Adv. & Appl. in Stat. **7**(2) (2007), 281-290

ESTIMATION OF PARAMETERS IN UNCONDITIONAL CATEGORICAL REGRESSION MODELS WITH INCOMPLETE DATA IN COVARIATES

K. AZAM[1], A. GRAMI[2], K. MOHAMMAD[1], GH. JANDAGHI[3], M. KARIMLOU[4] and A. KAZEMNEJAD[5]

[1]Department of Epidemiology and Biostatistics
School of Public Health and Institute of Public Health
Tehran University of Medical Sciences, Iran

[2]Faculty of Science, University of Tehran, Iran

[3]University of Tehran, Qom Campus, Iran

[4]University of Welfare and Rehabilitation, Iran

[5]University of Tarbiat Modarres, Iran

Abstract

In large–scale sampling, we are always facing non-responses item(s) non-response or unit(s) or both. In fitting a model to the data we have two groups of variables, namely dependent and independent variables. Non-response may occur for any of these groups of variables. In this paper we assume that Y as a categorical dependent variable, Z and X as independent variables. The first two variables are fully observed and we assume that the mechanism of missing-ness is random (MAR). In order to estimate parameters a model is devised based on likelihood function for the whole data set including missing data and the estimation of parameters are compared with those obtained by statistical software such as *S*-Plus which are only based on complete observed data and ignore missing units.

2000 Mathematics Subject Classification: 62J12, 62F10.

Keywords and phrases: missing at random, logistic regression, Goiter disease, maximum likelihood.

Received January 25, 2007

Our results show that the estimations obtained using maximum likelihood based model are superior to the standard estimations for the approach utilized by the soft wares. The comparison is made on a set of health survey data on Goiter disease carried out in Qazvin province.

1. Introduction

In the last decades, logistic regression models have played an important role in medical and epidemiological researches. Because of non-linearity of logistic models, the inference is made by maximum likelihood estimation, for there is no limitation considered about independent variables. In the cohort or cross-sectional studies in which no matching is done, the model parameters are estimated via unconditional maximum likelihood. Furthermore, in these types of studies, we may face data in which some part is missing due to unwillingness to respond, incompleteness of the questionnaire, incompleteness of the study frame and so on.

Generally, there are three mechanisms of missingness, missing at random (MAR), missing completely at random (MCAR), and non-ignorable missing (NI) (Little and Rubin [5]).

In MCAR, the missingness in one variable does not depend on itself or other covariates. So, we can eliminate the cases with missing values and do the analysis based on the remaining cases without any bias in estimates. In MAR, the missingness in one variable is independent of it but depends on the other covariates. For example, in studying the association between blood pressure and smoking, missingness in blood pressure depends on smoking but not on itself. In comparison to smokers, non-smokers, because of their sensitivity to their health, have more willingness to participate in the study.

As it is evident from example, the problem of missingness can mislead the analysis toward biased interpretations. Missingness can occur either in response variable or covariates. In our study the missingness in covariates with MAR mechanism of missing is of interest.

Treating missing values, generally three methods are used (Gao and Hui [4]). The simplest method which is used as a default in most statistical softwares is eliminating the cases with missing values and

doing the analysis on the remaining complete cases. This method usually introduces bias in estimates (Little and Rubin [5]). In the second method, one replaces the missing values by their means, using regression or other mechanisms, and does the analysis by standard methods of estimation. In this approach, if there is a large amount of data, one would face two important problems. Firstly, it may change the natural shape of the distribution of covariates and secondly, the mean and variance of the estimates of the parameters are changed. The third method which has recently captured the attention of the researchers is to determine a probability model for the variables with missing values. These probability models act similar to standard models with some changes in likelihood function.

In classical statistics, the logistic models are developed by maximum likelihood estimation and expectation maximization (EM) algorithm. This method, in addition to computational difficulties has some technical problems, i.e., one may reach a local instead of a global maximum or the iterative algorithm may not converge. Furthermore, we may face serious problems in analyzing small samples which would yield estimates, not sharing the asymptotic characteristics of MLE estimates.

Fuchs [3] and Little and Schluchter [6] used EM algorithm for maximum likelihood estimation of the parameters in logistic regression with missing discrete covariates and combined discrete-continuous missing covariates. This method generally needs many iterations and may be computationally intensive. Blackhurst and Schluchter [2] proposed a maximum likelihood method using EM algorithm which does not need iterations when the continuous missing covariate follows normal distribution.

Satten and Kupper [11, 12] generalized the analysis of logistic models with missing covariates and used alternative covariates to find information about the effects of missing covariates. Paik and Sacco [8] and Satten and Carroll [10] assumed a distribution for missing covariates and by introducing some changes in conditional and unconditional likelihood functions refined the estimates. Rathouz et al. [9] introduced a new class of estimates on the basis of modeling the missing covariates and mechanisms of missingness. Bayesian statisticians also have authored some papers on this issue.

In this study we have used the third method by determining a probability model for missing covariates and using Satten and Carroll [10] likelihood function.

2. Model, Definitions and Likelihood Function

Suppose Y_i is a binary response variable corresponding to subject i such that $Y_i = 1$ and $Y_i = 0$ denotes the diseased and non-diseased individual, respectively. We also suppose that the vector covariate **Z** is fully observed and the vector covariate **X** has missing values for some of its observations. Without loss of generality, we present our models with one Z and one X covariate. First, we assume both variables are fully observed. In this case the conditional probabilities given the covariates are defined based on a logistic model as follows:

$$P(Y = 1 \mid X = x, Z = z) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz)}. \tag{1}$$

$$P(Y = 0 \mid X = x, Z = z) = 1 - p(y = 1 \mid X = x, Z = z). \tag{2}$$

So, based on the above definitions the odds and odds ratio will be:

$$\theta(x, z) = \frac{P(Y = 1 \mid X = x, Z = z)}{P(Y = 0 \mid X = x, Z = z)} = \exp(\beta_0 + \beta_1 x + \beta_2 z + \beta_{12}\, xz), \tag{3}$$

$$\psi(x, z, x', z') = \frac{\theta(x, z)}{\theta(x', z')}, \tag{4}$$

where x' and z' are some levels of X and Z different from x and z.

The purpose of fitting a logistic regression model to data is to obtain an estimate of model parameters (here $\beta_0, \beta_1, \beta_2, \beta_{12}$) to determine the association between the response variable and a set of X and Z covariates. In case of complete data, we use standard methods of estimation. But if we assume that some values of X are missing, then we have

$$\tilde{\theta}(Z) = \frac{P(Y = 1 \mid Z = z)}{P(Y = 0 \mid Z = z)}. \tag{5}$$

In addition we define

$$\pi(x \mid z) = P(X = x \mid Y = 0,\ Z = z), \tag{6}$$

$$\rho(x \mid z) = P(X = x \mid Y = 1,\ Z = z), \tag{7}$$

where the functions π and ρ denote the probability distributions in non-diseased and diseased subjects, respectively. Satten and Kupper [11, 12] showed that

$$\tilde{\theta}(z) = \sum_{x} \theta(x, z) . \pi(x \mid z), \tag{8}$$

where the summation is taken over all values of $\mathbf{X}$. Their second result is

$$\rho(x \mid z) = \frac{\pi(x \mid z)\,\theta(x,\ z)}{\sum_{x} \pi(x \mid z)\,\theta(x,\ z)}. \tag{9}$$

In case of a continuous covariate X, we shall use integration instead of summation and $\pi(x \mid z)$ and $\rho(x \mid z)$ will be density functions.

The likelihood function for standard logistic regression when X and Z are fully observed is

$$L(\beta) = \prod_{i=1}^{n} \frac{[\theta(x_i,\ z_i)]^{y_i}}{1 + \theta(x_i,\ z_i)}. \tag{10}$$

If X contains missing values, we define the indicator variable Δ_i whose value is $\Delta_i = 1$ when x_i is observed and $\Delta_i = 0$ when x_i is missing. By this definition, the likelihood function is rewritten as follows:

$$P(Y,\ X,\ \Delta \mid Z) = P(Y \mid Z) P(\Delta \mid Y,\ Z) P(X \mid Y,\ Z,\ \Delta). \tag{11}$$

Under MAR *missingness*, we can assume that (Little and Rubin [5]):

$$P(X \mid Y,\ Z,\ \Delta) = P(X \mid Y,\ Z).$$

We also assume that the probability of missingness $P(\Delta \mid Y,\ Z)$ does not depend on vector parameter β. So, by elimination of this from the

likelihood function, the unconditional likelihood for observed data will become

$$L(\beta) = \prod_{i-1}^{n} \tilde{\theta}(z_i)^{y_i}[1+\tilde{\theta}(z_i)]^{-1}\pi(x_i \mid z_i)^{\Delta_i(1-y_i)}\rho(x_i \mid z_i)^{\Delta_i y_i}. \tag{12}$$

In the above equation, the distribution $\pi(x \mid z)$ is unknown, which can be derived once a model for π is chosen. A particularly attractive case is when X and Z take only finitely many values, Satten and Carroll [10].

$$\pi(x \mid z) = \frac{e^{\gamma_{xz}}}{\sum_{x'} e^{\gamma_{x'z}}} = \frac{e^{\gamma_0+\gamma_1 x+\gamma_2 x+\gamma_{12}xz}}{\sum_{x'} e^{\gamma_0+\gamma_1 x'+\gamma_2 z+\gamma_{12}x'z}} = \frac{e^{\gamma_1 x+\gamma_{12}xz}}{\sum_{x'} e^{\gamma_1 x'+\gamma_{12}x'z}}. \tag{13}$$

Substituting (3) through (9) in equation (13) and rewriting the likelihood function (12), we have

$$l(\beta \mid X,\ Z) = \prod_{i=1}^{n} \frac{(e^{\beta_0+\beta_1 x_i+\beta_2 z_i+\beta_{12}x_i z_i})^{\Delta_i y_i}}{1+\sum_x \left[e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})xz_i} \cdot \frac{1}{\sum_x e^{\gamma_1 x+\gamma_{12}xz_i}} \right]}$$

$$\times \left\{ \sum_x e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})xz_i} \cdot \frac{1}{\sum_x e^{\gamma_1 x+\gamma_{12}xz_i}} \right\}^{y_i(1-\Delta_i)}$$

$$\times \left(\frac{e^{\gamma_1 x_i+\gamma_{12}x_i z_i}}{\sum_x e^{\gamma_1 x+\gamma_{12}\,xz_i}} \right)^{\Delta_i}, \tag{14}$$

which is a function of parameters $\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1$ and γ_{12}. Taking logarithm of the both sides of (14), partially differentiating with respect to parameters, and setting the derivatives equal to zero, we obtain a nonlinear system of equations which is solved by numerical methods.

2.1. Estimation method

In this study, we used *R* software to program and solve the system of nonlinear equations, using iterative methods by following phases:

Phase 1. In this phase both X and Z covariates were completely observed and the model parameters were estimated using *R* code with precision of $e = 0.0001$ and compared with standard *S*-plus estimates.

Phase 2. We made some values of X variable missing using MAR mechanism and repeated the estimation procedure with both *R* and *S*-plus.

Phase 3. In this phase we did change both the percent of missingness and the precision and compared the results.

2.2. Example: Goiter disease data from health survey in Iran

We have used the data on Goiter disease collected during Health Survey in Qazvin province of Iran. The Goiter disease is one of the most prevalent diseases in Qazvin province. The data consisted of a bivariate response variable which is Goiter disease (Y) and place of residency (X) and sex (Z) as covariates which had been shown a significant relation to the disease (Noorbala and Mohammad [7]). 60% of the sample data (758 individuals) were diseased from which we drew a sample of $n = 100$ randomly due to the limitation of computer memory.

3. Results

For response variable (Y) we set $Y = 1$ for diseased (including 1A, 1B stage and above) and $Y = 0$ for non-diseased individuals, $X = 1$ for rural and $X = 0$ for urban residents, $Z = 1$ for females and $Z = 0$ for males. First by fitting standard logistic regression model (equation (9)) for full dataset all effects including interaction term were statistically significant. Then we applied our implemented *R* code to the data with 20 percent MAR in X variable and the precision $e = .0001$. The results are shown in Table 1. As can be seen the estimates of the parameters in both *S*-plus and *R* code are identical, which confirms the efficiency of the *R* code for full data. In addition for most cases the estimation of the standard errors

of the parameters in *R* code is less than those of standard logistic regression obtained from *S*-plus. After making 20% MAR in *X* covariate, the parameter estimates using Satten and Carroll likelihood of equation (14) (*R* code) and complete case analysis (*S*-plus) are shown in columns 4 and 5 of Table 1 respectively. The estimates in column 4 are much closer to the full data estimates (column 2) than the complete case estimates (column 5) without any significant difference in standard errors of estimations.

For more evaluation of our approach we did the analysis on the data with several percentages of missing and different precisions with two replicates. The results are shown in Table 2.

The maximum likelihood estimates obtained from cases with missing values are more reliable than those obtained from standard methods in which we ignore the cases with missing values.

Table 1. Parameter estimates for new and standard models using Goiter disease data of Qazvin province*

	Full data analysis		Data with 20% MAR missing in Area covariate	
parameters	New model	Standard model	New model	Standard model
β_0	–0.980852	–0.980829	–0.97996	–0.73317
	(0.49017)	(0.47859)**	(0.51021)	(0.52271)
β_1	1.791792	1.791759	1.79547	1.50339
	(0.75844)	(0.76819)	(0.69478)	(0.70279)
β_2	2.268725	2.268673	2.43209	1.81530
	(0.60241)	(0.62289)	(0.69211)	(0.68202)
β_3	–2.791966	–2.791921	–2.97879	–2.38279
	(0.96947)	(0.94599)	(0.92479)	(0.91049)

* Data is taken from "Health Survey in Iran 2000".

** The numbers in parentheses are standard errors of the estimates.

Table 2. Maximum likelihood estimates of parameters in both new and standard models. In terms of different percentage of missingness and different precisions (N_i and S_i represent i-th replication of estimates based on new and standard models respectively)

		e = 0.05				e = 0.005			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
r = 20%	N_1	–0.969 (0.084)	1.743 (0.162)	2.347 (0.127)	–2.728 (0.208)*	–1.044 (0.192)	1.831 (0.271)	2.131 (0.482)	–2.453 (0.547)
	S_1	–0.984 (0.300)	1.736 (0.340)	2.265 (0.534)	–2.700 (0.604)	–0.987 (0.132)	1.768 (0.225)	2.261 (0.334)	–2.634 (0.477)
	N_2	–0.967 (0.131)	1.699 (0.310)	2.247 (0.152)	–2.802 (0.398)	–0.891 (0.211)	1.589 (0.247)	2.280 (0.723)	–2.778 (0.825)
	S_2	–0.903 (0.448)	1.679 (0.800)	2.196 (0.518)	–2.843 (0.621)	–0.903 (0.191)	1.588 (0.257)	2.258 (0.591)	–2.706 (0.671)
r = 30%	N_1	–0.935 (0.082)	1.757 (0.207)	2.165 (0.084)	–2.612 (0.347)	–0.938 (0.271)	1.652 (0.317)	2.222 (0.472)	–2.655 (0.447)
	S_1	–0.823 (0.195)	1.663 (0.457)	2.074 (0.259)	–2.405 (0.548)	–0.822 (0.212)	1.527 (0.279)	2.030 (0.413)	–2.472 (0.520)
	N_2	–0.982 (0.249)	1.692 (0.562)	2.296 (0.114)	–2.737 (0.382)	–0.865 (0.101)	1.634 (0.271)	2.208 (0.421)	–2.708 (0.499)
	S_2	–0.883 (0.337)	1.636 (0.728)	2.107 (0.381)	–2.568 (0.855)	–0.998 (0.099)	1.738 (0.265)	2.210 (0.425)	–2.674 (0.459)

* The numbers in parentheses are standard errors of the estimates.

4. Discussion and Conclusion

In many situations where logistic regression is to be used to determine the effect of explanatory variables on a binary outcome, some of the explanatory variables are only available for a subset of study participants. In cross-sectional studies, the problem of dealing with MAR missing is of great importance because the elimination of missing-valued cases can lead to biased estimates of the parameters and misinterpretation of the results.

In this study we have demonstrated that a modification of the approach of Satten and Kupper [11] applies in this case, allowing likelihood-based inference for this type of data. This approach also has

the advantage that an unconditional likelihood can be easily constructed. As an example, we applied our approach to a cross-sectional study.

Besides the basic logistic model relating disease to covariates, the only thing that needs to be specified is a density or mass function $\pi(x \mid z)$ for the missing covariates which we used the Satten and Carroll [10] approach.

References

[1] P. Armitage and T. Colton, Encyclopedia of Biostatistics, John Wiley, New York, 1997.

[2] D. W. Blackhurst and M. D. Schluchter, Logistic regression with a partially observed covariate, Comm. Statist. Simul. 18(1) (1989), 163-177.

[3] C. Fuchs, Maximum likelihood estimation and model selection in contingency tables with missing data, J. Amer. Statist. Assoc. 77 (1982), 270-278.

[4] S. Gao and S. L. Hui, Logistic regression models with missing covariate value for complex survey data, Statistics in Medicine 16 (1997), 2419-2428.

[5] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, Second Edition, John Wiley, New York, 2002.

[6] R. J. A. Little and M. D. Schluchter, Maximum likelihood estimation for mixed continuous and categorical data with missing values, Biometrika 72 (1985), 497-512.

[7] A. A. Noorbala and K. Mohammad, National Health Survey in Iran, National Medical Research Center Publication, 2001.

[8] M. C. Paik and R. L. Sacco, Matched case-control data analyses with missing covariates, J. Roy. Statist. Soc. Ser. C 49 (2000), 146-156.

[9] P. J. Rathouz, G. A. Satten and R. J. Carroll, Semiparametric inference in matched case-control studies with missing covariate data, Biometrika 89(4) (2002), 905-916.

[10] G. A. Satten and R. J. Carroll, Conditional and unconditional categorical regression models with missing covariates, Biometrics 56 (2000), 384-388.

[11] G. A. Satten and L. Kupper, Inferences about exposure–disease associations using probability of exposure information, J. Amer. Statist. Assoc. 88 (1993a), 200-208.

[12] G. A. Satten and L. Kupper, Conditional regression analysis of the odds ratio between two binary variables when one is not measured with certainty: A method for epidemiologic studies, Biometrics 44 (1993a), 429-440.

[13] R. L. Stuart, P. Michael and E. Marium, Inference using conditional logistic regression with missing covariates, Biometrics 54 (1998), 295-303.

[14] M. R. Zali, K. Mohammad and K. Azam, Thyroid status based on health survey in Iran, J. Medical Council of Islamic Republic of Iran 13(2) (1995), 113-122.