Bangladesh J. Sci. Res. **26**(1&2): 75-82, 2013 (December)

COMPARISON OF CONDITIONAL LIKELIHOOD AND MODIFIED SCORE FUNCTION APPROACHES IN THE ANALYSIS OF MATCHED CASE-CONTROL DATA

Abdul Baten*, Taslim S. Mallick[1] and Jafar A. Khan[1]

Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh

Abstract

In this study, we compare Maximum Conditional Likelihood or MCL and Modified Score Function or MDS approach for modeling Antenatal Care (ANC) seeking behavior in matched case-control data obtained from Bangladesh Demographic and Health Survey 2007. The estimates of parameters are almost identical under both approaches. Due to the computational complexity of the MDS approach, one may prefer using the alternative computationally simpler MCL approach for the analysis of similar datasets.

Key words: Conditional logistic regression, matched case-control data, maximum conditional likelihood, modified score function

Introduction

A logistic regression model (Hosmer and Lemeshow 1989) is commonly used to study the relationship between a binary or dichotomous response variable and one or more explanatory variables. The classical maximum likelihood estimator of a logistic regression model works best when the degrees of freedom for the model is small compared to the number of observations. For large degrees of freedom, which occur in a matched case-control design, the conditional estimation approach is better (Breslow and Day 1980). Matched designs are commonly used in the situation where both the disease probability and the exposure of interest depend on a common set of variables (Sun *et al.* 2011). These common variables cannot be used as predictors. They are used as matching variables, so that the true relationship between the response and predictors is not confounded. Age and sex are commonly used as matching variables. Within each stratum, samples of cases (y = 1) and controls (y = 0) are chosen. The number of cases and controls need not be constant across the strata, but the most common matched designs include one case and M controls per stratum and are thus referred to as 1 *: M* matched studies (Hosmer and Lemeshow 1989).

Suppose, we have *1 : M_i (≥ 1)* matched case-control dataset, with n strata. Let the response (case-control) variable Y_{ij} take on value 1 or 0 accordingly to whether the J^{th} subject in the i^{th} matched set is a case or control, respectively and $X_{ij} = (X_{ij1}, ..., X_{ijp})^T$ is a $p \times 1$ vector of covariates associated with Y_{ij}. Also let S_i be the set of variables which are used for matching purpose in the i^{th} stratum.

Suppose, the disease risk model for the i^{th} stratum is

$$pr(Y_{ij} = 1 | S_i, X_{ij}) = H(\alpha_i(S_i + X_{ij}^T \beta)) \quad (1)$$

*Corresponding author: <pulokstat@yahoo.com>. [1]Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1000

for $j = 1, ..., M_i + 1$ and $i = 1, ..., n$ with $H(z) = \{1 + exp(-z)\}^{-1}$. Here M_i is the number of controls for each stratum, α_i is the stratum-specific parameter which is a function of S_i and $\beta = (\beta_1, ..., \beta_p)^T$ is the vector of parameters for the covariates X_{ij}.

Sun *et al.* (2011) used conditional logistic regression model for a low birth weight case-control study and used the MCL (Breslow and Day 1980) for the estimation of the parameters. They also extended Firth's (1993) modified score function (MDS) based approach for the estimation purpose in the conditional logistic regression model. They conducted simulation study to show that the MDS approach is more consistent than the MCL. In this study, we would like to compare the MCL and MDS approaches for modeling antenatal care (ANC) seeking behavior in matched case-control data obtained from Bangladesh Demographic and Health Survey 2007 (BDHS 2007). For the computational suitability of MCL, we use this approach to further explore the impact of the selected covariates on ANC.

Estimation techniques for matched case-control binary data

Conditional logistic regression works in nearly the same way as regular logistic regression, except that we need to specify which individuals belong to which matched set (e.g. which pair or stratum). The conditional analysis has a higher (less negative) log likelihood, which suggests a somewhat better "fit". The MDS method is applicable for a matched case-control study with varying number of controls in each stratum as long as M_i' s are bounded as $n \to \infty$.

Estimation of the parameter by MCL approach: For estimating the parameter β in equation (1), Sun *et al.* (2011) suggested to adopt the conditional logistic regression (Breslow and Day 1980), where the estimate of the parameters are obtained by maximizing the following likelihood function

$$L_{CLR}(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{M_i+1} p_{ij}^{Y_{ij}}, \qquad (2)$$

where $p_{ij} = \exp\left(x_{ij}^T \beta\right) / \sum_{k=1}^{M_i+1} \exp\left(x_{ik}^T \beta\right)$ is the conditional probability in which the j^{th} subject is a case and there is one case in the i^{th} stratum. The condition $\sum_{j=1}^{n} Y_{ij} = 1$ is a sufficient statistics for $\alpha_i (S_i)$, which is used to obtain the L_{CLR}.

Taking log on both sides of (2), we get log-conditional likelihood function as

$$\log L_{CLR}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} Y_{ij} \log p_{ij} . \qquad (3)$$

To obtain the MCL estimates of the parameters, we have to differentiate *log* L_{CLR} *(β)* with respect to β and set the resulting expression equal to zero.

Now in order to estimate the parameters by the Newton-Raphson method, we first have to obtain the score vector *U(β)* and the observed information matrix *I(β),* where *β* is the vector of the parameters. The score function for h^{th} element of *U(β)* may be obtained from (3) as

$$U_h(\beta)=\left[\frac{\partial \log L_{CLR}(\beta)}{\partial \beta_h}\right]_{p\times 1}=\left[\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\left(Y_{ij}-p_{ij}\right)X_{ijh}\right], \text{ for h = 1, ..., p.} \quad (4)$$

The $\left(h,l\right)^{th}$ element of the information matrix is obtained from (3) as

$$I_{hl}(\beta)=\left[-\frac{\partial^2 \log L_{CLR}(\beta)}{\partial \beta_h \partial \beta_l}\right]=\sum_{i=1}^{n}\left\{\sum_{j=1}^{M_i+1}X_{ijh}X_{ijl}^{T}p_{ij}-\left(\sum_{j=1}^{Mi+1}X_{ijh}p_{ij}\right)\left(\sum_{j=1}^{M_i+1}X_{ijl}p_{ij}\right)^{T}\right\}, \quad (5)$$

for *h, l = 1, ..., p.*

The maximum likelihood estimating equation is then *U(β) = 0*. This equation can be solved for β by Newton-Raphson method

$$\beta^{m+1}=\beta^{m}+\left[I\left(\beta^{m}\right)\right]^{-1}U\left(\beta^{m}\right), \quad (6)$$

where $I\left(\beta^{m}\right)$ and $U\left(\beta^{m}\right)$ are the information matrix and score vector respectively, evaluated at the estimates obtained at m^{th} iteration.

Estimation of the parameter by MDS approach: For cross-section studies generally MCL estimator (Firth 1993) is commonly used and regular logistic regression is used as a response model. But for the matched case-control studies, it is common to use a conditional logistic regression model (Sun *et al.* 2011) and the estimators are obtained by maximizing conditional likelihood which gives biased result if the sample size is not sufficiently large. To reduce this bias, Firth (1993) introduced a bias-preventive method by solving a modified score function under an unconditional logistic regression setup and later Sun *et al.* (2011) have used it under conditional logistic regression model for matched case-control data. Let U^{mod} *(β)* denote the modified conditional score function. Following Firth (1993) and Sun *et al.* (2011), the r^{th} component of the modified conditional score vector may be written as

$$U_r^{\text{mod}}\left(\beta\right)=U_r\left(\beta\right)+\frac{1}{2}\frac{\partial}{\partial \beta_r}\left\{\log\left|I\left(\beta\right)\right|\right\}$$

$$=U_r\left(\beta\right)+\frac{1}{2}tr\left\{I^{-1}\left(\beta\right)\frac{\partial I\left(\beta\right)}{\partial \beta_r}\right\}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\left(Y_{ij}-p_{ij}\right)X_{ijr}+$$

$$\frac{1}{2}tr\left[I^{-1}(\beta)\left\{ \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\left(X_{ij}-2\bar{X}_{i.}\right)X_{ij}^{T}\left(X_{ijr}-\bar{X}_{i.r}\right)p_{ij}\right\}\right] \qquad (7)$$

for $r = 1, ..., p$.

where, $\bar{X}_{i.r}=\sum_{j=1}^{M_i+1}X_{ijr}p_{ij}$ and $\bar{X}_{i.}=\sum_{j=1}^{M_i+1}X_{ij}p_{ij}$

Firth (1993) showed that the modified score estimator has the same asymptotic variance covariance matrix as that of the MCL estimator. Therefore, in order to obtain parameter estimates by MDS approach one may use (6) only by replacing $U(\beta)$ with $U^{mod}(\beta)$ using the same $I(\beta)$ as in (5).

Data and variables

The objective of this study is to compare MCL and MDS approaches for modeling antenatal care (ANC) seeking behavior using BDHS 2007 dataset.

It is known that age, education, type of place of residence, and wealth index play significant role for the utilization of maternal health care (Islam *et al.* 2004). Using logistic regression on BDHS 2004 data Rahman *et al.* (2008a) showed that mothers education, child ever born, wealth index, telling about pregnancy are the significant determinants of receiving ANC. Using multivariate logistic regression on BDHS 2004 data, Rahman *et al.* (2008b) showed that higher educated women were two and a half times more likely to receive assistance from medically trained personnel than women with no education. They also found that the main contributing factors likely to affect delivery practices were mass media exposure, household occupation, household quality index etc.

For the purpose of the study, we consider only four explanatory variables, namely place of residence, wealth index, birth order of child and education level to study their effects on ANC. We define ANC as dichotomous variable whether the respondent take antenatal care during pregnancy or not. The World Health Organization (WHO) recommends that pregnant women make at least four ANC visits, beginning during the first trimester of the pregnancy. Since the dataset comprises of women mostly from rural area of Bangladesh, where majority of them are not conscious about Health Care Centre (HCC) visits, we consider a pregnant woman to receive ANC if she has at least two antenatal visits. To be specific, we define case-control variable ANC denoted by Y_{ij} as follows:

$$Y_{ij}=\begin{cases} 1, \text{ if the } j^{th} \text{subject in the } i^{th} \text{matched set visits health care centre more than once} \\ 0, \text{ otherwise} \end{cases}$$

Prior to matching, the initial dataset with the above variables consisted of 4917 individuals. The distribution of these selected individuals with respect to different factors that may be thought of associated with ANC related issue are shown in Table 1.

From Table 1, it is clear that all the covariates are significantly associated with ANC (p -value < 0.01). To be specific, 63.3% women who live urban area take antenatal care, whereas 38.0% women who live in rural area take antenatal care. The rate of receiving antenatal care is significantly higher for the rich people as compared to the other wealth index categories. Higher frequency for the antenatal care has also been observed for the high educated class and for those women having their first pregnancy. This suggests the need of extensive counseling to the poor and less educated women for increasing consciousness about taking antenatal care.

Note that our objective is to analyze a matched case-control BDHS data for comparison of MCL and MDS approaches. For matching purposes, usually age, sex, ethnic group etc. are considered to be the matching variables. In this study we use age as a matching variable, because age is known to be a confounder that influences both the covariates and the response. As described by Table 1, initially in our case-control BDHS dataset, we have selected 4917 individuals. For a particular category of age i, suppose there are n_{1i} individuals with k cases and $(n_{1i} - k)$ controls. For a 1:2 matching, we have to select $2k$ controls randomly from $(n_{1i} - k)$ and discard the remaining controls from the dataset. Following this procedure for all categories of age, we have selected 3825 individuals, 1275 cases and 2550 controls, in our matched case-control dataset extracted from the BDHS 2007.

Table 1. Frequency wistribution of ANC by the set of selected explanatory variables.

Variable	Category	ANC		Total
		Yes (%)	No (%)	
Place of residence*	Urban	1103(63.3)	640(36.7)	1743
	Rural	1207(38)	1967(62.0)	3174
Wealth index*	Poor	544(28.3)	1380(71.7)	1924
	Middle	358(39.3)	522(60.7)	910
	Rich	1408(67.6)	675(33.8)	2083
Birth order of child*	1	938(60.9)	602(39.14)	1540
	2 - 4	1217(45.2)	1477(54.8)	2694
	≥ 5	155(22.7)	528(77.3)	683
Education levels*	No	287(22.7)	980(77.3)	1267
	Primary	569(37.8)	935(62.2)	1504
	High	1454(67.8)	692(32.2)	2146

*p value < 0.001.

A comparison between MCL and MDS estimation methods

The MDS approach (Sun *et al.* 2011) is expected to produce estimates with smaller bias as compare to the MCL approach when a conditional logistic regression is assumed for the matched case-control data. For the selected BDHS dataset, we have 1275 strata, (i.e, $i = 1, 2, ..., 1275$) and

we have considered $M_i = 2$ for all i. Under 1 : 2 matching, each stratum contains 3 individuals (j = 1, 2, 3), 1 case and 2 controls, therefore, we have 3825 individuals in our study.

In this section, we compare MCL and MDS approaches for estimating regression effects of the selected covariates on the ANC for the matched case-control BDHS 2007 data. We will consider 4 covariates, namely place of residence (Rural = 0, urban = 1), birth order of child (1or less = 0, 2 or more = 1), education level (No = 0, others = 1), wealth index (Poorest = 0, others = 1). Since MDS approach is computationally intensive, instead of considering all 4 covariates at once, we consider 6 different designs, each of which consists of 2 covariates. Therefore, for each design, we compare two regression estimates obtained by MCL and MDS approaches. The estimates obtained by MCL and MDS approaches along with their standard errors are reported in Table 2.

From Table 2, it was found that the estimates of parameters both in MCL and MDS approaches are approximately equal. As for example, for design 1, the MCL estimates for type of place of residence were found to be 1.072, which is obtained as 1.070 under MDS approach. On the other hand, under the MCL and MDS estimates, the effects of education level were found to be 1.427 and 1.423, respectively. The similarity between these two approaches may be due to the fact that we have large number of individuals in our sample. We therefore, conclude that the performance of MCL and MDS approaches are very similar for the BDHS 2007 data.

Table 2. Estimation of the conditional logistic regression model parameter for matched case-control BDHS 2007 by MCL and MDS approaches.

Design	Covariates	Estimation methods			
		MCL		MDS	
		$\hat{\beta}$	SE	$\hat{\beta}$	SE
1	Place of residence	1.072*	0.079	1.070	0.079
	Education level	1.427*	0.104	1.423	0.104
2	Place of residence	1.114*	0.077	1.113	0.078
	Birth order of child	−1.126*	0.107	−1.123	0.107
3	Place of residence	0.972*	0.078	0.970	0.078
	Wealth index	0.884*	0.108	0.880	0.109
4	Education level	1.423*	0.102	1.420	0.102
	Birth order of child	−1.038*	0.106	−1.036	0.106
5	Education level	1.329*	0.103	1.325	0.103
	Wealth index	0.891*	0.109	0.889	0.109
6	Birth order of child	−1.037*	0.104	−1.035	0.104
	Wealth index	1.062*	0.106	1.059	0.106

*p value < 0.001.

MCL estimates for BDHS 2007 matched case-control data

The findings from the last subsection motivate us to use simpler MCL approach for estimating the effects of several factors on ANC seeking behavior for the selected matched case-control BDHS

dataset. To be specific, we include all 4 covariates in the MCL approach in order to study their effects on ANC seeking behavior. Table 3 reports the categorization of the selected covariates and Table 4 reports the estimates obtained under MCL approach. Note that the selected categories of covariates, we have p = 7 parameters in the MCL estimating equation (6).

From Table 4, it is observed that, all the covariates have significant effect on the antenatal care seeking behavior. To be specific, women who are in urban family are more likely to receive antenatal care. The odds of receiving ANC for those in urban area are 1.828 times the odds for those in the rural area. The respondents who are from middle-class and rich families are more likely to receive antenatal care during pregnancy. The odds of women who are from middle and Rich family are 1.235 and 2.545 times, respectively than the odds for women from the poor family.

Table 3. Covariate categories for MCL approach.

Variables	Categorization
Place of residence	Urban =1
	Rural = 0
Birth order of child	1or less = 1
	2 - 4 = 2
	5 or more = 3
Education level	No = 0
	Primary = 1
	Secondary and higher = 2
Wealth index	Poorest and poorer = 1
	Middle = 2
	Richer and richest = 3

Table 4. Estimates, standard errors (SE) and odds ratios (OR) of the conditional logistic regression model parameters for matched case-control BDHS 2007 data by MCL approach.

Variable	$\hat{\beta}$	SE	OR
Place of residence (Ref = Rural)			
Urban	0.630*	0.060	1.828
Wealth Index (Ref = Poor)			
Middle	0.211*	0.091	1.235
Rich	0.934*	0.077	2.545
Birth Order of child (Ref = 1)			
2 - 4	−0.505*	0.055	0.604
≥ 5	−1.357*	0.104	0.257
Education level (Ref = No)			
Primary	0.652*	0.076	1.919
Secondary and High	1.459*	0.071	4.302

*p value < 0.001.

On the other hand, women tend to skip antenatal visits when she has more than one child already. More specifically the odds of receiving ANC for women with preceding 2 - 4 birth is 0.604 times the odds for those with the first pregnancy. This negligence of skipping ANC visits is

even more when she had preceding 5 births; the odds is 0.257 times than the odds of women with first birth order.

From Table 4, it was also found that the women with primary and high education are more likely to seek antenatal care during the pregnancy. For example, the odds of receiving ANC for high education group is 4.3 times the odds for no education group.

Conclusion

We have compared MCL and MDS approaches for modeling antenatal care (ANC) seeking behavior in matched case-control data from Bangladesh Demographic and Health Survey 2007. We found that the estimates under both approaches are quite similar - place of residence, wealth index and education levels have positive impact and birth order of child has negative impact on ANC seeking behavior. Due to the computational complexity of the MDS approach, one may prefer using the alternative computationally simpler MCL approach for the analysis of similar data sets.

Acknowledgement

The authors would like to thank Dr. Wasimul Bari for his help and valuable comments on the preparation of this manuscript.

References

Breslow, N.E. and N.E. Day. 1980. Statistical methods in Cancer Research, Volume 1 - The Analysis of Case-Control Studies. *IARC: Lyon.*

Bangladesh Demographic and Health Survey. 2007, National Institute of Population Research and Training (NIPORT), Dhaka, Bangladesh., Mitra and Associates, Dhaka, Bangladesh. And Macro International, Inc.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27-38.

Hosmer, D.W. and S. Lemeshow. 1989. Applied Logistic Regression. Wiley, New York.

Islam, M.A., R.I. Chowdhury, N. Chakraborty and W. Bari. 2004. A multistage model for maternal morbidity during antenatal, delivery and postpartum periods. *Statistics in Medicine* **23**: 137-158.

Rahman, M.M., RM. Islam and A.Z. Islam. 2008a. Rural urban differentials of utilization of antenatal health care services in Bangladesh. *Health Policy and Development* **6**(3): 117-125.

Rahman, M., T.I. Tarafder and G. Mostofa. 2008b. Modes of delivery assistance in Bangladesh. *Tanzania J. Health Res*. **10**(4): 246-252.

Sun, J.X., S. Sinha, S. Wang and T. Mati. 2011. Bias reduction in conditional logistic regression. *Statistics in Medicine* **3**0: 348-355.

(Manuscript received on 26 August, 2013; revised on 15 December, 2013)