1. Introduction
In dummy regression variable models, it is assumed implicitly that the dependent variable Y is quantitative whereas the explanatory variables are either quantitative or qualitative. There are certain type of regression models in which the dependent or response variable is dichotomous in nature, taking a 1 or 0 value. Suppose one wants to study the labor-force participation of adult males as a function of the unemployment rate, average wage rate, family income, education etc. A person is either in the labor force or not. Hence, the dependent variable, labor-force participation, can take only two values: 1 if the person is in the labor force and 0 if he or she is not.

There are several examples where the dependent variable is dichotomous. Suppose one wants to study the union membership status of the college professors as a function of several quantitative and qualitative variables. A college professor either belongs to a union or does not. Therefore, the dependent variable, union membership status, is a dummy variable taking on values 0 or 1, 0 meaning no union membership and 1 meaning union membership.

Similarly, other examples can be ownership of a house: a family owns a house or it does not, it has disability insurance or it does not, a certain drug is effective in curing an illness or it is not, decision of a firm to declare a dividend or not, President decides to veto a bill or accept it, etc.

A unique feature of all these examples is that the dependent variable is of the type which elicits a yes or no response. There are special estimation / inference problems associated with such models. The most commonly used approaches to estimating such models are the Linear Probability model, the Logit model and the Probit model. There are certain problems associated with the estimation of Linear Probability Models such as:

(i) Non-normality of the Disturbances (U’s)
(ii) Heteroscedastic variances of the disturbances
(iii) Nonfulfillment of $0 \leq \text{E}(Y \mid X) \leq 1$ (Possibility of $\hat{Y}$ lying outside the 0 - 1 range)
(iv) Questionable value of $R^2$ as a measure of goodness of fit; and

Linear Probability Model is not logically a very attractive model because it assumes that $P_i = \text{E}(Y = 1 \mid X)$ increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout. This seems sometimes very unrealistic. Therefore, there is a need of a probability model that has two features: (1) as X increases, $P_i = \text{E}(Y = 1 \mid X)$ increases but never steps outside the 0-1 interval, and (2) the relationship between $P_i$ and $X_i$ is non-linear, that is, approaches “one” which approaches zero at slower and slower rates as $X_i$ gets small and approaches one at slower and slower rates as X gets very large.
2. **Logit Model**

Logit regression (logit) analysis is a uni/multivariate technique which allows for estimating the probability that an event occurs or not, by predicting a binary dependent outcome from a set of independent variables.

In an example of home ownership where the dependent variable owns a house or not in relation to income, the linear probability Model depicted it as

\[ P_i = E(Y = 1 | X_i ) = \beta_1 + \beta_2 X_i \]

Where \( X \) is the income and \( Y = 1 \) means that the family owns a house.

Let us consider the following representation of home ownership:

\[
P_i = E(Y = 1 | X_i ) = \frac{1}{1 + \exp[-(\beta_1 + \beta_2 X_i)]} + \frac{1}{1 + \exp(-Z_i)} \quad \text{...(2.1)}
\]

where \( Z_i = \beta_1 + \beta_2 X_i \)

This equation (1) is known as the (cumulative) logistic distribution function. Here \( Z_i \) ranges from \(-\infty\) to \(+\infty\); \( P_i \) ranges between 0 and 1; \( P_i \) is non-linearly related to \( Z_i \) (i.e. \( X_i \)) thus satisfying the two conditions required for a probability model.

In satisfying these requirements, an estimation problem is created because \( P_i \) is nonlinear not only in \( X \) but also in the \( \beta \)'s. This means that one cannot use OLS procedure to estimate the parameters.

Here, \( P_i \) is the probability of owning a house and is given by

\[
\frac{1}{1 + \exp(-Z_i)}
\]

Then (1- \( P_i \)), the probability of not owning a house, is

\[
(1 - P_i ) = \frac{1}{1 + \exp(Z_i)}
\]

Therefore, one can write

\[
\frac{P_i}{1 - P_i} = \frac{1 + \exp(Z_i)}{1 + \exp(-Z_i)} \quad \text{...(2.2)}
\]

\( P_i / (1 - P_i) \) is the odds ratio in favour of owning a house i.e; the ratio of the probability that a family will own a house to the probability that it will not own a house.

Taking natural log of (2), we obtain

\[
L_i = \ln [P_i / (1 - P_i)] = Z_i = \beta_1 + \beta_2 X_i \quad \text{...(2.3)}
\]
That is, the log of the odds ratio is not only linear in \( X \), but also linear in the parameters. \( L \) is called the Logit.

### 2.1 Features of Logit Model

(i) As \( P \) goes from 0 to 1, the logit \( L \) goes from \( -\infty \) to \( +\infty \). That is, although the probabilities lie between 0 and 1, the logits are not so bounded.

(ii) Although \( L \) is linear in \( X \), the probabilities themselves are not.

(iii) The interpretation of the logit model is as follows: \( \beta_2 \), the slope, measures the change in \( L \) for a unit change in \( X \), i.e. it tells how the log odds in favour of owning a house change as income changes by a unit. The intercept \( \beta_1 \) is the value of the log odds in favour of owning a house if income is zero.

(iv) Given a certain level of income, say \( X^* \), if we actually want to estimate not the odds in favour of owning a house but the probability of owning a house itself, this can be done directly once the estimates of \( \beta_1 \) and \( \beta_2 \) are available.

(v) The linear probability model assumes that \( P_i \) is linearly related to \( X_i \), the logit model assumes that the log of odds ratio is linearly related to \( X_i \).

### 2.2 Estimation of Logit Model

In order to estimate the logit model, we need apart from \( X_i \), the values of logit \( L_i \). By having data at micro or individual level, one cannot estimate (3) by OLS technique. In such situations, one has to resort to Maximum Likelihood method of estimation. In case of grouped data, corresponding to each income level \( X_i \), there are \( N_i \) families among which \( n_i \) are possessing a house.

Therefore, one needs to compute

\[
\hat{P}_i = \frac{n_i}{N_i}
\]

This relative frequency is an estimate of true \( P_i \) corresponding to each \( X_i \). Using the estimated \( P_i \), one can obtain the estimated logit as

\[
\hat{L}_i = \ln \left( \frac{P_i}{1-P_i} \right) = Z_i = \beta_1 + \beta_2 X_i
\]

### 2.2.1 Steps in Estimating Logit Regression

(i) Compute the estimated probability of owning a house for each income level \( X_i \), as

\[
\hat{P}_i = \frac{n_i}{N_i}
\]

(ii) For each \( X_i \), obtain the logit as

\[
\hat{L}_i = \ln \left( \frac{\hat{P}_i}{1-\hat{P}_i} \right)
\]

(iii) Transform the logit regression in order to resolve the problem of heteroscedasticity as follows:

\[
\sqrt{W_i} L_i = \beta_1 \sqrt{W_i} + \beta_2 \sqrt{W_i} X_i + \sqrt{W_i} U_i \quad \text{... (2.4)}
\]

where the weights \( W_i = N_i \hat{P}_i / (1 - \hat{P}_i) \).
(v) Estimate (2.4) by OLS (WLS is OLS on transformed data)

(vi) Establish confidence intervals and / or test hypothesis in the usual OLS framework. All the conclusions will be valid strictly only when the sample is reasonably large.

2.3 Merits of Logit Model
(i) Logit analysis produces statistically sound results. By allowing for the transformation of a dichotomous dependent variable to a continuous variable ranging from $-\infty$ to $+\infty$, the problem of out of range estimates is avoided.
(ii) The logit analysis provides results which can be easily interpreted and the method is simple to analyse.
(iii) It gives parameter estimates which are asymptotically consistent, efficient and normal, so that the analogue of the regression t-test can be applied.

2.4 Demerits
(i) As in the case of Linear Probability Model, the disturbance term in logit model is heteroscedastic and therefore, we should go for Weighted Least Squares.
(ii) $N_i$ has to be fairly large for all $X_i$ and hence in small sample; the estimated results should be interpreted carefully.
(iii) As in any other regression, there may be problem of multicollinearity if the explanatory variables are related among themselves.
(iv) As in Linear Probability Models, the conventionally measured $R^2$ is of limited value to judge the goodness of fit.

2.5 Application of Logit Model Analysis
(i) It can be used to identify the factors that affect the adoption of a particular technology say, use of new varieties, fertilizers, pesticides etc, on a farm.
(ii) In the field of marketing, it can be used to test the brand preference and brand loyalty for any product.
(iii) Gender studies can use logit analysis to find out the factors which affect the decision making status of men/women in a family.

3. Probit Model
In order to explain the behaviour of a dichotomous dependent variable we have to use a suitably chosen Cumulative Distribution Function (CDF). The logit model uses the cumulative logistic function. But this is not the only CDF that one can use. In some applications, the normal CDF has been found useful. The estimating model that emerges from the normal CDF is known as the Probit Model or Normit Model.

Let us assume that in home ownership example, the decision of the $i^{th}$ family to own a house or not depends on unobservable utility index $I_i$, that is determined by the explanatory variables in such a way that the larger the value of index $I_i$, the greater the probability of the family owning a house. The index $I_i$ can be expressed as

$$I_i = \beta_1 + \beta_2 X_i, \quad \ldots(3.1)$$

where $X_i$ is the income of the $i^{th}$ family.
3.1 Assumption of Probit Model

For each family there is a critical or threshold level of the index ($I_i^*$), such that if $I_i$ exceeds $I_i^*$, the family will own a house otherwise not. But the threshold level $I_i^*$ is also not observable. If it is assumed that it is normally distributed with the same mean and variance, it is possible to estimate the parameters of (3.1) and thus get some information about the unobservable index itself.

In Probit Analysis, the unobservable utility index ($I_i$) is known as normal equivalent deviate (n.e.d) or simply Normit. Since n.e.d. or $I_i$ will be negative whenever $P_i < 0.5$, in practice the number 5 is added to the n.e.d. and the result so obtained is called the Probit i.e;

$$ \text{Probit} = \text{n.e.d} + 5 = I_i + 5 $$

In order to estimate $\beta_1$ and $\beta_2$, (3.1) can be written as

$$ I_i = \beta_1 + \beta_2 X_i + U_i $$ \hspace{1cm} \ldots (3.2)

3.2 Steps involved in Estimation of Probit Model

(i) Estimate $P_i$ from grouped data as in the case of Logit Model, i.e.,

$$ \hat{P}_i = \frac{n_i}{N_i} $$

(ii) Using $P_i$, obtain n.e.d ($I_i$) from the standard normal CDF, i.e. $I_i = \beta_1 + \beta_2 X_i$

(iii) Add 5 to the estimated $I_i$ to convert them into probits and use the probits thus obtained as the dependent variable in (3.2).

(iv) As in the case of Linear Probability Model and Logit Model, the disturbance term is heteroscedastic in Probit Model also. In order to get efficient estimates, one has to transform the model.

(v) After transformation, estimate (3.2) by OLS.

4. Logit versus Probit

(i) The chief difference between logit and probit is that logistic has slightly flatter tails i.e; the normal or probit curve approaches the axes more quickly than the logistic curve.

(ii) Qualitatively, Logit and Probit Models give similar results, the estimates of parameters of the two models are not directly comparable.