Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 6 (8): 900-909, 2011
ISSN 1819-3579 / DOI: 10.3923/tasr.2011.900.909

Adequacy of Multinomial Logit Model with Nominal Responses over Binary Logit Model

[1,2]Habshah Midi, [2]S.K. Sarkar and [2]Sohel Rana

[1]Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Laboratory of Computational Statistics and Operational Research, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

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

ABSTRACT

The aim of this study was to fit a multinomial logit model and check whether any gain achieved by this complicated model over binary logit model. It is quite common in practice, the categorical response have more than two levels. Multinomial logit model is a straightforward extension of binary logit model. When response variable is nominal with more than two levels and the explanatory variables are mixed of interval and nominal scale, multinomial logit analysis is appropriate than binary logit model. The maximum likelihood method of estimation is employed to obtain the estimates and consequently Wald test and likelihood ratio test have been used. The findings suggest that parameter estimates under two logits were similar since neither Wald statistic was significant. Thus, it can be concluded that complicated multinomial logit model was no better than the simpler binary logit model. In case of response variable having more than two levels in categorical data analysis, it is strongly recommended that the adequacy of the multinomial logit model over binary logit model should be justified in its fitting process.

Key words: Nominal response, multinomial logit, likelihood ratio test, Wald test, odds ratio, deviance

INTRODUCTION

Situations associated with various field of applied sciences involving categorical responses are quite common in practice. A categorical variable is one whose numerical values serve only as levels distinguishing different categories. Regression procedures aid in understanding and testing complex relationship among variables and is forming predictive equations. Generally, logistic regression technique is one such procedures and the most modern practice over discriminant analysis which allows categorically and continuously-scaled covariates to predict any categorically-scaled response (Darlington, 1990). Logistic regression analysis extends the techniques of multiple regression analysis to the situations in which the response variable is categorical and makes no distributional assumption about the variables. It is a direct probability model having capability to provide valid estimates regardless of study design (Harrell, 2001). Thus, it has become popular and widely used modeling technique with categorical data analysis in many fields. Logistic regression analysis is one of the most frequently used statistical techniques and is especially familiar in epidemiological research but subsequently the technique has been extended

and become tremendous growth in the use within social sciences, marketing applications, demographic and educational researches since the last two decades (Pohlmann and Dennis, 2003; Peng *et al*., 2002; Chuang, 1997).

It is not uncommon in practice that the categorical response has more than two levels. Several problems frequently encountered in the field of applied sciences involve a nominal response variable having multiple categories and a combination of interval and nominal scale explanatory variables. When response variable is nominal with more than two levels and the explanatory variables are mixed of interval and nominal scale, multinomial logit analysis is appropriate. A multinomial logit model is used for the data in which the response is often a set of choices and is therefore measured on a nominal scale. In the cases, where the responses are not ordinal in nature and the levels are unordered, multiple logistic regression is one often-used strategy to investigate the relationship between nominal responses and a set of explanatory variables. This modeling technique is flexible enough to deal with a variety of common applications and computationally affordable (Chan, 2005; Long, 1987).

Multinomial logit model is similar to dichotomous or binary logit model; except that the response variable is in the case will have multiple discrete responses instead of just two. This model is special case of discrete choice or conditional logit models introduced by McFadden (1974) which generalizes binary logistic regression by allowing more than two discrete responses. It is a generalized linear model that is used to predict the probabilities of the different possible responses of a categorically distributed response variable, given a set of explanatory variables. The multinomial logit assumes that data are case specific; that is, each response variable has a single value for each case. The multinomial logit model also assumes that the response variable cannot be perfectly predicted from the explanatory variables for any case. Collinearity is assumed to be relatively low, as it becomes difficult to differentiate between the impacts of several variables if they are highly correlated.

It is clear from the fact that multinomial logit model can be seen as a straightforward extension of binary logit model. Consequently levels of multinomial logit model can be collapsed and reduced the model as a binary one. This reduced model provides a good approximation to both the estimates of the coefficients and their corresponding standard errors (Begg and Gray, 1984). The purpose of the present study was to build a multinomial logit model with a response variable having more than two levels and test whether significant differences in the separate odds ratios produced by different logits over all model covariates and checks the adequacy or gain achieved with multinomial logit model over binary logit model.

MATERIALS AND METHODS

The Bangladesh Demographic and Health Survey (BDHS) 2007 was the fifth survey conducted under the authority of the National Institute for Population Research and Training (NIPORT) of the Ministry of Health and Family Welfare, Bangladesh since 24 March to 11 August, 2007. It was a nationally representative multistage cluster sample survey designed to collect data and provide information on basic national indicators of social progresses. The BDHS was implemented through a collaborative effort among which Macro International provided financial and technical assistance for the survey through United States Agency for International Development (USAID). BDHS is a periodic survey conducted in Bangladesh as a part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning and maternal and child health and serves as a source of population and health data for policymakers, program managers and the research community. The data were published for the research community on

March, 2009. A total of 10819 households were selected for the sample, of which 10400 were successfully interviewed. In those households, 11178 women were identified as eligible under reproductive age for the individual interview and interviews were completed for 10996 of them. But in the current analysis, only 2357 eligible women having two living children and able to bear and desire more children are considered on the ground of two children family norm campaign globally.

The selected eligible women were asked how long they would like to wait from now before the birth of a/another child. A response corresponding to wants within two years, wants after two years and wants no more are denoted as level 1, 2 and 3, respectively and used as a discrete choice outcome variable Y in the current study. Respondent's place of residence (X_1), wealth index (X_2), professional status (X_3) and presence of sex preference (X_4) are considered as potential covariates to develop a multinomial logit model with response variable Y. The explanatory variables, place of residence X_1 is leveled as 1 for 'urban' and 2 for 'rural', wealth index X_2 is leveled as 1 for 'poorest', 2 for 'poorer', 3 for 'middle', 4 for 'richer', 5 for 'richest', professional status X_3 is leveled as 1 for 'professional' and 2 for 'house wife' and presence of sex preference X_4 is leveled as 1 for 'having preferences about child's sex' and 2 for 'no preferences about child's sex' in the study.

MULTINOMIAL LOGIT MODEL

In statistics, multinomial logistic regression sometimes called the multinomial logit model is used for prediction of the probability of occurrence of an event by fitting data to a series of logit functions applying logistic distribution. To fit a multinomial logit model having more than two levels of response variable, one must pay attention to the measurement scale (Hosmer and Lemeshow, 2000). Levels associated with the response variable in the current study are nominal scale. Let Y be a multi-categorical response variable having L nominal levels. Generally, one value typically the first, the last, or the value with the highest frequency of the response variable is designated as the baseline or reference category. The probability of membership in other categories is compared to the probability of membership in the reference category. For a response variable with L categories, multinomial logit model describes:

$$^{L}C_{2} = \frac{L(L-1)}{2}$$

possible pairs of log-odds for comparisons but it is not necessary to develop all logistic regression models instead only some choice of (L-1) pairs are necessary and the rests are redundant.

To formulate the generalized multinomial logit model, let there are k explanatory variables and an intercept term denoted by the $X_i' = (x_{0i}, x_{1i}, x_{2i}, \cdots, x_{ki})$ vector of length (k+1) where $x_{0i} = 1$ in the analysis involving n independent subjects. The general expression for conditional probability of the lth level of response variable be present given the explanatory variables, is expressed by:

$$\eta_{il} = P(Y = l|X) = \frac{e^{X_i'\gamma_l}}{\sum_{t=1}^{L} e^{X_i'\gamma_t}};\ l = 1,2,\cdots,L\ ;\ i = 1,2,\cdots,n \qquad (1)$$

where, $\gamma'_1 = (\gamma_{l0}, \gamma_{l1}, \gamma_{l2}, \ldots, \gamma_{lk})$ is a vector of unknown parameters. Without loss of generality, L coded variables $Y_{i1}, Y_{i2}, \ldots, Y_{iL}$ with corresponding probabilities $\eta_{i1}, \eta_{i2}, \ldots, \eta_{iL}$ such that: $\sum_{l=1}^{L} Y_{il} = 1$ and: $\sum_{l=1}^{L} \eta_{il} = 1$ for the ith subject can be generated from the response variables Y having L nominal

levels. In order to construct the logit function, one level should be chosen as the baseline or referent level and all other levels can be compared to it. The choice of referent level though is arbitrary, generally the last level having highest frequency of the response variable is chosen as referent level (Kutner *et al*., 2005). Multinomial logit model in general not a linear model in the parameters, logit transformation can be used to make it approximately linear by the principle of generalized linear model (McCullagh and Nelder, 1989). Using the last level L as referent level, only (L-1) meaningful comparison can be done with respect to the referent level to describe the relationship between the response variable and the explanatory variables. Thus the logit for the lth such comparison with respect to referent level is given by:

$$\psi_l = \log_e\left[\frac{\eta_{il}}{\eta_{iL}}\right] = \gamma_{l0} + \gamma_{l1}x_{1i} + \gamma_{l2}x_{2i} + \cdots + \gamma_{lk}x_{ki} = X_i'\gamma_l; l = 1,2,\cdots,(L-1) \tag{2}$$

In terms of the logit-function and using the condition $\Psi_L = 0$ the general expression for the conditional probability given in Eq. 1 can be written as:

$$\eta_{il} = P\left[Y = l|X\right] = \frac{e^{\psi_l}}{1+\sum_{t=1}^{L-1} e^{\psi_t}}; l = 1,2,\cdots,(L-1) \tag{3}$$

Since the multinomial logit model with nominal responses is a straightforward generalization of binary logit model, it can be easily collapsed into a binary logit model considering pooling multiple outcome categories into a binary 'ever' versus 'never' outcome, in case of no gain achieved by the multinomial logit model.

MAXIMUM LIKELIHOOD ESTIMATION

After formulation of the multinomial logit model, the next step is to describe the methods for obtaining estimates of the (L-1) vectors of parameters $\gamma_1, \gamma_2, \ldots, \gamma_{(L-1)}$ given by:

$$\gamma_1 = \begin{bmatrix} \gamma_{10} \\ \gamma_{11} \\ \gamma_{12} \\ \vdots \\ \gamma_{1k} \end{bmatrix}, \gamma_2 = \begin{bmatrix} \gamma_{20} \\ \gamma_{21} \\ \gamma_{22} \\ \vdots \\ \gamma_{2k} \end{bmatrix}, \ldots\ldots, \gamma_{(L-1)} = \begin{bmatrix} \gamma_{(L-1)0} \\ \gamma_{(L-1)1} \\ \gamma_{(L-1)2} \\ \vdots \\ \gamma_{(L-1)k} \end{bmatrix}$$

Multinomial logit model quantifies the effect of an explanatory variable in terms of the log-odds ratio using Maximum Likelihood Estimation (MLE). The more efficient and precise approach from the statistical viewpoint is to obtain estimates of the (L-1) logits simultaneously instead of sequential binomial logits. To do so, the likelihood for the full data set is required. In order to construct the likelihood function, let the lth category for the response variable Y is selected for the ith response. More specifically, for the ith case, $Y_{i1} = 0$, $Y_{i2} = 0,\ldots$, $Y_{il} = 1,\ldots$, $Y_{iL} = 0$. The probability of this response is given by:

$$P(Y_{il} = 1) = \eta_{il} = (\eta_{i1})^0 \times (\eta_{i2})^0 \times \cdots \times (\eta_{il})^1 \times \cdots \times (\eta_{iL})^0 = \prod_{l=1}^{L} (\eta_{il})^{Y_{il}} \tag{4}$$

For n independent observations and L levels for the response variable Y, the likelihood function can be constructed by:

$$\varphi = \prod_{i=1}^{n} P(Y_{il}) = \prod_{i=1}^{n}\left[\prod_{l=1}^{L}(\eta_{il})^{Y_{il}}\right] \tag{5}$$

Taking natural logarithm and using the fact that:

$$\sum_{l=1}^{L} Y_{il} = 1$$

for each i, the log-likelihood function is given by:

$$\text{Log}_e(\varphi) = \sum_{i=1}^{n}\left[\sum_{l=1}^{L-1}(Y_{il}\psi_l) - \text{Log}_e\left\{1+\sum_{l=1}^{L-1}e^{\psi_l}\right\}\right] \tag{6}$$

The likelihood equations can be found by taking the first partial derivatives of $\text{Log}_e(\varphi)$ with respect to each of (L-1) × (k+1) unknown parameters. The maximum likelihood estimates Υ_1, Υ_2,. .., $\Upsilon_{(L-1)}$ of are those values of $\hat{\gamma}_1, \hat{\gamma}_2, \cdots, \hat{\gamma}_{(L-1)}$, that maximize Eq. 6 and can be obtained by setting the likelihood equations equal to zero and solving for the vectors of parameters. These likelihood equations are non-linear in parameters and can be numerically solved by Newton-Raphson method. Hence one must rely on standard statistical software programs and iterative computation that is used to obtain these estimates. On the other hand, in order to test the significance of covariates, the matrix of second partial derivatives is required to get the information matrix and the estimator of the covariance matrix and consequently the standard error of the maximum likelihood estimators. The generalized form of the elements in the matrix of second partial derivatives is given in Eq. 7 and 8, respectively.

$$\frac{\delta^2 \text{Log}_e(\phi)}{\delta\gamma_{lv}\delta\gamma_{lv'}} = -\sum_{i=1}^{n} x_{vi}x_{v'i}\eta_{il}(1-\eta_{il}) \;;\; v \neq v' = 0,1,\cdots,k \tag{7}$$

$$\frac{\delta^2 \text{Log}_e(\phi)}{\delta\gamma_{lv}\delta\gamma_{l'v'}} = \sum_{i=1}^{n} x_{vi}x_{v'i}\eta_{il}\eta_{il'} \;;\; l \neq l' = 1,2,\cdots,(L-1) \tag{8}$$

The estimated or observed information matrix, denoted by $I(\hat{\gamma})$, is the (L-1) (k+1) × (L-1) (k+1) matrix whose elements are the negatives of the values obtained from the Eq. 7 and 8 evaluated at $\hat{\gamma}$. The estimated standard error (SE) of the maximum likelihood estimator is obtained from the positive square root of principal diagonal of inverse of the observed information matrix like $S.E(\hat{\gamma}) = \sqrt{I^{-1}(\hat{\gamma})}$. Although computationally different, the multinomial logit model produces results that are nearly identical to the general 2×2 contingency table having observed cell frequencies a, b, c, d (Collett, 1991). It is notable that in the multinomial logit model, the MLE estimation of the standard error of the estimate is quite close to the estimated standard error derived by using Woolf (1955) approach given by:

$$S.E(\hat{\gamma}_{lv}) = \left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)^{\frac{1}{2}} \;;\; l = 1,2,\cdots,(L-1) \;;\; v = 0,1,2,\cdots,k \tag{9}$$

Hence the estimator of the variance of the difference between two coefficients, $(\hat{\gamma}_{lv} - \hat{\gamma}_{l'v} ; l \neq l')$ is given by:

$$v(\hat{\gamma}_{lv} - \hat{\gamma}_{l'v}) = v(\hat{\gamma}_{lv}) + v(\hat{\gamma}_{l'v}) - 2\,cov(\hat{\gamma}_{lv}, \hat{\gamma}_{l'v}) \tag{10}$$

The values for the estimates of the variances and covariances can be obtained from software program SPSS through a listing of the estimated asymptotic covariance matrix. The form of this matrix is a little different from the covariance matrix in the binary setting. There are two matrices containing the estimates of the variances and covariances of the estimated coefficients in each logit and a third containing the estimated covariances of the estimated coefficients from the different logits. Such matrix for the multinomial logit model is not exhibited in the current analysis. In order to interpret the effect of covariates on response variable, a measure of association called odds ratio, a powerful analytic tool should be defined. The odds ratio, denoted OR, defined as the ratio of odds for a specific level to the odds for the referent level. In a multinomial outcome setting, the odds ratio of outcome Y = *l* versus outcome Y = L for a specific covariate x = r versus x = s is defined by:

$$OR_{lv}(r,s) = \frac{P(Y = l|x = r)/P(Y = L|x = r)}{P(Y = l|x = s)/P(Y = L|x = s)} ; l = 1,2,\cdots,(L-1) \tag{11}$$

In a multinomial logit model, the response variable Y having L distinct nominal levels, (L-1) logits are generated and consequently, (L-1) parameter estimates and corresponding odds ratios are found for each of the covariates. In case of any significant differences among the parameter estimates or the corresponding odds ratios under different logits are found, the adequacy of the multinomial logit model over binary model will be established.

RESULTS AND DISCUSSION

In order to display the findings of current study, suppose there are L = 3 levels in the response variable and the third level having the highest frequency is considered as referent level. SPSS multinomial logit output for nominal response is exhibited in Table 1 which contains the estimated regression coefficients, estimated approximate standard errors, the Wald test statistics with associated p-values, the estimated odds ratios, 95% confidence intervals for the odds ratios for the two estimated logits or linear predictors. The results of multinomial logit model can be expressed in the form of odds ratios, telling us how much change there is in the probability of being certain level under study, given a unit change in any other given covariate but holding all others covariates in the analysis constant. More simply, the results tell us how much a hypothesized cause has affected this response, taking the role of all other hypothesized causes into account. A preliminary indication of the importance of the explanatory variables in the model under different logits can be assessed through the Wald statistic. The Wald test is obtained by comparing the maximum likelihood estimates of the slope parameters $\hat{\gamma}_{lv}$, to the estimates of their corresponding standard errors $S.E(\hat{\gamma}_{lv})$. The estimates of the standard errors of the estimated parameters can be obtained from Eq. 9. The resulting ratios:

$$W_{lv} = \frac{\hat{\gamma}_{lv}}{S.E(\hat{\gamma}_{lv})}$$

Table 1: Estimated coefficients, estimated standard errors, Wald chi-square statistics with degrees of freedom and p-values, odds ratios and 95% confidence interval of odds ratios for the multinomial logit model to the BDHS-2007 data

								95% CI for OR_l	
Logit	Explanatory variables	$\hat{\gamma}_{lv}$	$S.E(\hat{\gamma}_{lv})$	W_{lv}	df	Sig.	OR_l	LB	UB
1 (1/3)	Intercept	-3.087	0.190	265.2	1	0	-	-	-
	X_1 = Rural	0.344	0.190	3.3	1	0.071	1.41	0.97	2.05
	X_1 = Urban	-	-	-	-	-	1.00	-	-
	X_2 = Poorest	0.997	0.276	13.0	1	.000	2.71	1.58	4.66
	X_2 = Poorer	0.995	0.260	14.7	1	.000	2.70	1.63	4.49
	X_2 = Middle	0.485	0.274	3.2	1	.076	1.62	0.95	2.79
	X_2 = Richer	0.423	0.263	2.6	1	.108	1.53	0.91	2.56
	X_2 = Richest	-	-	-	-	-	1.00	-	-
	X_3 = Professional	-0.363	0.172	4.5	1	0.034	0.7	0.5	0.98
	X_3 = House wife	-	-	-	-	-	1.00	-	-
	X_4 = Having sex preference	1.740	0.184	89.2	1	.000	5.7	3.9	8.17
	X_4 = No sex preference	-	-	-	-	-	1.00	-	-
2 (2/3)	Intercept	-2.377	0.139	290.9	1	.000	-	-	-
	X_1 = Rural	0.305	0.144	4.5	1	0.035	1.36	1.02	1.8
	X_1 = Urban	-	-	-	-	-	1.00	-	-
	X_2 = Poorest	0.88	0.214	16.8	1	.000	2.41	1.583	3.67
	X_2 = Poorer	0.882	0.200	19.5	1	.000	2.42	1.634	3.58
	X_2 = Middle	0.674	0.199	11.4	1	.001	1.96	1.328	2.89
	X_2 = Richer	0.391	0.198	3.9	1	0.048	1.48	1.003	2.18
	X_2 = Richest	-	-	-	-	-	1.00	-	-
	X_3 = Professional	-0.508	0.136	14.1	1	.000	0.6	0.461	0.78
	X_3 = House wife	-	-	-	-	-	-	-	-
	X_4 = Having sex preference	1.879	0.147	163.9	1	.000	6.55	4.91	8.73
	X_4 = No sex preference	-	-	-	-	-	1.00	-	-

under the null hypotheses $H_0 : \gamma_{lv} = 0$, will follow a standard normal distribution and hence equivalently:

$$W_{lv} = \frac{\hat{\gamma}_{lv}^2}{\left[S.E(\hat{\gamma}_{lv})\right]^2}$$

will follow chi-square distribution with single degree of freedom. Examination of the Wald statistics in Table 1 suggests that each of the explanatory variables may contribute to the model. From the statistical point of view, the findings exhibited in Table 1, all the explanatory variables irrespective of their levels under different logits are significantly associated with the response variables at 10% level of significance.

For polychotomous explanatory variable we can expand the number of odds ratios to include comparisons of each level of the variable to a reference level for each possible logit function. Thus the four estimated coefficients for the design variable wealth index (X_2) which estimate the log odds for poorest, poorer, middle and richer versus the reference value of richest, suggest that two categories poorest and poorer are similar, middle and richer are also similar since neither Wald statistics are significant. The sign and magnitude of the estimated coefficients for the accumulated design variables poorest with poorer and middle with richer suggest that the log odds of poorest

with poorer and middle with richer differ significantly from richest and are of similar magnitude within each of the two logit functions (test are not presented). The likelihood ratio test after accumulation of poorest with poorer and middle with richer yields the value G = 2.22 with four degrees-of-freedom with p = 0.70 indicates the accumulation of insignificant levels is worthwhile for polychotomous explanatory variables.

Hauck and Donner (1977) examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject the null hypothesis when the coefficient was significant. They recommended that the more robust likelihood ratio test should be used to justify the significance of individual predictor. Jennings (1986) has also looked at the adequacy of inferences in logistic regression based on Wald statistics. In order to avoid the uncertainty of inferences, he suggested that both the likelihood ratio test G and the Wald test W_{lv} require to test the significance of the maximum likelihood estimates for $\hat{\gamma}_{lv}$. The likelihood ratio test G is nothing but the change in the deviance of a model with single covariate and a full model where minus twice the log likelihood is known as deviance and denoted by D (Agresti, 2002). Under the same null hypotheses, likelihood ratio tests G follow chi-square distribution with (L-1)×(M-1) degrees of freedom. Here L and M are the number of levels of response variable and the corresponding explanatory variable, respectively. The output of the likelihood ratio tests are shown in Table 2 and it can be concluded that all the explanatory variables included in the model are significantly associated with the response variable at 5% level of significance.

In order to fit a model, it is important to have tools to test for lack of fit, especially important for the multinomial logit model, whose fit is notoriously difficult to visualize. Such tools are remarkably scarce in multinomial logistic regression applications (Goeman and le Cessie, 2006). In such a situation, Deviance and Pearson's chi-square goodness-of-fit test can be employed whether the model adequately fits the data. In these tests, lack of fit is indicated by the significance value less than 0.05. To support the adequacy of the fitted model, a significance value greater than 0.05 is needed. If no warning message is given from the program or the number of subpopulations with zero frequencies is small with p>0.05, it may be concluded that the model fits the data well. In the current study, Deviance chi-square value is 60.00 having 64 degrees of freedom with significance value 0.62 and Pearson's chi-square value is 56.71 having 64 degrees of freedom with significance value 0.73. The large p-values for both the goodness-of-fit tests signify the adequacy of the fitted multinomial logit model.

The main objective of this study was to test the equality of the two odds ratios, $H_0 : OR_{1v} = OR_{2v}$ (v = 1,2,3,4) under two different logits which is equivalent to a test that the log-odds for Y = 2 versus Y = 1 is equal to zero, simply $H_0 : \gamma_{lv} = \gamma_{2v}$. The simplest way to obtain the point and interval estimate is from difference between the two estimated slope coefficients in the multinomial logit model. Using the output of the asymptotic variance-covariance matrix produced by the multinomial logit model, it can easily be obtained the estimator of the variance of the difference between the two estimated coefficients and the endpoints of a 95% confidence interval for this difference and summarized in Table 3.

Unfortunately, the findings with high p-values suggest that there is no significant difference between the logits over the entire set of explanatory variables. Equivalently, the confidence intervals exhibited in Table 3 for all the explanatory variables include zero and hence cannot concluded that the log odds for Y = 2 is different from the log odds for Y = 1. In practice, if there is no difference in the separate odds ratios over all model covariates then one should consider pooling

Table 2: Likelihood ratio test for the significance of overall and individual importance of explanatory variables in the model

Model with explanatory variables	D = -2Log likelihood	G	df	Sig.
Null model	594.432	299.861	14	<0.001
X_1	301.241	6.669	2	<0.050
X_2	333.400	38.828	8	<0.001
X_3	311.386	16.814	2	<0.001
X_4	490.982	196.410	2	<0.001
Full model	294.572	-	-	-

Table 3: Estimated coefficients under different logits, their differences, significant or p-values, 95% confidence interval for the difference in multinomial logit model

Variables	Coeff. of Logit 1	Coeff. of Logit 2	Difference	Sig.	95% CI for the difference
X_1 = Rural	0.344	0.305	0.039	0.86	(-0.386, 0.464)
X_2 = Poorest	0.997	0.88	0.117	0.71	(-0.497, 0.730)
X_2 = Poorer	0.995	0.882	0.113	0.71	(-0.475, 0.701)
X_2 = Middle	0.485	0.674	-0.191	0.54	(-0.799, 0.417)
X_2 = Richer	0.423	0.391	0.032	0.92	(-0.562, 0.626)
X_3 = Professional	-0.363	-0.508	0.145	0.47	(-0.247, 0.537)
X_4 = Having sex preference	1.74	1.879	-0.139	0.46	(-0.511, 0.233)

multi response levels into a binary 'ever' versus 'never' responses and more complicated multinomial model is no better than the simpler dichotomous logit model.

CONCLUSION

Logistic regression is a form of regression analysis that is specifically tailored to the situation in which the response variable is dichotomous or polychotomous. Response variable having more than two levels is a situation frequently faced in the categorical data analysis. If the levels of the response variable are nominal scale, nominal multinomial logistic regression is appropriate. Multinomial logistic regression is increasingly common, involving analyses in which the possible causal effects of explanatory variables on a categoric response variable having more than two response categories are assessed via comparison of a series of dichotomous responses. The multinomial logit model is a generalization of dichotomous logit model but complicated in terms of fitting process and interpretation. In case of significant difference is found in the separate odds ratios produced by the different logits over the entire set of explanatory variables, it may be concluded that multinomial logit model adequately fit the data with response variable having more than two levels, otherwise response levels should be pooled into binary levels for ease of computation, mathematical tractability and ease of interpretability. In the current study, there is no significant difference between the parameter estimates under different logits and it may be finally concluded that dichotomous logit model is preferred choice than complicated multinomial logit model.

REFERENCES

Agresti, A., 2002. Categorical Data Analysis. 2nd Edn. John Wiley and Sons, Inc., Publication, New Jersey, ISBN: 9780471360933.

Begg, C.B. and R. Gray, 1984. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika, 71: 11-18.

Chan, Y.B., 2005. Multinomial logistic regression (Biostatistics 305). Singapore Med. J., 46: 259-269.

Chuang, H.L., 1997. High school youths' dropout and re-enrolment behavior. Econ. Educ. Rev., 16: 171-186.

Collett, D., 1991. Modeling Binary Data. 2nd Edn., Chapman and Hall, London, ISBN: 9781584883241, pp: 387.

Darlington, R.B., 1990. Regression and Linear Models. McGraw-Hill, New York, ISBN: 0070153728, pp: 542.

Goeman, J.J. and S. le Cessie, 2006. A goodness-of-fit test for multinomial logistic regression. Biometrics, 62: 980-985.

Harrell, F., 2001. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis. Springer, New York, ISBN: 0-387-95232-2, pp: 568.

Hauck Jr, W.W. and A. Donner, 1977. Wald's test as applied to hypotheses in logit analysis. J. Am. Stat. Assoc., 72: 851-853.

Hosmer, W.D. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., John Wiley and Sons, New York, ISBN-10: 0471356328, pp: 392.

Jennings, D.E., 1986. Judging inference adequacy in logistic regression. J. Am. Stat. Assoc., 81: 471-476.

Kutner, M.H., C.J. Nachtsheim, J. Neter and W. Li, 2005. Applied Linear Statistical Models. 5th Edn., McGraw-Hill, New York, ISBN: 0-07-310874-X, pp: 1424.

Long, J.S., 1987. A graphical method for the interpretation of multinomial logit analysis. Sociol. Methods Res., 15: 420-446.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapmann and Hall, London, ISBN: 0-412-31760-5, pp: 536.

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In: Frontiers in Econometrics, Zarembka, P. (Ed.). Academic Press, New York, pp: 105-142.

Peng, C.Y., K.L. Lee and G.M. Ingersoll, 2002. An introduction to logistic regression analysis and reporting. J. Educ. Res., 96: 3-14.

Pohlmann, J.T. and W.L. Dennis, 2003. A comparison of ordinary least squares and logistic regression. Ohio J. Sci., 103: 118-125.

Woolf, B., 1955. On estimating the relation between blood group and disease. Ann. Human Genet., 19: 251-253.