Chapter - 7

Problems related to Elementary Schooling in Rural Uttar Dinajpur District: An Empirical Exercise II

7. Introduction & Relevance of the study

One of the important findings of literacy analysis in the previous chapter shows that education itself is an important significant factor that can enhance the development of further educational scenario at least at household level. The age composition of the literate members of the study villages indicates an important feature. It has been observed that out of total literate members (407) in all the villages belonging to age group of 5 years and above, 54.8% (223) belongs to the schooling age group (5 to 14 years), 17% (69) in the age group of 15-25 years and the remaining 28.3% (115) in the age group of above 25 years. It is expected that majority of the parents of school going age children belong to the age group of above 25 years. It has been found that out of 120 households in the study villages, 43 fathers and 58 mothers are completely illiterate. The age wise literacy scenario is shown in Table-7.1.

Table-7.1: Age wise Literacy Character of the study Villages

	Kantigach	Juropani	Dangipara	Uttar Bhagalpur	All Villages
No of Family members in the age group 5 to 14years	59	72	58	61	250
No of Literate Family members in the age group 5 to 14years	48	62	56	57	223
No of Family members in the age group 15 to 25years	24	30	34	27	115
No of Literate Family members in the age group 15 to 25years	6	7	32	24	69
No of Family members aged above 25 years	59	56	87	59	261
No of Literate Family members aged above 25 years	6	11	63	35	115
Literacy Rate 5 to 14 years	81.36	86.11	96.55	93.44	89.20
Literacy Rate 15 to 25 years	25.00	23.33	94.12	88.89	60.00
Literacy Rate 25 years and above	10.17	19.64	72.41	59.32	44.06

Source: Calculated from Field Survey

It is also observed that in all the villages the literacy rate is found to be highest among the schooling age (5-14 years) children and it diminishes with the increase in age. It is not a unique feature of the study villages. The interesting observation is that in the low literate villages (Kantigach and Juropani) the gap in literacy rate between the schooling age children and the other family members is comparatively higher. It is obvious that the family members belonging to the age group 5-14 years having a minimum literacy level will hardly have any significant impact if there is no assurance that their school education cycle will be completed. This means that eight years of schooling is a prima-facie to have meaningful educational level. This requires, among other things, universal enrolment and the school dropout rate as low as possible. The primary target in such an educationally deprived area is to have minimum of eight years of schooling for all children so that within a near future this meaningful literate section will further generate a more educationally advanced human resource for the society. Thus it appears that it is not an overnight journey to bring an educationally underdeveloped area under the umbrella of educationally advanced area. This final section of the dissertation thus concentrates exclusively on the problem of elementary schooling in rural West Bengal.

Child schooling, unlike literacy rate, is rather complex in nature (Duraisamy, 2001). In the literature of economics of education, the researchers generally deal with several educational outcomes, such as dropout rate, enrolment rate, grade completion, school attendance rate, retention rate, etc.. These educational outcomes sometimes may take the form of qualitative rather than quantitative nature. In the literature of economics of education, situations involving qualitative outcomes are quite common. For example, enrolment decision that is usually taken by the parents cannot be quantified. The parents in this situation may have two options - either to enroll their wards or not to enroll them (Majumdar, 2001). Consider also the enrolment scenario of the children in a household, there may be a situation where some (> or = 1) of the schooling age children were never enrolled in any school as opposed to the event that all the children got admitted in a school. Again for example, suppose a household has three children in the age group 5 to 14 years and among them the 13 year old child has never been enrolled in any school. Again, the opposite scenario may be that in a household with two such children, all of them are enrolled in school. The summary of these responses is that in case of the first household, the event of never enrolment is present and for the second household, the event of never enrolment is absent. This helps in making an assessment of universal enrolment of children in school. Universal enrollment becomes more meaningful by studying such decision making behaviour of the households. The response variable may be assigned two values, 1, if some (> or = 1) of the schooling age children are never enrolled in any school and '0', if all the children got admitted in a school. It may however be noted here that the 0, 1 values of the response are chosen arbitrarily for mathematical convenience. Similarly, if it is to assess whether any children of a particular household, who was/were enrolled in the past, has dropped out or not, then the regressand or the dependent variable becomes qualitative in nature. Study of such an option covers a range of dimension where one can interpret whether all the schooling age children get enrolled in school and also the enrolled children are continuing their school education, i.e. the probability of retention is also covered here. Thus by incorporating these two variables in such a qualitative approach, universal enrollment and universal retention can be deciphered. In studying the elementary schooling, the above two qualitative response variables thus have been incorporated. A similar study was carried out in the same state where the probability of child school participation was predicted defining the enrolment decision (y) as y=1 if the child was ever enrolled in school/attending school and =0 if the child was never enrolled in school/had dropped out of school (Sengupta et.al. 2002). The study therefore, does not split the enrolment and dropout scenario.

When the dependent variable becomes categorical and binary (0-1) in nature, it assumes a dichotomous nature. Econometric investigation of such binary response is rather complex. Ordinary Least Squares can model binary response variables using linear probability models (LPM) (Menard, 1995; Cohen, Rea and Lerman, 1970 cited in Gujrati, 2003). Although the linear probabity model ($Yi = \alpha0 + \alpha iXi + ui$ or $Pi = E(Yi/Xi) = \alpha0 + \alpha iXi$) has the advantage of simplicity, it suffers from several problems such as non-normality of ui, heteroscedasticity of ui, possibility of estimated probability of the dependent variable to lie out side the 0-1 range and possibility of becoming the R2 value very low. However, one can resolve these problems following some statistical techniques such as- i) by applying WLS to resolve the heteroscedasticity, ii) by increasing the sample size to minimize the non-normality of ui, iii) restricted least-square may also be helpful in making the estimated probability to lie within the reasonable range (0-1). Even then the LPM suffers from a fundamental problem as because it assumes $Pi = E(Yi/Xi)$ increasing linearly with the increase in Xi.

But with a binary dependent variable, need arises for a model where as Xi increases, the estimated probability of dependent variable will increase but never step outside the 0-1 interval. Secondly, a non-linear relationship is expected between Pi and Xi. It has rightly been remarked that Pi approaches zero at slower and slower rates as Xi gets smaller and smaller and approaches one at slower and slower rates as Xi gets very large (Aldrich and Nelson, 1984). Thus instead of a straight line (assumed in LPM Model), it seems preferable to fit some kind of sigmoidal curve that resembles an inverted S laid on its side. The both tails of a sigmoid curve level off before reaching Pi = 0 or 1, so that the problem of impossible values of Pi can easily be avoided. Popular methods used to analyze such a sigmoid curve, include the probit model, linear discriminant function analysis, and logistic regression. The logit and probit actually yield similar results although not identical (Aldrich and Nelson, 1984; Hosmer and Lameshow, 1989, page 168). Probit regression, based on the probability integral transformation, lacks natural interpretation of regression parameters (Press and Wilson, 1978; Hailpern & Paul F. Visintainer, 2003). They also argue that although the discriminant analysis is computationally simpler than the probit model, the later assumes that predictor variables are normally distributed and that variables jointly assume a multivariate normal distribution. In a situation where some of the predictor variables are dichotomous or discrete, the assumption of normality is often violated. In contrast, the logistic regression model makes no assumption about the variable distribution. It is a direct probability model because it is stated in terms of $\Pr\{Y = 1|X\}$. Another advantage of the logit model is its ability to provide valid estimates, regardless of study design (Harrell, 2001). Considering all these, we have finally preferred to go with the logit model partly because it is comparatively easy to work with mathematically, and partly because it leads to a model that is easy to interpret. Logistic regression was proposed as an alternative in the late 1960s and early 1970s (Cabrera, 1994), and it became routinely available in statistical packages in the early 1980s.

Given the above relevance, the aim of the present chapter of the research is to empirically investigate the significant indicators of two schooling aspects, namely problem of never enrolment in school and dropout of school. As stated earlier, the logistic regression technique proceeds with a presumption that the result may improve upon the earlier empirical works in the related literature. Two central concepts have been conveyed that have a closer link with the process of such empirical investigation - to convey the concepts of logistic regression as simply as possible and to demonstrate how the logistic regression technique has been applied in the present data set. Apart from this econometric investigation, a qualitative search on the issue of school enrolment is also being discussed with a view to compare the econometric result with the qualitative observation of data set.

7.1 Logistic Regression Exercise

7.1.1 Modeling a Categorical Dependent Variable

Logistic regression analysis (LRA) extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical. Categorical socio-economic variables may be binary response where the response, Y, of a subject can take one of two possible values, denoted usually by 1 for positive response and 0 for negative response (Hosmer and Lameshow, 1989) and as well as it may take the form of ordinal response where the response, Y, of a subject can take one of m ordinal values, denoted by 1, 2,.....m (Agresti, 1990) data.

In the model, two binary response variables have been stated earlier. Binary response variables are Bernoulli distributed. For such dichotomous response variable, dependent variable can be assigned only two values - 0 and 1. So, modeling the actual values of Y is not exactly of interest and accordingly it does not make any sense. Rather, modeling the probability that each individual in the population responds with 0 or 1 (Karen Grace-Martin, 2001) is what is required. Thus, instead of simply regressing the actual value of dependent variable (which could take only two values) it would be appropriate to model the probability of occurrence of positive or negative response (Karen Grace-Martin, 2001). The application of logistic regression technique accrues from this concept of regressing the probability instead of predicting the actual value of the dependent variable.

Keeping in mind that probability is a measure of likelihood ranging from 0 (an impossibility) to 1 (a certainty), the basic principle of logistic regression is to find an equation similar to that used in linear regression to predict the probability of events falling into one category as opposed to another (e.g. the probability of dropout compared to the probability of non-dropout), but to ensure that probabilities cannot be predicted below 0 or above 1. Accordingly, in case of a dichotomous/binary response variable, there is a sigmoidal (S-shaped) relation between the probability of occurrence of a response and the explanatory variables, because, probability can only fall within the boundaries of 0 and 1. In such a situation adopting a specific model that is designed to handle the specific requirements of binary dependent variables is conclusive.

7.1.2 The Log of Odds Transformation in Logistic Regression

In order to ensure a probability constraint (0 to 1), logistic regression makes two transformations to the dependent variable (Y). First, one has to transform the probability into odds and then to transform the odds using logarithms.

Assuming the probability of the event that one or more children in a household are never enrolled in school is equal to p, then it may be said that (1-p) will be the probability of the event where all children are enrolled. Similarly, if one assumes that the probability of the event that one or more children in a household are dropped out of school, then (1-p) will be the probability of the event that all the enrolled children in a household are currently attending school. Odds are a way of presenting probabilities where the odds of an event happening [Odds (Y=1)] is the probability that the event will happen [p(Y=1)] divided by the probability that the event will not happen [1-p(Y=1)], i.e.-

$$\text{Odds}\,(Y = 1) = \frac{P(Y = 1)}{[1 - P(Y = 1)]} \text{-------}(1)$$

Unlike probabilities, which cannot exceed 1, odds have no such theoretical maximum constraint [substituting the P(Y=1) =1 odds stretches to ∞]. This means that as the probability that Y = 1 tends towards 1, the odds become positive and increasingly large and finally stretching into infinity. So within the maximum probability constraint [P(Y=1) =1], the odds do not have the maximum constraint criterion so far. But this has only resolved part of the problem, for odds may be constrained to zero if one substitutes P(Y=1) = 0 in equation (1). To deal with this problem, second transformation has to be performed by taking the natural logarithm (log with a base which has a constant value of 2.72---) of the odds that Y = 1. This log of odds of the event is reffered as logit which is actually the contraction of the term logistic and unit (Fraas, Drushal and Graham, 2002). This is often written as logit (p) and accordingly the equation for logit (p) is -

$$\text{logit}\,(p) = \ln[\text{Odds}\,(Y = 1)] = \ln[\frac{P(Y = 1)}{1 - P(Y = 1)}] \text{------}(2)$$

Combining the above two expressions, one can write –

$$\ln\left(\text{odds}(\text{event})\right) = \ln\left(\frac{\text{prob}(\text{event})}{\text{prob}(\text{non event})}\right) = \text{logit}\,(p) = \ln[\frac{P(Y=1)}{1-P(Y=1)}] \text{------}(3)$$

Taken together, transforming the two extreme limits (1 and 0) of probability that Y = 1 into odds [i.e., substituting 1 and 0 for P(Y=1) in equation-1] and then taking the natural log of the odds [i.e., transforming the odds value in equation-(3)], the transformed variable, logit (p), will have a maximum and minimum values as +∞ to -∞, but the predicted probability will never produce impossible results below 0 or above 1. This means that when the best fitting line is calculated, it cannot produce impossible values of Y since the value of logit Y has no maximum or minimum values.

7.1.3 Logit-link Function - The Regression Equation in the Logit Model

By transforming the dependent variable into the natural logarithm of the Odds (Y=1), the procedure ensures that the logit has no upper or lower limit i.e., it has an unlimited range of values. As such log of odds transformation of the dependent variable ensures that the familiar linear regression equation can be retained with the log of odds of dependent variable (not with the dependent variable itself). Accordingly, with the logit transformation as link function, the regression equation can be written as a linear combination of the predictors (McCullagh and Nelder, 1983) in the model which is as follows-

$$\text{logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = \propto + \beta_1\ X_{i1} + \beta_2\ X_{i2} + \dots\dots + \beta_k X_{ik} - - - (4)$$

where $i = 1,2,3, \dots, n =$ numbers of observations (120 in the present model) and X1, X2, X3 ... are the set of explanatory variables. Although the left hand side of equation (4) differs slightly form that of conventional Multiple Regression (MR) equation, the right hand side of the equation is linear and similar to a MR equation. Here the Log-odds/ logit (p) are a linear function of the predictors. The above equation is identical to the multiple linear model except that the log-odds in favor of Y = 1 replaces the expected value of Y. Instead of multiple linear regression equation, it is called Logistic Regression equation which predicts the log of odds/logit of the dependent variable assuming a linear relationship between the predictors and log of odds/logit of the dependent (Karen Grace-Martin, 2001). However it does not assume the linearity between the predictors and the dependent variable.

Thus it is seen that whereas OLS regression has an identity link function, logistic regression has a logit link function (i.e., logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as OLS regression does). Logistic regression is related to, and answers the same questions as, multiple regression analysis traditionally does with a discrete dependent variable, and even a multiple regression analysis with a dichotomous dependent variable. However, logistic regression is more flexible than the other techniques. So it is not that the logistic regression technique is being applied in the present for the categorical nature of the dependent variable only. In logistic regression, there is no need to assume any particular distributional pattern for the predictor variables in the model. In logistic regression, the predictors do not have to be normally distributed, linearly related, or it even does not require the assumption of equal variance within each group. Logistic regression analysis is especially useful when the distribution of responses on the dependent variable is expected to be nonlinear with one or more of the independent variables.

To solve equation-(4) so that one can return to the probability that Y = 1, the calculation needs to be reversed which turns the probability into odds. This is done by taking the anti-log. Accordingly, by taking the antilog, Equation-(4) may be written as –

$$P = \frac{\exp^{\text{logit}(p)}}{1 + \exp^{\text{logit}(p)}} = \frac{\exp^{\propto + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}}}{1 + \exp^{\propto + X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}}} = \frac{\exp^Z}{1 + \exp^Z} - - - (5)$$

Where, $Z = \propto + \beta_1\ X_{i1} + \beta_2\ X_{i2} + \dots\dots + \beta_k X_{ik}$ = the utility function of the event P and it is nothing but the estimated logit (p) or log odds of the dependent variable. The β terms are the logistic regression coefficients, also called parameter estimates which are actually partial slope coefficients (Gujarati Damodar N., 2003). A greater value of Z implies a greater probability for the event to take place. When Z approaches infinity, P approaches 1, indicating a high likelihood for the event to occur. When Z approaches negative infinity, P approaches 0 thereby indicating a low likelihood for the event to occur and if Z equals zero, the probability is 0.50, implying a 50/50 chance for the event to occur.

7.1.4 Odds, Log (odds) and Probability - A Relationship

A probability and the odds are both measures of likelihood. They have a classic mathematical relationship. Odds are simply a ratio of two probabilities and hence basically a number. As such it can be written as –

$$\text{odss} = \exp^{\ln(\text{odds})} = \exp^{\text{logit}(p)}$$

Substituting $\exp^{\text{logit}(p)}$ by odds, we may re-write equation-(5) as -

$$P = \frac{e^{\text{logit}(p)}}{1 + e^{\text{logit}(p)}} = \frac{e^Z}{1 + e^Z} = \frac{\text{odds}}{1 + \text{odds}} - - - - - - (6)$$

There are two basic reasons underlying the development of the model above. First, probabilities and odds obey multiplicative, rather than additive, rules. However, taking the logarithm of the odds allows for the simpler, additive model since logarithms convert multiplication into addition. And, second, there is a (relatively) simple exponential transformation for converting log-odds back to probability (equation-6). This way of expressing the probability results in understanding the meaning of the regression coefficients. With a one unit change in a particular explanatory variable (Xs), the β_js represent the amount the logit (log-odds) changes holding the effects of others explanatory variables held constant. But this does not make much sense. By transforming the values of the coefficients estimated in equation- (1), one can easily get the logit value or log odds which again may be transformed in to P by using equation- (5). As such every respondent may be given a probability of the event of never enrolment and dropout. Thus, it is seen that the estimated coefficients can easily be transformed so that their interpretation makes sense.

7.1.5 Regression Estimation

The statistical procedure for estimating the parameters (α and βs) of a LR model is quite different from that of ordinary least square method applied to estimate the MLR model. The most used method for estimating the logistic model is maximum likelihood estimation procedure (Hosmer and Lemeshow, 1989 and Ryan, 1997). The maximum likelihood (ML) method is preferred over the weighted least squares approach by several authors, such as Haberman (1978) and Schlesselman (1982). Instead of minimizing the error terms with least squares, the logit coefficients (β terms) are estimated by maximum likelihood (ML), i.e., by searching for that set of βs which will make the observed responses maximally likely, i.e., a set of β that will in general assign a high probability to 1-responses and a low probability to 0-responses. In a process known as iteration, estimates of the parameters are calculated again and again in the hope that they will 'converge' to stable values and will produce the optimal set of βs. The iteration process will not successfully converge if the independent variables are too highly correlated. This is the familiar multicollinearity problem sometimes encountered in OLS regression.

The ML estimator is consistent. As the sample size grows large, the probability that the ML estimator differs from the true parameter by an arbitrarily small amount tends toward 0. The ML estimator is asymptotically efficient, which means that the variance of the ML estimator is the smallest possible among consistent estimators. The ML estimator is asymptotically normally distributed, which justifies various statistical tests. This indicates that the desirable properties of consistency, normality and efficiency will be retained as the sample size approaches infinity (Greene, 2003, ch.17, Judge et al.). Although there is no hard and fast rule for sample size, it is generally suggested that it is risky to use ML with samples smaller than 100 and 10 observation per parameter seems to be reasonable for modeling the logistic regression using ML method (Long Scott J., 1997).

7.2 The Variables in the Logit Model

It has already been stated in chapter-VI that data collection were limited to four villages (two each from one educationally backward and developed blocks) in the district of Uttar Dinajpur, West Bengal. All the villages are almost equally equipped with a primary school within the village, although the other basic amenities like, road, bus connectivity, electricity facility, are quite different. The detail of the same has been discussed in chapter-VI. The respondents were primarily the head of the family and thus comprise the male members of the family.

Logistic regression supports only a single dependent variable and for binary logistic regression, this response variable can have only two categories. Accordingly, two separate models using the same predictor variables are applied. The regression model assumes - i) either a household has at least one children who has never been enrolled or not and ii) either a household has at least one children who has been dropped out of school or it has not.

The description of variables with their notation is presented in a tabular form below:

Variables	Notation	Description
Binary Response		
Never enrolled	PBENRLMNT	1= if at least 1 child within a household was never enrolled, 0= no children within the household were ever enrolled
Dropped out of school	PBDOSCH	1= if at least 1 child within a household dropped out of school, 0= no child within the household dropped out of school
Independent Covariates		
Economic dependency ratio	ECONDEP	Ratio of non-earners to earners at household level
Proportion of educational expenditure	EDNTOTEX	Educational expenditure as a % of total expenditure
Mothers' empowerment	EMPMTH	1= if mothers' empowerment index is greater than or equal to that of father's, 0 = less than that of father's
Fathers' education level	FTHEDN	education in completed number of years
Female work participation	FWP	1= if female members work, 0= no female works
Pattern of income	INREGIRREG	1= if the main income of a household is regular, 0= if the main income is not regular
Mother's education level	MTHEDN	education in completed number of years
Opportunity cost of schooling	OPTNTCOST	1= if there is any children in the age group below 18 years and working, 0= if there is no children in the age group below 18 years and working
Monthly per capita expenditure	MPCE	Monthly per capita expenditure at household level (in multiples of Rs.10)

It is expected that as proportion of educational expenditure to total expenditure (EDNTOTEX), monthly per capita expenditure of the household (MPCE), educational level of father and mother (FTHEDN, MTHEDN) increase, the probability of never enrolled and school dropout will decrease continuously as a function of these variables and households with regular income (INREGIRREG) and with mothers' empowerment will also show the increase in the probability of never enrolled and school dropout. However it is assumed that increasing economic dependency representing the ratio of non-earners to earners (ECONDEP) and household dependency measured as the ratio of old age and sibling members to total number of schooling age children (HHDEP) will lead to an increase in the probability of the response variables and it will also be higher for the households where the female members work (FWP) in the paid market and where there is an opportunity for the children to be engaged as child labour (OPTNTCOST).

7.2.1 Selection of variables

It is important to note here that in a small sample size of 120 households, the number of predictors (10 explanatory variables proposed to be included) in the model is large. A small sample with a large number of predictor variables can cause problem of model 'overfit' in the analysis. Among others, a regression model is a situation where the aim is to find the 'best', most 'parsimonious1', model to predict the dependent variable or explain the variation. In such a situation, the resultant model will more likely to be numerically stable and also more easily to be generalized (Hosmer and & Lemeshow, 2000, pp 92). Hosmerand Lemeshow (2000, pp 95) suggest to perform a univariate analysis of each potential independent variable. Those whose univariate test has a p-value < 0.25 should be considered as a candidate for the multivariable model. Applying this methodology of variable selection, female work participation (FWP) and empowerment of mother (EMPMTH) has been excluded from determining the probability of never enrolment (PBENRLMNT). The table of univariate test is shown in Appendix-1.

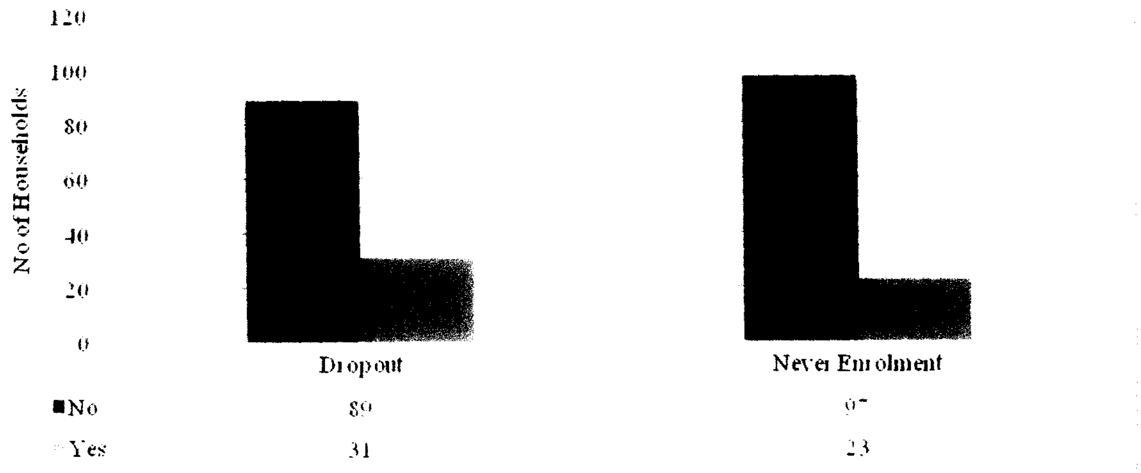
7.3 Reporting and Interpreting Logistic Regression2 Results

The results in different dimensions are being presented below. All statistics reported herein use 3 decimal places in order to maintain statistical precision.

7.3.1 Frequency Table: Binary Responses

Dealing with the two dependent variables in order to observe the magnitude of the problem that is to be predicted in the logistic regression model, calculation of the frequency table (annexed in Appendix-II) of the binary response variable has been undertaken (in SPSS, from Data Editor > Analyze > Descriptive Analysis > Frequencies) and the same has been graphically presented in Figure - 7.1.

Fg-7.1 : Frequency of Binary Response



Thirty one households, out of total 120 households surveyed, were found where one or more children, once enrolled in school had dropped out of school. This represents that in case of 25.8% households, the problem of school dropout is somehow related. On the other hand, the problem of never enrolled is less common than school dropout. Less than 20% (23 out of 120 households) of the households are lagging in this case. This observation makes a comparison between two educationally disadvantaged scenarios. The whole representation briefs the problem as what is the likelihood or probability of a household to have or not have the problem of school dropout or never enrolment so far as the school age children within the household is concerned. The logistic regression technique actually does the job and statistically predicts this likelihood on the basis of some related covariates of the problem.

7.3.2 Evaluations of the Logistic Regression Model

After choosing a logistic regression model, although it would be more appropriate or rather conventional to interpret the logistic regression coefficients, after assessing how effective the model will be or how consistent the model will be with the data. An effective evaluation of the Logistic Regression Model can be judged by reporting on overall model evaluation, goodness-of-fit statistics, statistical tests of individual predictors, and validations of predicted probabilities (Peng, Lee, & Ingersoll, 2002).

7.3.2.1 Goodness-of-fit indicators for overall model/Overall significance of the model

Overall model evaluation is a statistical test where it is investigated whether a logistic model provides a better fit to the data if it demonstrates an improvement over the intercept-only model (also called the null model). An intercept-only model serves as a good baseline because it contains no predictors. An improvement over this baseline is examined by using three inferential statistical tests: the likelihood ratio, score, and Wald tests. In general, the likelihood ratio statistic (often referred to as -2 Log likelihood) is superior to the Wald statistic (in the sense that it gives more reliable results), and accordingly the exercise will concentrate on the likelihood ratio (LR) statistic. (Larsen Pia Veldt, 2008). Before applying this statistic in the model, it is appropriate to define it at this juncture.

The likelihood ratio (not likelihood ratio statistic) is defined as –

$$\text{likelihood ratio} = \frac{LR}{LF} \text{ - - - - - - } (7)$$

The numerator (LR) corresponds to the maximum probability (likelihood) of an observed outcome under the null hypothesis (i.e. under the reduced model). The denominator (LF) corresponds to the maximum probability of an observed outcome under the full model (i.e. varying parameters over the whole parameter space). By taking natural log on both sides of equation-(7), it follows as -

$$\text{Log likelihood ratio} = \text{Log LR} - \text{Log LF}$$

This implies that –

$$\begin{aligned} \text{Likelihood ratio statistic (LR Statistics)} &= -\ 2 \text{ Log likelihood ratio} \\ &= -2(\text{Log LR- Log LF}) \\ &= (-\ 2 \text{ Log LR}) - (-2 \text{ Log LF}) \end{aligned}$$

It is very evident that a regression model provides a better fit to the data if the likelihood of the reduced model is less than the likelihood of the full model. Hence, a smaller likelihood ratio indicates a better fit model. If the likelihood ratio becomes smaller (within the range of 1 to 0) the natural log of the likelihood ratio will become negative with greater magnitude and hence, minus the natural log of the likelihood ratio will be a bigger positive number3. So it is twice minus the natural log of the likelihood ratio. It turns out that under the null hypothesis, the LR Statistic [−2 log (LR / LF)] has an approximate chi-square distribution, with degrees of freedom equal to the number of explanatory variables in the model (Hosmer and Lemeshow, 1989). A finding of significance indicates that the model with the predictors is significantly different from the model with the intercept only.

In SPSS version 13, the table Iteration History in Block-1 of logistic regression output provides the initial – 2 Log Likelihood and also the – 2 Log Likelihood under full model. The LR Statistics (model chi-square) is calculated by subtracting the model – 2 Log Likelihood from initial – 2 Log Likelihood. In SPSS, the output referred as Omnibus Tests of Model Coefficients gives the model chi-square. In the present analysis, the likelihood ratio test results a significant chi-square value for both the model (Table-7.2).

Table-7.2: Overall Model Evaluation (Omnibus Tests of Model Coefficients)

Test	Chi-square	df	p	Model
Likelihood ratio test	56.880#	10	.000	PROBDOSC
	60.220##	8	.000	PRENRLMNT
Score test	38.990	10	.000	PROBDOSC
	41.424	8	.000	PRENRLMNT

# (initial – 2 Log Likelihood = 137.117, model – 2 Log Likelihood = 80.233, hence LR Statistics = (137.113 – 80.233) = 56.880
## (initial – 2 Log Likelihood = 117.271, model – 2 Log Likelihood = 57.052, hence LR Statistics = (117.272– 57.052) = 60.220

A finding of significance, as mentioned above, indicates that at least one of the predictors is significantly related to the response variable. One can estimate logistic models using block entry of variables or any of the stepwise methods (forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald) as suggested in SPSS version 13. However, in the present model block entry of variables has been applied in the estimation process and as such there is no difference in the result of Omnibus Tests in step, block, or model chi-square values.

7.3.2.2 Hosmer and Lemeshow goodness of fit

Goodness-of-fit statistics assesses whether the fitted model adequately describes the observed outcome experience in the data (Hosmer and Lemeshow 2000).This statistic measures the correspondence between the actual and predicted values of the dependent variable. In this case, better model fit is indicated by a smaller difference in the actual and predicted values on the dependent variable. The Hosmer-Lemeshow test groups observations into deciles based on predicted probabilities. In SPSS this is shown by Contingency Table for Hosmer and Lemeshow Test, which for the present model is shown in Appendix-III. It then computes a chi-square from observed and expected frequencies. A probability (p) value is also computed from the chi-square distribution4 with J-2 degrees of freedom (where j= number of groups dividing the observations) to test the fit of the logistic model. A good model fit is indicated by a non-significant (if the significance value is more than 0.05) chi-square value indicating that the model prediction is not significantly different from observed values. This inferential goodness-of-fit test yields a non-significant chi-square value for both the models in our analysis. i.e., both the model shows no evidence of lack of fit based on the Hosmer and Lemeshow chi-square test of goodness of fit statistic.

This statistic is the most reliable test of model fit for SPSS binary logistic regression, because it aggregates the observations into groups of "similar" cases and the test is considered more robust than the traditional chi-square test, particularly if continuous covariates are in the model or sample size is small which seems to be applicable for the present model (Garson G. David, 2009).

Table-7.3: Hosmer and Lemeshow Goodness-of-Fit Tests

Model	H-L Statistic	df	Sig
PROBDOSC	7.451	8	.489
PRENRLMNT	2.217	8	.974

7.3.3 R^2 for Logistic Regression

In OLS, to assess how well a model explains the data, R-square statistics or the coefficient of determination indicates the proportion of variation in the dependent variable explained by predictors in the model. But there is no equivalent statistic in logistic regression (Cohen Jacob, 2003) that makes a sense that proportion of variance accounted for as R^2 does in OLS. Nonetheless, a number of measures (SPSS uses R^2 like measures as Nagelkerke and Cox and Snell, EViews uses McFadden R-squared) have been proposed in logistic regression as an analog to R^2 of multiple linear regression and these are some times referred as Pseudo R-squares.

One widely accepted method is offered by Cox and Snell (1989), but it suffers from the problem that the index of measure as suggested by them does reach the maximum value of 1. However, the value of the index as found in the present two referred models is presented in Table-7.4. Nagelkerke (1991) has tried to adjust the problem of restricted maximum value of R^2 index as developed in Cox and Snell measure. As such the value of Nagelkerke R^2 statistics is always fairly higher and the same has been represented for the present models in the same Table-7.4.

Table-7.4 Model Summary

PROBDOSC				PRENRLMNT			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square	Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	57.052(a)	.395	.633	1	80.233(b)	.377	.554

Note: a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.
b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

However McFadden R^2 is another measure of the goodness-of-fit of a logistic model that corresponds to pseudo R^2 as a measure of goodness-of-fit of models estimated by OLS (Krznar, 2004). McFadden R-squared is the likelihood ratio index computed as –

$$1 - \frac{LL(\beta)}{LL(\beta_0)}$$

where LL(β) is is the maximized value of the log likelihood function and LL(β0) is the restricted log likelihood or the maximized log likelihood, when all slope coefficients (except for the constant term) are restricted to zero. It has the property that it always lie between zero and one. A closer value of the statistic to 1 indicates a better fit model. From Table- 7.4, in both the models, the values are moderately high (0.415 for probability of dropout and 0.514 for probability of never enrolment). Combining all the results, it may however be said that the regression equations fitted for the two models is moderately accounted for explaining the dependent responses. But, based on these statistics, it would not be very scientific to say much about the variance of the dependent explained by the predictors in the model (Long, 1997, pp. 104–109; Menard, 2000). In addition, these measures of model fit does not bear correspondence to predictive efficiency or can be tested in an inferential framework (Peng, Chao-Ying Joanne Lee, Kuk Lida Ingersoll, Gary M., 2002).

7.3.3.1 Predictive accuracy of the Model: (Expectation-prediction table)

After assessing the significance of the variables in the model, one can explain the Predictive accuracy of the model. The predictive success of the logistic regression can be assessed by looking at the classification table, showing correct and incorrect classifications of the binary dependent variable. The logistic regression predicts the log of odds of the dependent which is the natural log of the odds (or probability/ [1–probability]) and hence it can be transformed back to the probability scale. The predicted probabilities can then be compared with the actual outcome to measure the validations of predicted probabilities. The degree to which predicted probabilities agree with actual outcomes is expressed as a classification table and vis-à-vis the present analyses, it is shown in Tables 7.5 & 7.6. From the Table 7.5, it is calculated that in determining the probability of dropout, 80 observations with the y=0 (nonevents) and 20 observations with the y=1 (events) are correctly classified by the estimated model. Now, the proportion of observation with y=1 that are correctly predicted is termed the sensitivity, while the fraction of y=0 observations that are correctly predicted is known as specificity. Overall, the estimated model correctly predicts 83.33% observations with a sensitivity of 64.52 percentage points and a specificity of 89.89 percentage points. The estimated model improves on the Dependent =1 predictions by 64.52 percentage points, but does more poorly on the Dependent=0 predictions (by 10.11 percentage points). Overall, the estimated equation is 9.17 percentage points better at predicting responses than the constant probability model. This represents a 35.48 percent improvement over the default model.

Table-7.5: Expectation-prediction table for Dependent Variable: PBDOSC

Dependent Variable: PBDOSC

Sample: 1 120

Included observations: 120

Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	80	11	91	89	31	120
P(Dep=1)>C	9	20	29	0	0	0
Total	89	31	120	89	31	120
Correct	80	20	100	89	0	89
% Correct	89.89	64.52	83.33	100.00	0.00	74.17
% Incorrect	10.11	35.48	16.67	0.00	100.00	25.83
Total Gain*	-10.11	64.52	9.17			
Percent Gain**	NA	64.52	35.48			

Similarly, in determining the probability of never enrolment, the estimated model correctly predicts 90.00% observations with a sensitivity of 65.22 percentage points and a specificity of 95.88 percentage points (Table-7.6). Here, the estimated equation is 9.17 percentage points better at predicting responses than the constant probability model.

Table-7.6: Expectation-prediction table for Dependent Variable: PBENRLMNT

Dependent Variable: PBENRLMNT
Sample: 1 120
Included observations: 120
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	93	8	101	97	23	120
P(Dep=1)>C	4	15	19	0	0	0
Total	97	23	120	97	23	120
Correct	93	15	108	97	0	97
% Correct	95.88	65.22	90.00	100.00	0.00	80.83
% Incorrect	4.12	34.78	10.00	0.00	100.00	19.17
Total Gain*	-4.12	65.22	9.17			
Percent Gain**	NA	65.22	47.83			

7.3.4 Multicollinearity in the Model

Presence of collinearity causes similar problems in logistic regression as it causes in linear regression model. The maximum-likelihood estimation is not too accurate in the case of multicollinearity and the logistic model becomes unstable when there exist strong dependence among the predictors (Hosmer and Lemeshow, 1989 and Ryan, 1997). The detection of collinearity in linear regression model (e.g. tolerance statistics, Eigen values, etc.) is straightforward and available in most of the statistical packages (SPSS, STATA, Eviews). But no such standard statistics for logistic regression is available in the popular statistical packages so far. However, Menard, Scott (2002, page 76) has suggested to run an OLS regression model using the same dependent and independent variables that have been used in logistic regression model. He notes, "Because the concern is with the relationship among the independent variables, the functional form of the model for the dependent variable is irrelevant to the estimation of collinearity." The detection and diagnosis of collinearity in logistic regression in a similar way as linear regression have also been discussed in Hosmer and Lemeshow (1989). It is suggested that large standard errors associated with the logistic coefficients could be a collinearity warning.

Diagnostic information for multicollinearity (tolerance statistics, VIFs and also eigen values) for the predictors has already been calculated and shown in chapter-VI and no serious concern of collinearity was found so far. In addition to this, the standard errors associated with the logistic coefficients both for the two models are not very large (ranging from a lowest of 0.006 to a highest of 1.346; Table7.7 & Table-7.8). Hence, by prior selection of independent variables following a reasonable criterion and by investigating the standard errors it is expected that statistically fair estimates will be derived in the prescribed model. Sometimes, presence of collinearity tends to produce the β coefficients unreasonably high. It is suggested that (Menard, 2002, page-76) an un-standardized logistic coefficient greater than 2 or a standardized logistic coefficient greater than 1 is a caution for the presence of collinearity among the regressors. This criterion is also almost satisfied (Table7.7 & Table-7.8) in connection with the present models.

7.4 Interpretation of Logistic Regression Coefficients

After observing the Significance Tests (likelihood ratio, score, Wald, and Hosmer & Lemeshow tests) that suggests no such lack of model fit, it is time to concentrate on the basic results of regression where it will be seen the significant factors explaining the probability of dropping out of school and never enrolment can be identified. It may be noted here that the reference age group of the children has been assumed as 5 to 14 years. For each observation, the data available are the households grouped separately in two opposite categories (coded as 1 if at least one never enrolled or dropped out child belongs to the household, 0 if no5) along with a series of independent covariates stated earlier. By default, in SPSS, the logistic procedure will predict the "1" category of the dependent variable, making the "0" category the reference category. It is of primary concern to target those households where the event of school dropout or never enrolment occurs. This occurrence, as discussed earlier, is a chance expressed in probability manner when logistic regression equation is fitted. So, identifying the factors that determine this chance is the primary concern.

In the present analysis, the binary response variable in the first model (Model-I) is school dropout and in second model (Model-II) the response variable is never enrolment in school. In SPSS output, the parameter estimates appear in the "B" column of the "Variables in the Equation" table. Table 7.7 presents the results for the predicted logit of school dropout and Table7.8 for the never enrolment. Each Table shows the estimated logistic coefficient (β coefficients) its standard error (S.E.), the Wald statistic, degrees of freedom, p significance level for the constant and each predictor in the model. By observing the Wald chi-squared statistics, the result shows that the important variables contributing to the probability of dropping out of school are- proportion of educational expenditure (EDNTOTEX), Pattern of income (INREGIRREG), Mother's education level (MTHEDN) and Opportunity cost of schooling (OPTNTCOST).

Hence, the estimated logit equation can be written as -

Predicted logit of (dropout) = (2.088 - EDNTOTEX x 0.258 - INREGIRREG x 1.350 - MTHEDN x 0.611+ OPTNTCOSTx 1.476)

Table-7.7: Variables in the Equation- (Model-I)

		B	S.E.	Wald	df	Sig.	Exp(B)
1		2	3	4	5	6	7
Step 1(a)	ECONDEP	.066	.363	.033	1	.856	1.068
	EDNTOTEX	-.258*	.107	5.784	1	.016	.773
	EMPMTH	.743	.640	1.347	1	.246	2.102
	FTHEDN	-.086	.134	.409	1	.522	.918
	FWP	-.363	.611	.352	1	.553	.696
	INREGIRRE	-1.350*	.649	4.330	1	.037	.259
	MTHEDN	-.611*	.261	5.469	1	.019	.543
	OPTNTCOST	1.476*	.594	6.172	1	.013	4.374
	MPCE	-.036	.037	.952	1	.329	.964
	HDEP	-.215	.615	.122	1	.727	.807
	CONSTANT	2.088	1.540	1.839	1	.175	8.067

a Variable(s) entered on step 1: econdep, edntotex, empmth, fthedn, fwp, inregirreg, mthedn, optntcost, mpce, hdep.

Similarly, from the regression coefficients and its associated significance level in Table- 7.8, the Wald chi-squared values for Fathers' education level (FTHEDN), Monthly per capita expenditure (MPCE) and Household dependency ratio (HDEP) are significant in predicting the probability of never enrolment. It may be noted that the variables were not at all significant in explaining the log of odds of the probability of dropout. It signifies that the socio-economic correlates are not the same for explaining the school level outcome attributes (e.g. dropout and never enrolment are being predicted significantly not by the same variables). Secondly, it may also be said that the household related socio-economic variables are of not much importance in determining the enrolment/non-enrolment decision taken by the households. This opens the scope of including supply related schooling facilities as additional variables in determining the enrolment decision of the households. However it is beyond the scope of the present analysis. The estimated logit equation in predicting the probability of never-enrolment can be written as -

Predicted logit of (never enrolment) = (2.950- FTHEDN x 0.660 - MPCE x0.159 + HHDEPx2.179)

Table-7.8: Variables in the Equation- (Model-II)

		B	S.E.	Wald	df	Sig.	Exp(B)
1		2	3	4	5	6	7
Step 1(a)	ECONDEP	.455	.437	1.087	1	.297	1.576
	EDNTOTEX	.040	.108	.138	1	.710	1.041
	FTHEDN	-.660*	.277	5.693	1	.017	.517
	INREGIRREG	-1.265	.804	2.477	1	.115	.282
	MTHEDN	.126	.232	.293	1	.588	1.134
	OPTNTCOST	-.642	.741	.749	1	.387	.526
	MPCE	-.159*	.067	5.720	1	.017	.853
	HDEP	2.179*	.928	5.505	1	.019	8.833
	CONSTANT	2.950	2.199	1.799	1	.180	19.102

a Variable(s) entered on step 1: econdep, edntotex, fthedn, inregirreg, mthedn, optntcost, mpce, hdep.

The sign of the estimated β coefficients (indicating sign of partial effects of each predictor) corresponding to the significant variables is very important in analyzing the logit result. According to the first model (Table-7.7), the log of the odds of a child dropping out of school is inversely related to proportion of educational expenditure (EDNTOTEX), Pattern of income (INREGIRREG) and Mother's education level (MTHEDN) and and positively related with Opportunity cost of schooling (OPTNTCOST). In other words, the higher the proportion of expenditure on education, the less likely it is that a child would be dropped out of school. Similarly, the higher the chance of a household to have regular income, less likely it is that a child would be dropped out of school. The households where the level of mother's education is higher, it would be less likely that the children of the household will drop out of school. Again, in a situation where there is an opportunity of children to be engaged in child labour market, there will be a higher chance that a child is dropped out of school.

On the other hand, the log of the odds of a child being never enrolled in school is inversely related to Fathers' education level (FTHEDN) and Monthly per capita expenditure (MPCE) thereby representing that a higher level of father's education and monthly per capita expenditure of the households would lower the probability of never enrolment of the children and the same would be higher if household dependency on the children of schooling age becomes higher.

To determine the significant factors by observing the Wald Chi-square value and to make a judgment over the direction of relationship by assessing the sign of the estimated β coefficients, is just beginning of interpretation of logistic regression coefficients. Contribution of significant explanatory variables to the dependent variable is the central point to be discussed in the interpretation of logistic regression coefficients.

Logistic regression has a logit link function and accordingly, in any logistic regression equation, the parameter estimates (β coefficients) associated with explanatory variables represent contributions to estimated log-odds (logits are the log odds of the event occurring). To put it differently, the estimated β coefficient assigned for a particular independent variable measures the change in estimated logit in favor of $Y = 1$ with one unit change in that particular variable on average, with other independent variables held constant. For example, for a 1 % increase in proportion of educational expenditure, the estimated logit that the households will have at least one dropped out child decreases by a factor 0.258 controlling for other variables present in the model (from Table 7.7). Thus, the logistic regression calculates changes in the logit or log of odds of the dependent variable, and not changes in the dependent variable itself as OLS regression does. Also, if all predictors are set equal to 0, the predicted log-odds in favor of $Y = 1$ would be the constant term α. The β coefficients vary between plus and minus infinity, with 0 indicating the given explanatory variable does not affect the logit (that is, makes no difference in the probability), whereas, a positive or a negative β coefficient indicates that the explanatory variable will increase or decrease the logit of the dependent. It is not natural to think in terms of logit/log-odds as because it sounds more statistical than real life expression. One actually needs to provide information that can be used to judge the practical significance of these parameter estimates.

As an alternative of representing the β coefficient directly, the parameter estimates of a logistic regression can be interpreted in terms of odds ratios which is simply the exponential transformation of β coefficients, i.e., odds ratio = exp(β) which implies that β = ln(odds ratio). This exponential transformation of β coefficients i.e., exp (β) is termed as odds ratio and this transformed "coefficient" is pretty useful and more easily explainable. So, if one takes the exponent constant (about 2.72) and raises it to the power of β then the odds ratio is obtained. For example, if the regression result indicates the partial regression slope (β coefficient) is 0.75, the odds ratio is approximately 2.12 (because, $exp^{0.75} = 2.72^{0.75} = 2.12$). In SPSS, odds ratios appear as "Exp(B)" in the "Variables in the Equation" table.

In the present analysis of result, the Odds Ratios will be used to interpret the result. In the analysis, the dependent variable (both) has been coded as 0, 1. Thus the higher category has been predicted and the lower category is used as the comparison of reference by default. The odds ratios corresponding to the estimated coefficients are shown in the last columns of Table-7.7 & 7.8. It may be of worth to note here that interpretation of odds ratio corresponding to any quantitative variable (continuous covariates), can simply be expressed as a percent change in odds, i.e. for continuous variables, the odds ratio represents the change in odds (p/1-p) for a one-unit change in the independent variable controlling for other variables in the model. Complexity arises when all or some independent variable/s are also dichotomous (our model contains both continuous and dichotomous independent variable). Odds ratios will differ depending on whether a dichotomy is entered as a dichotomous covariate or as a categorical variable in the estimation procedure. If a dichotomy is declared categorical, then the prediction is for the lower category and the higher category is the reference. If, on the other hand, the dichotomous independent is left as a dichotomous covariate, then the prediction is for the higher category (usually 1 category, reference category is the lower category usually the 0 category). For example, if covariate coding of sex (0=male, 1=female), and if the odds ratio is 1.751, we can say that the odds of occurring the event compared to non-event are increased by a factor of 1.751 for being female rather than male, controlling for other variables in the model.

In entering such independent variables Mothers' empowerment (EMPMTH), Pattern of income (INREGIRREG) and Opportunity cost of schooling (OPTNTCOST) in the model, preference is to enter them as a dichotomous covariate, coding '1' in favour of the event, and '0' for the non-event. This methodology of interpretation of the continuous and dichotomous independent variables has been followed in the present analysis. In general, the β coefficients can vary between plus and minus infinity, with 0 indicating the given explanatory variable does not affect the logit6 (that is, makes no difference in the probability of the dependent value equaling the value of the event, usually 1); whereas, positive or negative b coefficients indicate the explanatory variable increases or decreases the logit of the dependent.

7.4.1 The Use of Odds Ratio in the Analysis: A Discussion

7.4.1.1 Household Expenditure

While examining the effect of households' expenditure (assumed to be a proxy of income of the households), the partial slope coefficient in Table-7.8 shows that monthly per capita expenditure is reasonably significant exerting its negative impact on the logit of never enrolment and hence on the probability of never enrolment too.

But the important question is how much variance in dropout is explained by the MPCE. The estimated β coefficient here is as low as -0.159 that results an odds ratio equal to 0.853. With respect to odds, the influence of each predictor is multiplicative. Thus, for each 1 unit (Rs.10.00, because MPCE is measured in multiples of Rs.10) increase in Monthly per capita expenditure (MPCE) the predicted odds of never enrolment decreases by a factor 0.147 (1 – 0.853). This explains that a 1 unit increase in MPCE (Rs. 1.00 because MPCE is measured in rupees) will lower the odds of the event that at least one child within a family is never enrolled in school by 14.7%.

Educational expenditure as proportion of total expenditure (in percentage term), on the other hand, is found to be significant having a negative sign too in predicting the event of dropout only. In other words we can say that higher the expenditure on education as proportion of total expenditure, the less likely it is that a child would be dropped out from school within a household. The odds ratio is = Exp (β) = $e^{-0.258}$= 0.773 which explains that for each 1 per cent increase in Educational expenditure, the odds of dropout of school by one or some children within a household decrease by 22.7% (1 – 0.773) or by a factor 0.227, adjusting for other variables in the model. Comparing the above, one can say that MPCE, which is calculated from total annual expenditure of the family (crude measure of family income), remains significant in determining the probability of never enrolment but has no impact in determining the probability of children's dropout from school. While proportion of educational expenditure, although significant in determining the probability of children's dropout from school, has practically insignificant in determining the probability of never enrolment.

7.4.1.2 Parental Education

Educational level of parents is entered in the model as a continuous covariate measured in terms of completed years of education. A higher level of parental education is expected to lower the chance of probability corresponding to the occurrence of both never enrolment and dropout of school in our model. The Wald chi-squared values for father's education are significant with its expected sign in explaining the probability of never enrolment while mothers' education remains insignificant here. Whereas, mothers' education has emerged as statistically significant also with expected negative sign of β coefficient in explaining the event of some of the children is dropped out of school but insignificant in predicting the probability of occurrence of the event of never enrolment/enrolment. This suggests that an educated father is more active in taking the decision that the children will be enrolled in school, while an educated mother takes the active part in ensuring the continuation of the children's education or dropping out of school.

The odds ratio (Table 7.8) corresponding to β coefficient of father's education is 0.517 (<1) which indicates that the odds of never enrolment compared to all children enrolled decreases by a factor of 0.483 (1-0.517) for each one year of additional increase in level of father's education, controlling for other variables in the model. Similarly, the odds ratio (Table-7.7) corresponding to β coefficient of mother's education is 0.543 (<1) which indicates that the odds of dropping out compared to all enrolled children attending school decreases by a factor of 0.457 (1-0.543) or by 45.7% for each one year of addition to level of mother's education, controlling for other variables in the model.

7.4.1.3 Nature of income

Nature of income has been assumed to be a dichotomous covariate in our model. It has been assigned with the score 1 for the households with regular pattern of income and 0 if the main income of the family is irregular. The β coefficient of this variable (-1.350) is significant for predicting the probability of dropout. The corresponding odds ratio is 0.259. We would therefore say that the odds of one or some children within a household compared to all enrolled children attending school are decreased by a factor of 0.259 when the respondent's main income is regular compared to those with irregular income pattern, controlling for other variables in the model. Briefly, it suggests that in households with regular income pattern, less likely will be the chance of school dropout and vice-versa provided the other covariates of dropping out remains the same.

7.4.1.4 Opportunity Cost of Schooling

The availability of work opportunity where the children may be engaged is often reason enough for households to withdraw their child/children from school and send them to the labour market as wage labourers. In economic terms, this is opportunity cost of sending the children (i.e. the income to be accrued from child labour) which has to be sacrificed if instead the parents send their ward/s to school. This opportunity cost of schooling is not directly calculated here. Instead, a dummy has been introduced in the form of yes/no category. If it is found that one or some schooling age children in a household work in the labour market then the household was given a score 1 and 0 otherwise with the assumption that it will have an adverse impact on schooling outcome. The variable is found to be significant in predicting the dropout pattern of the children within a household although it is not found to be significant in predicting the enrolment decision. The official age (5+ years) of child to be enrolled in the first grade of any primary school in this state does not actually appear to be favourable for a child to work as a wage earner. Accordingly, statistical result also supports the logic. However, in the advanced stage of primary grades and during the upper primary schooling, the enrolled children may be withdrawn for financial support of the family. The regression coefficient has the positive sign (1.476) which indicates that the presence of job opportunity for children will increase the likelihood of dropping out of school. The corresponding odds ratio is 4.374 representing that the problem of dropout of school will be four times more likely if there is job opportunity of children for which the households send their ward/s to work rather than to send their ward to school.

7.4.1.5 Household Dependency

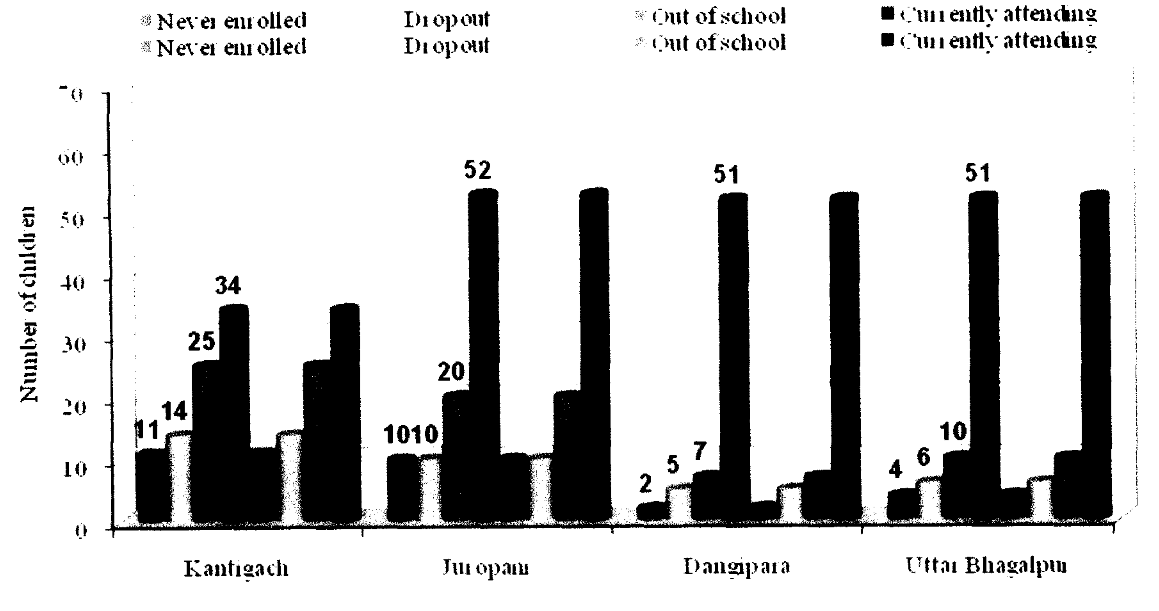
Household dependency has been constructed as a ratio between the total number of siblings and old age members to total number of schooling age children [(siblings+old age)/ children] and introduced in the regression equation as a continuous covariate with the assumption that it will adversely affect the schooling behavior/decision. Statistical findings indicate that this non-financial burden shouldered by the schooling age children in a family is significant in predicting the probability of never enrolment while making it insignificant for the dropping out of school as such. The regression coefficient (2.179) and corresponding odds ratio (8.833) suggests that for each one unit of increase in household dependency ratio there will be an increase in the odds of never enrolment compared to all children enrolled in school by about 9 times.

7.5 School Enrolment- Parental Viewpoint: A Qualitative Analysis of Survey Data: Reasons for Enrolment

Enrolment and dropout are two important educational outcomes that have a close correlation with the parental behavior. Actually, to enroll a child in school is mostly a parental decision guided by some socio-economic correlates. The event is realized partly by the consciousness of the parents about the benefits of education (both direct and indirect). This section is primarily concerned with "Why" a child is being enrolled by the parents. In order to investigate this particular parental behavior, it was of special interest to review the response of the guardian on the reasons for enrolling children to school, based on the interviews with the respondents (parents). This qualitative analysis may provide some additional information on parental behavior for the education of their children. This also may convey some of the parental consciousness on the direct and indirect benefits of education.

The survey conducted in four villages and thirty households in each village were selected randomly from all the households in each village. As such 120 households comprise the sample size of the study. In these 120 households, there are total 250 children in the age group of 5-14 years of which 51.2% (total - 128) belongs to the primary age group and the remaining 48.8% (total- 122) in the upper primary age group (9+ to 14 years). Out of these total children, 89.2% (223 in total) have been once enrolled in school and they have been declared as literate. The remaining 27 children have never been enrolled in any school. These 27 children come from among 23 households for which the response on the question whether one or some children has/have never been enrolled in school was positive and coded accordingly as '1'. For the remaining 97 households the responses have been negative. Now, of the total enrolled children (223), all are not currently attending the school. It was found that 35 enrolled children were not currently attending school. They are literate but creating another educational problem as they had dropped out of school before completing the eight years of schooling. These 35 children came from 31 households for which the response on the question whether one or some children has/have been dropped out of school was positive and coded accordingly as '1'. These two are the dependent binary responses which have been separately predicted from a series of household related socio-economic covariates in the logistic regression analysis. The empirical findings have already been discussed in the previous section of this chapter. However the current status of the children is represented in Figure-7.2.

Fig:7.2: Status of Schooling age children



Source: - calculated from Field Survey

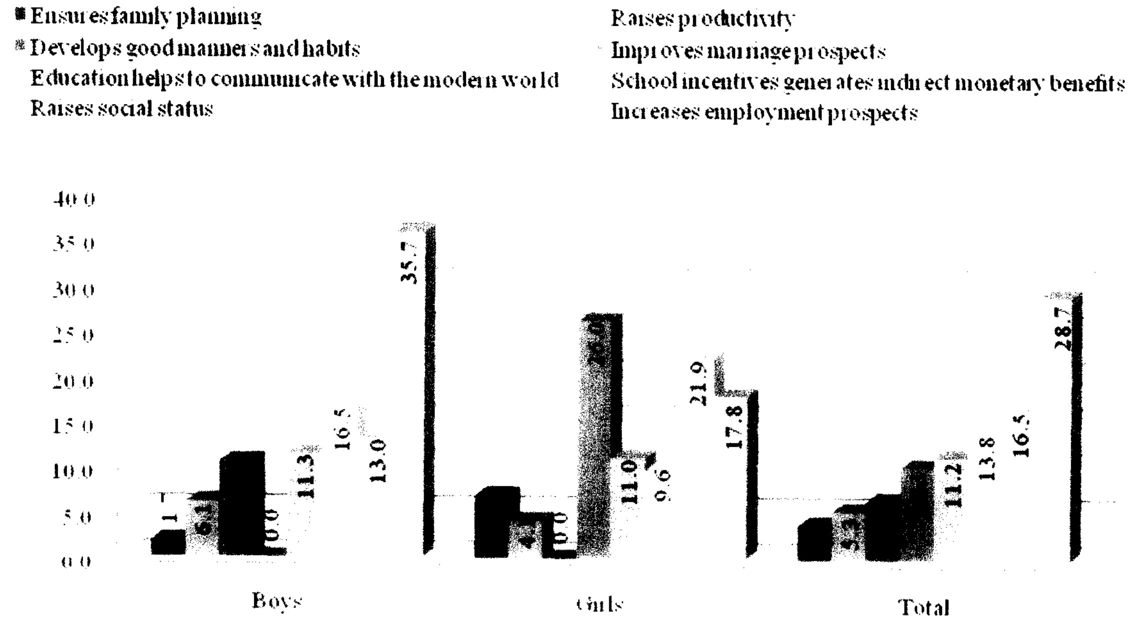
In absolute percentage, 24.80% (62 out of 250) children are found to be out of school of which 10.80% (27 out of 250) are never enrolled and the remaining 14.00% (35 out of 250) are dropped out of school. The problem of both never enrolment and dropout of school is more prominent in Kantigach followed by Juropani and Uttar Bhagalpur. In Kantigach, 42.37% children are out of school followed by Juropani (27.78%). It is earlier noted in these two villages, children were seen to be engaged in labour market for the availability of work. The empirical findings in the previous section thus support this view.

It has been an experience that most of the Indian parents (more than 90%) want their children to be educated (PROBE Report, 1999 The Pratichi Education Report, Number-I, 2002). But in practice, this aspiration is not usually seen to be realized. In order to find the reasons of this gap, it will be interesting to know the parental view on the question that why they get their children admitted to school. The reasons that have been mostly considered by the parents in this respect have been collected from the survey. For the purpose of this, a specific question was asked where the respondents were asked to cite three reasons for enrolment of their children to school. This question was built-in with some ready responses broadly categorized as economic and non-economic reasons. Among the economic reasons, the responses were incorporated with the following alternatives - increases employment prospects, school incentives generates indirect monetary benefits, enhances earnings, raises productivity and others. The non-economic response categories with predetermined alternatives such as - raises social status, improves marriage prospects, develops good manners, ensures family planning, enhances the feeling of security, education helps to communicate with the modern world and others. If a respondent fails to cite any answer, his/her response was recorded as no reasons given. The reasons that a particular respondent cited have been ordered and the first one for each of the child is collected. Among the above two broad categories, eight types of responses were cited as first reason for enrolling the children by the respondent apart from the reason 'no reasons given'. As such, information collected for this purpose for 188 children in total who are currently attending school from the sample villages. Out of this, 115 children are boys and 73 are girls. The information collected in this respect has been depicted in Figure-7.3.

It is seen from this figure that there are several socio-economic issues that have been considered by the parents in sending their children to school. In total, most of the children (around 28.7%) are getting enrolled considering their better future employment prospect. This particular reason appears to be more important for the boys' education (35.7%). The second major reason (for 16.5% children) that has mostly been considered by the rural parents is that education will raise social status of the family. This social aspect is mostly associated with girls' education (21.9%).

In the recent period, especially after the introduction of DPEP, school incentives for the children in the form of free text-book, uniform, mid-day-meal etc. have been introduced in the government run schools with the objective that these will create an indirect economic benefit for the parents in terms of opportunity cost of schooling and accordingly it will also enhance the school enrollment and retention rate. As such it is considered as an economic reason, rather than an institutional one. Our objective is also to have an idea about the explanatory power of this issue. It has been observed that out of total 188 enrolled children, 26 (13.8%) were enrolled and continued with education with the consideration of this reason. Again education as a social variable that helps to communicate with the modern world is also taken to be an important consideration (for 11.2%) on the part of the parents in enrolling their children in school. A major reason cited by the parents (for 9.5%) for the education of their girl children. These four major reasons together have been considered as the basic reasons (for around 70% of the total children) for the enrolment of children in school. Apart from these, four other socio-economic issues are also being considered in this respect. The detail has been shown in Figure -7.3.

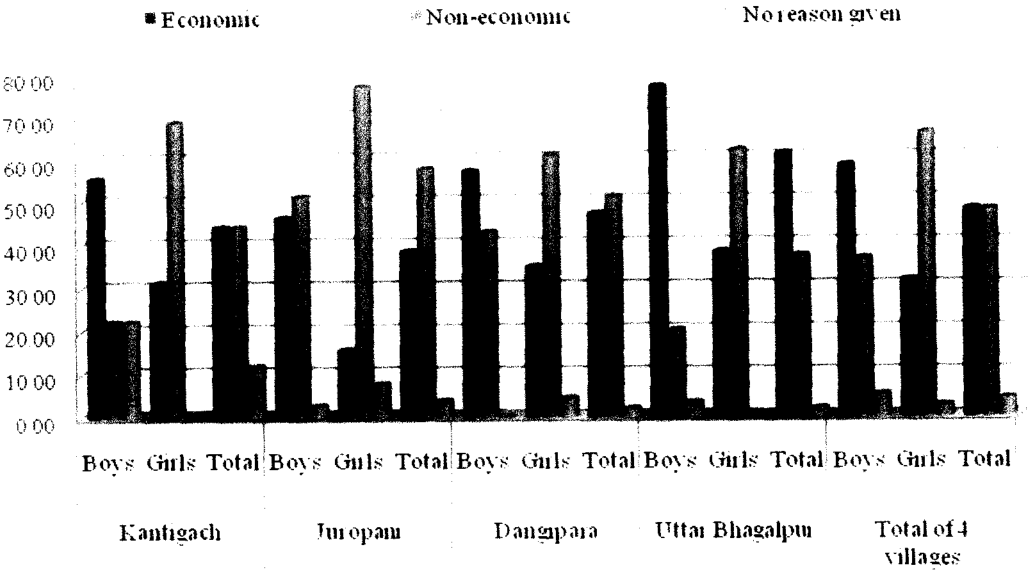
Fig-7.3: Socio-Economic Reasons for School Enrolment



Source: - calculated from Field Survey; N= No. of responses = 188

All the cited reasons when classified into economic and non-economic categories, produces an interesting result. Among the enrolled girl children (73 in total), 65.75% (48in total) have been enrolled in school for the reasons associated with non-economic purpose. It may however be noted here that the non-economic reasons are being considered with a view to comprehending the future benefits of education. These are, in a sense, indirect benefit accruing from educating a child. This consideration is more important for bringing a sound educational development in an area. However, this consideration varies from village to village and of course from household to household. A village level data on the reasons cited by the parents broadly categorized into economic and non-economic reasons are shown in Figure-7.4.

Fig-7.4: Reasons for School Enrolment- Economic & Non-economic (in %)



Source: - calculated from Field Survey; N= No. of responses = 188

It is seen from the village level data that in all the villages, boys' education is mostly connected with economic reasons, while the girls are enrolled with the consideration of non-economic benefits from education. Girls' education is also viewed for economic reasons too. But its intensity is smaller than the boys. Summarily, in all the study villages (irrespective of the sex of the children), however, the non-economic reasons are being given equal importance while getting the children admitted in school except in Uttar Bhagalpur village. Figure-7.4 provides an additional important observation. In spite of low educational background of the parents, only eight in case of eight children, respondents were not able to cite any reasons for the purpose.

Note

1. Parsimony is also a factor in statistics. In general, mathematical models with the smallest number of parameters are preferred as each parameter introduced into the model adds some uncertainty to it. Additionally, adding too many parameters leads to "connect-the-dots" curve-fitting which has little predictive power. In general terms, it may be said that applied statisticians (such as process control engineers) value parsimony quite highly.
2. The logistic regression analysis has been carried out by the Binary Logistic procedure in SPSS/version 13.0 (SPSS Inc. 1989-2004) in the Windows XP Professional version 2002. However, in order to evaluate the logistic regression model which needs different significance tests for binary logistic regression, the statistical package EViews version 3.1 (Quantitative Micro Software, 1994-98) is also used.
3. But one should not make it as an objective sense as the LR Statistic is a cumulative measure across all cases and its size is therefore highly dependent on number of observation.
4. Hosmer and Lemeshow (1989, p.141) report that extensive simulation indicates that the distribution when the model is correct is well approximated by a χ^2-distribution.
5. We have coded 1 as a response for those households where at least one child, once enrolled, currently not attending school. It may be the case for the households where more than one child is not currently attending school. If such response coded separately, then multinomial logistic regression would have to be applied.
6. When β =0, Exp (β) =1, and an odds ratio of 1 corresponds to an explanatory variable represents that it does not affect the dependent variable. If β moves to positive infinity then exp (β) will also moves to positive infinity. But if β moves toward negative infinity, the exp (β) converges to 0. This implies that exp (β), i.e., the value of the odds ratio ranges between 0 to positive infinity.

Reference

Agresti, A. 1990. *Categorical Data Analysis*, John Wiley & Sons, New York

http://www.toodoc.com/introduction-to-categorical-data-analysis-agresti-a-ebook.html

Aldrich John H. and Nelson Forest 1984. *Linear Probability, Logit and Probit Models*, Sage Publications, pp-26, 1984

Cabrera, A. F. 1994. Logistic regression analysis in higher education: An applied perspective, *Higher Education: Handbook of Theory and* Research, Vol. 10, 225-256.

Cohen Jacob 2003. *Applied Multiple Regression-Correlation Analysis for the Behavioral Sciences,* 3rd Edition, Lawrence Erlbaum Associates, Inc., New Jersey, 07430

Cox, D.R. and E. J. Snell 1989. *Analysis of binary data*, (2nd edition). London: Chapman & Hall.

Duraisamy, Malathy 2001. Demand for and Access to Schooling in Tamil Nadu, in Vaidyanathan A. and P R Gopinathan Nair (eds) *Elementary Education in Rural India: A Grassroots View,* Sage Publication, New Delhi.

Garson, G. David 2009. *Logistic Regression"* Retrieved 8/7/09 from

http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm#estimates

Greene, W.H. 2003. *Econometric Analysis*, 5th edition, Prentice Hall.

Gujarati, N. Damodar 2003. *Basic Econometrics' fourth edition (international),* The McGraw-Hill Companies, Inc

Hailpern S. M. and P. F. Visintainer 2003. Odds ratios and logistic regression: further examples of their use and interpretation, *The Stata Journal* (2003) 3, Number 3, pp. 213–225.

Harrell, F. 2001. *Regression Modeling Strategies,* New York: Springer.

Hosmer D.W. and Lemeshow S. 1989: *Applied Logistic Regression*, New York, John Wiley & Sons. 307p.

Hosmer, D. W., & Lemeshow, S. 2000. *Applied logistic regression* (2nd Ed.). New York: John Wiley & Sons, Inc.

http://books.google.co.in/books?id=Po0RLQ7USIMC&pg=PR3&dq=hosmer+and+lemeshow+logistic+regression&source=gbs_selected_pages&cad=4#v=onepage&q=&f=false

John W. Frass, J. Michael Drushal and Jeff Graham 2002. *Expressing Logistic Regression Coefficients as Changes in Initial Probability Values: Useful Information for Practitioners,* Office of Educational Research and Improvement, Educational Resources Information Centre (ERIC), US Department of Education.
http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED469762

Karen Grace-Martin 2001. *Stat News #44: Introduction to Logistic Regression*, Office of Statistical Consulting, Cornell University, Division of Nutritional Sciences and the College of Human Ecology, and faculty in the College of Agriculture and Life Science;
http://www.human.cornell.edu/Admin/StatCons/.
http://www.human.cornell.edu/admin/statcons/resource/topics/logistic.htm

Krznar, I. 2004. *Currency Crisis: Theory and Practice with Application to Croatia*, Croatian National Bank Working Paper, Vol.12, August, pp.1-46.

Larsen Pia Veldt 2008. *Master of Applied Statistics, ST111: Regression and analysis of variance,* http://statmaster.sdu.dk/courses/st111

Long Scott J. 1997. *Regression Models for Categorical and Limited Dependent Variables (Advanced Quantitative Techniques in the Social Sciences*, Sage Publications ISBN 0-8039-7374-8

McCullagh, P. and J.A. Nelder 1989. *Generalized Linear Models, 2nd Edition*, Chapman & Hall.

Majumdar Manabi 2001. Educational Opportunities in Rajasthan and Tamil Nadu: Despair and Hope, in Vaidyanathan A. and P R Gopinathan Nair (eds) *Elementary Education in Rural India: A Grassroots View*, Sage Publication. New Delhi.

Menard, Scott 2002. *Applied Logistic Regression Analysis*, 2nd Edition, (Quantitative Applications in the Social Sciences)', Sage University Paper, Series No. 07-106. Thousand Oaks (CA): Sage. http://books.google.co.in/books?id=EAI1QmUUsbUC&pg=PP9&lpg=PP9&dq=Menard+S.+(1995):+%E2%80%98Applied+Logistic+Regression+Analysis&source=bl&ots=4RGNM2rXGW&sig=8uqmC-h-

Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination' *Biometrika, 78*, 691–692.

Peng, Chao-Ying Joanne Lee, Kuk Lida Ingersoll, Gary M. 2002. An Introduction to Logistic Regression - Analysis and Reporting, *Journal of Educational Research*, September/October 2002 [Vol. 96(No. 1)] pp. 3-13.

Press, S. J., & Wilson, S. 1978. Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association, 73*, 699–705

Ryan, T.P. 1997. *Modern Regression Methods*, Wiley, New York

Sengupta, Piyali and Jaba Guha 2002. Enrolment, Dropout and Grade Completion of Girl Child in West Bengal, *Economic and Political Weekly*, Vol. XXVII No. 17, 1621-37.

APPENDIX

Appendix-7.I Result of Univariable Analysis

Model-I, Dependent Variable- PBDOSC

PBDOSC		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ECONDEP	0.338	0.202	2.800	1	0.094	1.401
Step 1(a)	EDNTOTEX	-0.301	0.086	12.341	1	0.000	0.740
Step 1(a)	EMPMTH	-0.513	0.422	1.478	1	0.224	0.599
Step 1(a)	FTHEDN	-0.332	0.101	10.715	1	0.001	0.717
Step 1(a)	FWP	-0.620	0.431	2.070	1	0.150	0.538
Step 1(a)	INREGIRREG	-1.943	0.533	13.274	1	0.000	0.143
Step 1(a)	MTHEDN	-0.656	0.205	10.227	1	0.001	0.519
Step 1(a)	OPTNTCOST	1.514	0.441	11.775	1	0.001	4.543
Step 1(a)	MPCE10	-0.076	0.026	8.690	1	0.003	0.927
Step 1(a)	HDEP	0.406	0.322	1.591`	1	0.207	1.501

Model-II, Dependent Variable- PBENRLMNT

PBENRLMNT		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ECONDEP	0.948	0.255	13.843	1	0.000	2.580
Step 1(a)	EDNTOTEX	-0.183	0.076	5.732	1	0.017	0.833
Step 1(a)	EMPMTH*	-0.440	0.467	0.888	1	0.346	0.644
Step 1(a)	FTHEDN	-0.637	0.210	9.176	1	0.002	0.529
Step 1(a)	FWP*	-0.159	0.467	0.116	1	0.733	0.853
Step 1(a)	INREGIRREG	-1.384	0.545	6.453	1	0.011	0.251
Step 1(a)	MTHEDN	-0.456	0.179	6.468	1	0.011	0.634
Step 1(a)	OPTNTCOST	0.669	0.471	2.017	1	0.156	1.952
Step 1(a)	MPCE10	-0.163	0.046	12.639	1	0.000	0.849
Step 1(a)	HDEP	1.540	0.410	14.135	1	0.000	4.665

*Not included in the model; Sig. > 0.25

Appendix-7.II Frequency Table- Logistic Regression

Statistics

		pbdosc	pbenrlmntt
N	Valid	120	120
	Missing	0	0

pbdosc

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	89	74.2	74.2	74.2
	1.00	31	25.8	25.8	100.0
	Total	120	100.0	100.0	

pbenrlmntt

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	97	80.8	80.8	80.8
	1.00	23	19.2	19.2	100.0
	Total	120	100.0	100.0	

Appendix-7.III Contingency Table for Hosmer and Lemeshow Test

Model-I, Dependent Variable- PBDOSC

		pbdosc = .00		pbdosc = 1.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	12	11.998	0	.002	12
	2	12	11.968	0	.032	12
	3	11	11.842	1	.158	12
	4	12	11.573	0	.427	12
	5	12	10.871	0	1.129	12
	6	10	10.012	2	1.988	12
	7	8	8.684	4	3.316	12
	8	5	6.109	7	5.891	12
	9	5	3.782	7	8.218	12
	10	2	2.162	10	9.838	12

Model-II, Dependent Variable- PBENRLMNT

		pbenrlmntt = .00		pbenrlmntt = 1.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	12	12.000	0	.000	12
	2	12	11.999	0	.001	12
	3	12	11.984	0	.016	12
	4	12	11.922	0	.078	12
	5	12	11.717	0	.283	12
	6	11	11.223	1	.777	12
	7	11	10.550	1	1.450	12
	8	7	8.374	5	3.626	12
	9	7	5.648	5	6.352	12
	10	1	1.582	11	10.418	12