Journal of Mathematics and Statistics 5 (4): 387-394, 2009
ISSN 1549-3644
© 2009 Science Publications

Optimization Techniques for Variable Selection in Binary Logistic Regression Model Applied to Desire for Children Data

S.K. Sarkar and Habshah Midi
Laboratory of Applied and Computational Statistics, Institute for Mathematical Research,
University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Abstract: Problem statement: The population problem is the biggest problem in the world. In the global and regional context, Bangladesh population has drawn considerable attention of the social scientists, policy makers and international organizations. Bangladesh is now world's 10th populous country having about 140 million people. The recent experience of Bangladesh shows that fertility can sustain impressive declines even when women's lives remain severely constrained. Recent statistics also suggest that, despite a continuing increase in contraceptive prevalence rate (56%), the expected fertility decline in Bangladesh has stalled. **Approach:** The purpose of this study was to explore the possibility of further fertility decline in Bangladesh with special attention to identify some social and demographic factors as predictors which are responsible to desire for more children using stepwise and best subsets logistic regression approaches. The study had compared two approaches to determine an optimum model for prediction of the outcome. **Results:** It had been found, excess desire for children is solely responsible for the stalled fertility. **Conclusion:** To overcome the situation, the policy makers of Bangladesh should pay their attention to eliminate the regional variations of desire for more children and introduce awareness programs among rural women about the positive impact of smaller family.

Key words: Best subsets, stepwise logistic regression, design variables, Mallow's C_p, score test

INTRODUCTION

The population problem is the biggest problem in the world today. It makes every other problem worse and harder to solve. The world's population is expected to grow by another 2.3 billion, from 6.8 billion in 2009 to 9.1 billion in 2050. Most of this growth will take place in the developing countries. In global and regional context, Bangladesh population has drawn considerable attention of the social scientists, policy makers and international organizations. Bangladesh is now world's 10th populous country having about 140 million people. According to the United Nations and other agencies, the population growth rate of Bangladesh is still 1.65%. If this rate continues, the population of Bangladesh will double in 2050. Unless action is taken to accelerate the reductions in the rates of growth, the population of the world will not stabilize and certain region and countries like Bangladesh will go far beyond the limits consistent with political stability and acceptable social and economic conditions. However, recent statistics suggest that, despite a continuing increase in contraceptive prevalence rate (55.8%), the fertility decline in Bangladesh has stalled. The total fertility rate is still 3.1 and it is far beyond the replacement level fertility rate 2.1. Further fertility decline is required to achieve stable population in Bangladesh[14].

The purpose of this study is to explore the possibility of further fertility decline in Bangladesh with special attention to identify some crucial social and demographic factors as predictors which are responsible to desire for more children. The study provides a simple explanation and demonstration of how to obtain a best subsets solution in logistic regression and interpret the results.

The criteria for including a variable in a model may vary from one problem to the next and from one scientific discipline to another. The traditional approach to statistical model building involves seeking the most parsimonious model that still explains the data. There are several steps one can follow to aid in the selection of variables for a logistic regression. The present study will discuss stepwise and best subset logistic regression for variable selection and compare them to determine a parsimonious model. Variables must be selected carefully so that the model makes accurate predictions, but without over fitting the data. Selecting variables by

Corresponding Author: S.K. Sarkar, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

hand is a laborious task and can over look important predictors. Thus, it is important that variable selection be automatic. The problem of variable selection is often addressed by sequential methods that start with a set of variables and attempt to grow or shrink the set by selecting which predictor should be added or removed from the set. This approach has been traditionally called stepwise selection. The method is frequently employed in sociological, demographical, educational and psychological research, both to select useful subsets of variables and to evaluate the order of importance of variables[3,10].

A crucial aspect of using stepwise logistic regression is the choice of an 'alpha' level to judge the importance of variables. Let p_e denote the level of entry for the important variables. The choice of p_e determines how many variables eventually are included in the model. Bendel and Afifi[1] have studied the choice of p_e for stepwise linear regression and Costanza and Afifi[2] have studied the choice for stepwise discriminant analysis. More recently by using Monte Carlo simulation Lee and Koval[8] examined the issue of significance level for forward stepwise logistic regression. The results of this research have shown that the choice of $p_e = 0.05$ is too stringent, often excluding important variables from the model. Choosing a value for p_e in the range from 0.15-0.20 is highly recommended by Hosmer and Lemeshow[4]. On the other hand the program requires the second pre-chosen level p_r to remove the variable from the model, which indicates some minimal level of continued contribution to the model. Whatever value one may chose for p_r, it must exceed the value of p_e to guard against the possibility of having the program enter and remove the same variable at successive steps. Since there is no theory Hosmer and Lemeshow[4] strongly recommended to use $p_e = 0.15$ and $p_r = 0.20$ for stepwise logistic regression program to locate the important variables which are related to the outcome.

An alternative to stepwise selection of variables for a model is best subsets selection. The procedures identify a group of subset models that give the best values of a specified criterion without requiring the fitting of all possible subset logistic regression. A best subsets approach would allow for the identification of these competing models. A statistical algorithm is adopted which produces some type of summary statistic for every possible combination of predictors. Hosmer *et al.*[6] proposed such an algorithm for estimating a best-subsets logistic regression. The method was reiterated in Hosmer and Lemeshow's[4] popular book on applied logistic regression.

The objective of this study is to pare down a large number of predictor variables to a subset which meets theoretical or predictive standards on the basis of stepwise logistic regression and best subsets approaches. After selection of the important predictor variables, the study will compare the two approaches to identify an optimum model for prediction of the outcome.

MATERIALS AND METHODS

The Bangladesh Demographic and Health Survey (BDHS) is part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning, maternal and child health. The BDHS is intended to serve as a source of population and health data for policymakers and the research community. The study will utilize the data from BDHS 2004. Macro International Inc. of Calverton, Maryland, USA, provided technical assistance to the project as part of its International Demographical and Health Surveys program and financial assistance was provided by The United States Agency for International Development (USAID). A total of 10,523 households were selected for the sample and 11,440 eligible women were completed their interview. The women under sterilization, declared in fecund, divorced, widowed, having more than and less than two living children are not involved in the analysis. Only 2216 eligible women who have two living children and able to bear and desire more children are only considered here during the period of global two children campaign.

The variable age of the respondent, region of residence, fertility preference, place of residence, level of education, expected number of children and sex preference are considered in the analysis. The variable fertility preference involving responses corresponding to the question, would you like to have (a/another) child? The responses are coded 0 for 'no more' and 1 for 'have another'. This variable is treated response variable Y as desire for children in the analysis. The age of the respondent (X_1), region of residence (X_2) is coded 1 for 'Barisal, 2 for 'Chittagong', 3 for 'Dhaka' 4 for 'Khulna' 5 for 'Rajshahi' and 6 for 'Sylhet', place of residence (X_3) is coded 0 for 'urban' and 1 for 'rural', level of education (X_4) is coded 0 for 'no education', 1 for 'primary level', 2 for 'secondary level' and 3 for 'higher level', sex preference (X_5) is coded 0 for 'no preferences' and 1 for 'preferences' and expected number of children (X_6) is coded 0 for 'two or less' and 1 for 'more than two'. In the study, the number of categories for each predictor varies from two to six. For instance, the region of residence (X_2) and level of education (X_4) have more than two categories. Both predictors have more than two discrete values and the scale of measurement is nominal. We know that it is inappropriate to model a nominal scale variable in logistic regression as if it were an interval scale variable. Therefore, we must form a set of design variables to represent the categories of the predictors. A reference cell coding technique is used to generate design variables. For X_2 we have five design variables as X_2_c, X_2_d, X_2_k, X_2_r and X_2_s. For X_4, the three design variables are X_4_n, X_4_p and X_4_s.

There are two methods that may be used to select variables from a summary table. The first method is based on p-value for entry at each step while the second is based on a likelihood ratio test of the model at the current step versus the model at the final step. Let q denote an arbitrary step in the procedure. In the first method we compare $p_{x_q}^{(q-1)}$ to a pre-chosen significance level such as $p_e = 0.15$. The subscript of p refers to the variable that has been added to the model and the superscript (q-1) refers to the step. If the value $p_{x_q}^{(q-1)}$ is less than p_e, then we move to the step q. We stop at the step when $p_{x_q}^{(q-1)}$ exceed p_e. We consider the model at the previous step for further analysis. In this method the criterion for entry is based on a test of the significance of the coefficient for X_q conditional on $X_1, X_2, \ldots X_{q-1}$ being in the model. In general, we may test the conditional null hypothesis $H_0{:}\beta_q = 0$ such that $X_1, X_2, \ldots X_{q-1}$ in the model, against $H_1{:}\beta_q \neq 0$. To test the null hypothesis the test statistic

$$G = \left\{\left(-2\ln L_{q-1}\right) - \left(-2\ln L_q\right)\right\} \tag{1}$$

where, L_q and L_{q-1} are the log-likelihoods for the step q and q-1 respectively. G is approximately distributed as chi-square with 1 degree of freedom depending on whether X_q is continuous or dichotomous and k-1 degrees of freedom whether X_q is polychotomous with k categories. The software SPSS version 11.5 uses the score test for selection and the likelihood ratio test for removal of covariates. Table 1 presents the p-values as a result of applying stepwise variable selection. At each step, the p-values from the score test to enter are below the horizontal line and the p-values for the likelihood ratio test to remove are above the horizontal line. The asterisk denotes the maximum p-value to remove at each step.

In the second method, we compare the model at the current step q to the model at the final step. We evaluate the p-value for the likelihood ratio test of these two models and proceed in this fashion until this p-value exceeds p_e. This tests that the coefficients for the variables added to the model from step q to the final step s are all equal to zero. That is we would like to test the null hypothesis $H_0{:}\beta_q = \beta_{q+1} = \ldots = \beta_s = 0$, s≤p against H_1: At least two of the coefficients are not zero. To test the null hypothesis, the test statistic is:

$$G = \left\{\left(-2\ln L_q\right) - \left(-2\ln L_s\right)\right\} \tag{2}$$

where, Lq and Ls are the log-likelihood for the step q and the final step s, respectively. The statistic G is approximately distributed as chi-square with degrees of freedom depending on the number of parameters to be tested from step q+1 to the final step s. At any given step it has more degrees-of-freedom than the test employed in the first method. For this reason the second method may possibly select a larger number of variables than the first method. The summary statistics for the above two methods of variable selection are illustrated in Table 2. At each step, each method provides a test of a different hypothesis. The number of parameters being tested for the second method is larger than the first method except for the last step. The second method may possibly select more variables than the first method. In cases where this occurs, one should carefully examine the additional variables and include them if they seem socially relevant to the outcome. In such case we would undoubtedly opt for the richer model selected by second method. In the present study, both methods select the same set of variables for further analysis.

Table 1: The p-values to enter (below the horizontal line) and p-values to remove (above the horizontal line) the covariates

Variable\step	0	1	2	3	4	5
Expected number of children (X_6)	0.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Age of respondent (X_1)	0.000	0.000	**0.000**	**0.000**	**0.000**	**0.000**
Region of residence (X_2)	0.000	0.000	0.000	**0.000**	**0.000**	**0.000**
Place of residence (X_3)	0.000	0.002	0.063	0.017	**0.017***	**0.015**
Sex preference (X_5)	0.000	0.087	0.102	0.169	0.150	**0.145***
Level of education (X_4)	0.000	0.076	0.433	0.204	0.366	0.380

Note: *: Denotes the maximum p-value to remove the explanatory variable at each step

Table 2: The log-likelihood and likelihood ratio test statistics (G), degrees of freedom (df) and p-values for two methods of selecting variables for a final model

			Method 1			Method 2		
Step	Variable entered	-2 log-likelihood	G	df	p-value	G	df	p-value
0		2761.26				882.85	9	0.0000
1	Expected number of children (X_6)	2051.68	709.58	1	0.000	173.27	8	0.0000
2	Age of respondent (X_1)	1953.23	98.45	1	0.000	74.82	7	0.0000
3	Region of residence (X_2)	1886.26	66.97	5	0.000	7.85	2	0.0197
4	Place of residence (X_3)	1880.53	5.73	1	0.017	2.12	1	0.1450
5	Sex preference (X_5)	1878.41	2.12	1	0.145			

The conclusion of the stepwise selection process has only identified a collection of variables which seem to be statistically important because the procedure identifies variables as candidates for a model solely on statistical grounds. We can observe from Table 2 that the stepwise procedure for variable selection terminates at step 5, because no further predictors can be added with the resulting p-values less than 0.15. Thus the variables X_6, X_1, X_2, X_3 and X_5 have been selected by stepwise logistic regression procedure for further analysis.

On the other hand Hosmer *et al.*[6] have shown that best subsets logistic regression may be performed in a straight forward manner using any program capable of best subsets linear regression. Applying best subsets linear regression software to perform best subsets logistic regression is most easily explained using vector and matrix notations. Let X denotes the n×(p+1) matrix containing the values of all p covariates with the first column containing 1 to represent the constant term. The n×n diagonal matrix is denoted as V with general element $v_i = \hat{\pi}(1-\hat{\pi}_i)$ where $\hat{\pi}_i$ be the estimated logistic probability computed using maximum likelihood estimate of is $\hat{\beta}$. Symbolically:

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

and

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Pregibon[12] verified that $\hat{\beta} = (X'VX)^{-1}X'VZ$ where $Z = X\hat{\beta} + V^{-1}r$ and r is the vector of residuals given by $r = (Y + \hat{\pi})$. This representation of $\hat{\beta}$ provides the basis for use of linear regression software. It is easy to verify that any linear regression package, that allows weights, produces coefficient estimates identical to $\hat{\beta}$ when used with z_i as the dependent variable and case weights v_i, equal to the diagonal elements of V. To replicate the results of the maximum likelihood fit from a logistic regression package using a linear regression package, we calculate for each case, the value of a dependent variable and the corresponding case weight are as follows:

$$z_i = \text{In}\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)} \tag{3}$$

$$v_i = \hat{\pi}_i(1-\hat{\pi}_i) \tag{4}$$

It can be shown that the weighted residual sum squares produced by the program is:

$$\sum_{i=1}^{n} v_i (z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\pi})^2 / \hat{\pi}_i(1-\hat{\pi}_i) \tag{5}$$

The expression in (5) is a Pearson chi-square statistic from a maximum likelihood logistic regression program. The subsets of variables selected for best models depend on the criterion chosen for best. Hosmer and Lemeshow[4] recommended using the best subsets linear regression that was developed by Mallow[11]. This is a measure of predictive squared error denoted by C_q. This measure is originally denoted by C_p. We use q instead of p because p refers to a total number of possible variables while q refers to some subsets of variables. Hosmer *et al.*[6] justified that when best subsets logistic regression is performed via a best subsets linear regression package Mallow's C_q has the same intuitive appeal as it does in linear regression. They showed that for a subset of q of the p variables:

$$C_q = \frac{\chi^2 + \lambda^*}{\chi^2 / (n-p-1)} + 2(q+1) - n \tag{6}$$

Where:

χ^2 = The Pearson chi-square obtained from (5)

λ^* = The multivariate Wald test statistic for the hypothesis that the coefficients for the p-q variables not in the model are equal to zero

Under the assumption that the model fit is the correct one, the approximate expected value of χ^2 and λ^* are (n-p-1) and (p-q) respectively. Substitution of these approximate expected values into the expression for C_q yields C_q = q+1. Hence models with C_q near q+1 are candidates for a best model. The best subsets linear regression program selects as best that subset with the smallest value of C_q. If the C_q criterion is to be employed, the five best subsets according to this criterion are to be identified. This algorithm search for the five subsets of predictors variables with the smallest C_q values using much less computational efforts than when all possible subsets are evaluated. Table 3 represent the results of the five best models selected using C_q as the criterion obtained from the output of Statistical Analysis System (SAS) 9.1 for windows. Using only the summary statistics, we would select model 3 as the best model since it has the smallest value of C_q.

Hosmer and Lemeshow[5] also show how an approximation to Mallow's C_q can be obtained from score test output in a survival time analysis. A similar approximation can be obtained from C_q for logistic regression. First, Pearson chi-square statistic is replaced by its mean $\chi^2 \approx$ (n-p-1). Next the Wald statistic for the p-q excluded covariates may be approximated by the difference between the values of the score test for all p covariates and the score test for q covariates, namely $\lambda_q^* \approx s_p - s_q$. These results produce an approximation to (6) as follows:

$$C_q \approx s_p - s_q + 2q - p + 1 \tag{7}$$

The value of s_p is the score test for the model containing all p covariates and is obtained from the computer output. The value of s_q is the score test for the particular subset of q covariates and its value is also obtained from the output. Table 4 illustrated the five best models that are identified based on score test and Mallows C_q criterion.

The results of Table 4 selected the five best models using C_q, as the main criterion and we would select model 3 as the best model since it has the smallest value of C_q. We observed that the best subsets selected by both approximation and stepwise logistic regression procedures identified the same set of predictors.

Table 3: Five best models identified using Mallow's C_q.

Model	Model covariates	Mallows C_q
1	X_1, X_5, X_6	67.32
2	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3	10.04
3	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3, X_5, X_6	10.01
4	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3, X_4_n, X_4_p, X_4_s, X_6	12.94
5	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_4_n, X_4_p, X_4_s, X_5, X_6	15.39

Table 4: Five best models identified using the score test approximation to Mallow's C_q (s_{12} = 850.11)

Model	Model covariates	Score s_q	C_q
1	X_1, X_5, X_6	796.20	48.91
2	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3	331.11	522.00
3	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3, X_5, X_6	848.32	8.79
4	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_3, X_4_n, X_4_p, X_4_s, X_6	848.79	12.32
5	X_1, X_2_c, X_2_d, X_2_k, X_2_r, X_2_s, X_4_n, X_4_p, X_4_s, X_5, X_6	847.35	13.76

Intensive analysis: Consider a collection of q predictors selected with stepwise logistic regression approach be denoted by the vector $X' = (X_1, X_2 \ldots X_q)$. Suppose the conditional probability that the outcome is present be denoted by $P(y = 1 | X) = \pi$. Then the log-odds of having Y = 1 is modeled as a linear function of the predictor variables as:

$$\text{In}\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q ; 0 \le \pi_i \le 1 \tag{8}$$

where, the function $\pi_i = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_q X_q)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_q X_q)}$ is known as logistic function. The most commonly used method of estimating the parameters of a logistic regression model is the Maximum Likelihood (ML). Suppose (y_1, y_2….y_n) be an independent random sample of size n from the corresponding independent random variables (Y_1, Y_2…Y_n). The response Y_i is a Bernoulli random variable with probability mass function $f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$; Y_i = 0 or 1; i = 1, 2…n. Since the Y_i's are assumed to be independent the sample likelihood function is defined as the joint probability function of the random variables as $g(Y_i, Y_2, \ldots Y_n) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$, the log-likelihood function as L(β_0, β_1,… β_q) = L_i (say):

$$\begin{aligned} &= \sum_{i=1}^{n} Y_i\,(\beta_0 + \beta_1 X_1 + \ldots. + \beta_q X_q) \\ &- \sum_{i=1}^{n} \text{In}\left\{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots. + \beta_q X_q)\right\} \end{aligned} \tag{9}$$

In multivariable logistic regression, the likelihood equations are non-linear explicit function of unknown parameters. Therefore, we use a very effective and well known Newton-Raphson iterative method to solve the equations which is known as iteratively reweighted least square algorithm. The solution of the likelihood equations requires special software that is available in most statistical packages. In the study, SPSS 11.5 for windows is used.

RESULTS

Once the particular multiple logistic regression model has been fitted, we should begin the process of model assessment. This usually involves formulation and testing of a statistical hypothesis to determine whether the predictors in the model are significantly associated to the response variable. Fortunately, the likelihood ratio test for the overall significance of all coefficients for the predictors as well as significance of single predictor in the model is shown in Table 2. Considering the results from Table 2 we may conclude that except the predictor sex preference (X_5) all other four predictors are significantly associated with the response variable at 5% level of significance.

In order to find the overall goodness-of-fit, Hosmer and Lemeshow[7] and Lemeshow and Hosmer[9] proposed grouping based on the values of the estimated probabilities. Using this grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic, $\hat{C}$ is obtained by calculating the Pearson chi-square statistic from the g×2 table of observed and estimated expected frequencies. A formula defining the calculation of $\hat{C}$ is as follows:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k^{'}\bar{\pi}_k)^2}{n_k^{'}\bar{\pi}_k(1-\bar{\pi}_k)} \qquad (10)$$

Where:

g = Denotes the number of groups
$n_k^{'}$ = The number of observations in the kth group
o_k = The sum of the Y values for the kth group
$\bar{\pi}_k$ = The average of the ordered $\hat{\pi}$ for the kth group

Hosmer and Lemeshow[7] demonstrated that under the assumption the fitted logistic regression model is the correct model and the distribution of the statistic $\hat{C}$ is well approximated by the chi-square distribution with g-2 degrees of freedom. The value of the Hosmer-Lemeshow goodness-of-fit statistic computed from the frequencies in Table 5 is $\hat{C}$ = 6.61 and the corresponding p-value computed from the chi-square distribution with 8 degrees of freedom is 0.58. The large p-value signifies that there is no difference between the observed and the predicted values of the outcome. This indicates that the model seems to fit quite well. A comparison of the observed and expected frequencies in each of the 20 cells in Table 5 shows close agreement within each decile.

Table 5: Contingency table for Hosmer and Lemeshow goodness-of fit test

		Desire for no more children		Desire for more children					
Deciles	$\bar{\pi}_k$	Observed	Expected	Observed (o_k)	Expected	Total ($n_k^{'}$)	$\hat{C} \sim \chi^2$	df	p-value
1	0.0459	208	207.04	9	9.96	217	6.61	8	0.58
2	0.0784	206	208.29	20	17.71	226			
3	0.1051	194	199.57	29	23.44	223			
4	0.1309	186	191.20	34	28.80	220			
5	0.1588	184	180.85	31	34.15	215			
6	0.1977	183	179.72	41	44.28	224			
7	0.2702	170	162.75	53	60.25	223			
8	0.4902	118	113.17	104	108.83	222			
9	0.7644	45	52.30	177	169.71	222			
10	0.8968	24	23.11	200	200.89	224			

Table 6: Analysis of maximum likelihood estimates

Predictors	Coefficients β	S.E	Wald chi-square statistics	df	p-value	Odds ratio exp (β)
X_6	2.915	0.188	241.252	1	0.000	18.443
X_1	-0.087	0.010	68.528	1	0.000	0.9170
X_2	-	-	68.539	5	0.000	-
X_2_c	0.695	0.217	10.221	1	0.010	2.0000
X_2_d	-0.191	0.205	0.8630	1	0.353	0.8260
X_2_k	-0.434	0.226	3.7010	1	0.050	0.6480
X_2_r	-0.502	0.210	5.7050	1	0.017	0.6050
X_2_s	0.804	0.257	9.8050	1	0.002	2.2350
X_3	0.306	0.127	5.8650	1	0.015	1.2350
X_5	-0.297	0.206	2.0690	1	0.150	0.7430
Constant	0.574	0.351	2.6820	1	0.101	1.7750

Table 6 shows the coefficients β's, their standard errors, the Wald chi-square statistics, associated p-values, odds ratio i.e., exp (β). In order to determine the worth of the individual regressor in logistic regression, the Wald statistic denoted as $W = \frac{\hat{\beta}_1^2}{\left[S.E(\hat{\beta}_i)\right]^2}$ [13]. Under the null hypothesis $H_0:\beta_i = 0$, the statistic W is approximately distributed as chi-square with single degree of freedom. The Wald chi-square statistic from Table 6 almost agree reasonably well with the likelihood ratio test for individual predictors that all the coefficients except X_2_d (Dhaka region) and X_5 (sex preference) have significant contribution to predict the response variable. In order to interpret the results of Table 6 we need to introduce a measure of association named 'odds ratio'. The odds of the outcomes being present among individuals with x = 1 is defined as $\pi_1/1-\pi_1$. Similarly, the odds of the outcome being present among the individuals with x = 0 is defined as $\pi_0/1-\pi_0$. The odds ratio, denoted as OR, defined as the ratio of the odds for x = 1 to the odds for x = 0 and given by $OR = \frac{\pi_1/1-\pi_1}{\pi_0/1-\pi_0}$. In general, the relationship between the odds ratio and the logistic regression coefficient is OR = exp (β) and the relationship is widely used to interpret the fitted values.

The odds ratio corresponding to the estimated coefficient for the variable expected number of children (X_6) is 18.44 indicates the change in log-odds of desire for more children among the women having more than two desired children. The result suggests those women who expect more than two children are 18 times as likely to desire another child as other women in the study population.

The odds ratio for the variable age of respondent (X_1) is 0.917, indicates the change in log-odds of desire for more children per one year increase in age among the study population. The odds ratio suggests the desire for more children significantly goes down by 8% for one year increase in age.

The odds ratio corresponding to the estimated coefficient for design variables X_2_c (Chittagong), X_2_d (Dhaka), X_2_k (Khulna), X_2_r (Rajshahi) and X_2_s (Sylhet) are 2.000, 0.826, 0.648, 0.605 and 2.235 respectively. Except X_2_d (Dhaka) all the coefficients are statistically significant at 5% level of significance. The results indicate the regional change in log-odds of desire for more children with respect to the reference region. The results suggest the desire for more children is approximately 2 times as likely as prevail among the women in Chittagong and Sylhet region. On the other hand, women under Khulna and Rajshahi region about 40% as less likely to desire for more children as reference region.

The odds ratio corresponding to the estimated coefficient for the variable place of residence (X_3) is 1.235 indicates the change in log-odds of desire for more children among rural women with respect to urban women. The result can be interpreted as the desire for more children significantly rise about 24% among rural women as compared to urban women.

The variable sex preference (X_5) is not statistically significant. Theoretically there is no advantage to include the variable in the model. In fact, sex preference is known to be 'socially important variable' to determine desire for children. The response corresponding to the variable may be latent. Son preference is the most significant factor which continued to exert a great influence on bearing a third child even during the period of the two children campaign. Hence there is a further scope for analysis with such socially significant variable.

DISCUSSION

Model selection is a fundamental task in data analysis. Methods such as stepwise and best subsets logistic regression are tools that build and compare suits of logistic regression models. Stepwise logistic regression is mainly designed to identify the most parsimonious set of predictors that are effective in predicting the response variable. The procedure indicates covariates with statistically significant effect, simultaneously adjusting for the other covariates in the logistic regression model. So the procedure is best viewed as a data screening tool. On the other hand, the best subsets technique is versatile because it allows for the consideration of a number of issues, statistical and theoretical, in comparing candidates models. The best subsets procedure is a time saving algorithms and have been developed to identify the most promising models, without having to evaluate all possible candidate models. Though the best subsets approach is introduced as time saving, as the number of models to compare grows rapidly as increase the number of potential predictors, the algorithms may require excessive computer time. Selection of best subsets of variables based on some criteria like Mallow's C_q need fitting a lots of models. It can be very expensive because each fit requires an iterative procedure. So, users should not be lured into accepting the variables suggested by a best subset strategy without considerable critical evaluation. In contrast, stepwise logistic regression outputs only a single set of predictors, thus fostering the notion that the chosen model is the best model. Under this condition, stepwise logistic regression procedures may need to be employed to assist in the selection of predictor variables. Stepwise methods are sequential, hence cheaper than best subsets methods.

The output of the intensive analysis suggest that the desire for more children is significantly higher in Chittagong and Sylhet regions and significantly lower in Khulna and Rajshahi regions as compare to reference region. It is also higher among rural women than among urban women. It is established in this study that urban women have less desire for more children than rural women. It is also observed that desired family size is still significantly larger among the study population. In fact, effective population control cannot be achieved until there is a change in the society's attitude toward desired family size.

CONCLUSION

In this respect, the Government of Bangladesh should highlight to the rural women that limiting family size has positive effects on the mother's health, domestic peace, happiness and well-being. The policymaker should pay their attention ensuring female educational programs that can provide young women with gainful employment. This strategy also delaying age at marriage as well as age at first birth which is important for fertility decline. Finally, it is important for the Government of Bangladesh, instead of propagating the two-child norm across the board, emphasize those policies that actively enhance women's status through education as well as involving them in the workforce and change their attitudes toward family size.

REFERENCES

1. Bendel, R.B. and A.A. Afifi, 1977. Comparison of stopping rules in forward "stepwise" regression. J. Am. Stat. Assoc., 72: 46-53. DOI: 10.2307/2286904
2. Costanza, M.C. and A.A. Afifi, 1979. Comparison of stopping rules in forward stepwise discriminant analysis. J. Am. Stat. Assoc., 74: 777-785. http://www.jstor.org/stable/2286399
3. Davis, L.J. and K.P. Offord, 1997. Logistic regression. J. Person. Assess., 68: 497-507. DOI: 10.1207/s15327752jpa6803_3
4. Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., John Wiley and Sons, Inc., ISBN: 9780471356325, pp: 373.
5. Hosmer, D.W. and S. Lemeshow, 1999. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd Edn., Wiley, Inc., New York, ISBN: 9780471754992, pp: 392.
6. Hosmer, D.W., B. Jovanovic and S. Lemeshow, 1989. Best subsets logistic regression. Biometrics, 45: 1265-1270. http://www.jstor.org/stable/2531779
7. Hosmer, D.W. and S. Lemeshow, 1980. Goodness of fit tests for the multiple logistic regression model. Commun. Stat., 9: 1043-1069. DOI: 10.1080/03610928008827941
8. Lee, K. and J.J. Koval, 1997. Determination of the best significance level in forward stepwise logistic regression. Commun. Stat., 26: 559-575. DOI: 10.1080/03610919708813397
9. Lemeshow, S. and D.W. Hosmer, 1982. A review of goodness-of-fit statistics for use in the development of logistic regression models. Am. J. Epidemiol., 115: 92-106. PMID: 7055134
10. Lottes, I.L., M.A. Adler and A. DeMaris, 1996. Using and interpreting logistic regression: A guide for teachers and students. Teach. Sociol., 24: 284-298. http://www.jstor.org/stable/1318743
11. Mallows, C.L., 1973. Some comments on C_p. Technometrics, 15: 661-676. http://www.math.tau.ac.il/~yekutiel/MA%20seminar/Malows%202000.pdf
12. Pregibon, D., 1981. Logistic regression diagnostics. Ann. Stat., 9: 705-724. DOI: 10.1214/aos/1176345513
13. Rao, C.R., 1973. Linear Statistical Inference and Its Applications. 2nd Edn., Wiley, New York, ISBN: 9780471218753, pp: 625.
14. UN., 2008. Population division, world population prospect: The 2008 revision. http://esa.un.org/unpp/