from $-\infty$ to $+\infty$. It means that the logits are not so bounded, (ii) L is linear in X_i but the probabilities are not, (iii) There can be many independent variables or regressors in the model, as the casemay be according to the theory, and (iv) The slope i.e. β_2 in the above Equation (4) measures the changes in L for a unitchange in X_i.

For the estimation purpose, Equation (4) can further be rewritten as:

$$L = \ln \left[\frac{p}{(1-p)} \right] = \beta_1 + \beta_2 X_i + u_i \ldots\ldots (5)$$

Since it is notpossible to estimate the equation with the OLS method due to some of the drawbacks, hence, the Maximum Likelihood Method (MLE) will be used to estimate the parameters in the given equation.

The logit model for the present study can be written as,

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

OR,

$$L = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + u_i$$

Where, L = 1 or 0 1 = Acceptance of vaccination

0 = Refusal of vaccination

X_i = Independent variables used in the study

β_i = Logit coefficients

u_i = Error term

Here, it can be seen that it is not possible to directly compute the value of parameters through the standard OLS and hence, the value of p =1 foraccepting the vaccine and p = 0 for refusal of vaccine cannot be put directly. Therefore, it is always preferred to use Maximum Likelihood Estimation (MLE) in the case of a binary logistic regression model.

Before moving to the next section, it is quite important to discuss a few things regarding the methodology and techniques of the present study:

i) R^2 is one of the widely used conventional measures to check the goodness of fit of the model, but it is not generally accepted to use R^2 in the case of a binary regression model. So instead of that, Pseudo R^2 will be used in the current analysis.

ii) To check the propositions framed in the study, the Likelihood Ratio test or statistic will be used which is a counterpart to the F statistic in the case of a linear regression model.

iii) In the Maximum Likelihood method, standard errors obtained are asymptotic which is generally in the case of a large sample.

After checking for the regression analysis, Hosmer and Lemeshow test is used to check the

goodness of fit of the model. This statistic helps to determine how accurately the model is described by the data. It is one of the most widely used statistics in the caseof a binary regression model.

3.3.2 Multicollinearity Check

Multicollinearity in the model arises when two or more explanatory variables in the case of multiple regression models are correlated with each other. There arises a problem in the case of multicollinearity because independent variables should be independent and if they are correlated then the problem occurs with the fitness of the model. If multicollinearity present in the model is moderate, then it may not be a trouble. But, high multicollinearity can cause trouble and serious obstacles in the model.

In the case of the present research study, the multicollinearity is checked for the explanatory variables used in the study to understand whether the variables such as attitude towards a vaccine, trust in the vaccine, behavioral biases is correlated witheach other. This is done by using the correlation matrix. The correlation coefficient matrix helps to understand the correlation coefficient values among the explanatory variables used along with the level of significance.

3.3.2.1 Variance Inflation Factor

There are various measures to detect multicollinearity in the model. Since the regression model used in the study is binary logistic regression analysis, so the best way to detect thepresence of multicollinearity is the Variance Inflation Factor (VIF). The variance inflation factor depicts how much the variance of the explanatory variable or the behavior of the explanatory variable is inflated by the correlation of other independent or explanatory variables. It helps to understand quickly how much the variable is affecting the standard error. In the case of SPSS, VIF value is depicted to detect the presence of multicollinearity but some software for data

analysis provides the value of Tolerance (TOL), which is reciprocal or reverse of VIF. Following is the procedure to calculate VIF:

i) First, it is required to run the OLS regression where one of the explanatory variable act as a dependent variable and is a function of other explanatory variables:

$$X = b_0 + a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n + e$$

Here, b_0 is a constant term and e signifies the error term.

ii) After running the OLS regression, the value of VIF can be obtained by using the formula for VIF:

$$VIF = 1/ 1- R^2_i$$

The value of VIF should be checked to predict the presence or absence of multicollinearity. If the value of VIF obtained is equal to 1 then the explanatory variables are not correlated, if the value ranges between 1 and 5, then the moderate multicollinearity is detected. There is a presence of strong multicollinearity if the value exceeds 5. If strong multicollinearity is detected in the model, then various remedial measures should be used to solve the problem of multicollinearity.