Scientific Research and Essay Vol.4 (12), pp. 1463-1467, December, 2009
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 © 2009 Academic Journals

Full Length Research Paper

Polychotomous logistic model with missing values

M. Karimlou[1], Gh. Jandaghi[2]*, K. Azam[3], A. Grami[4] and K. Mohammad[3]

[1]Department of Computer and Biostatistics, University of Social Welfare and Rehabilitation Sciences, Iran.
[2]University of Tehran, Qom Campus, Iran.
[3]Department of Epidemiology and Biostatistics, Tehran University of Medical Sciences, Iran.
[4]University of Tehran, Sciences Campus, Iran.

Accepted 26 October, 2009

In health studies, we often face some variable missing. This missingness can happen in either response or other covariates. In this paper, the discussion focuses on missing covariates. A method is proposed for analysis of logistic regression models in which the response variable is polychotomous and some covariates' values are missing at random. The maximum likelihood function of the model is derived and the results are compared with the routine methods based on elimination of missing cases. Both the proposed method and the usual method are compared on a real dataset of goiter disease and is shown that the proposed method acts significantly better than usual method.

Key words: Missing at random, logistic regression, polychotomous response, goiter disease, likelihood function.

INTRODUCTION

Logistic regression is an analytical tool generally applied in medical and epidemiological researches (Stuart et al., 1998). In epidemiological researches, the researcher wants to calculate the odds and odds ratio for a disease. Since in logistic regression models the estimated parameters results in odds ratios, this paper intends to study a special case in logistic regression where the response variable has more than two categories and the covariates have missing values. In many medical datasets we may face some missingness in some covariates such as denying to respond, lack of information in files and incompleteness of study frame. In such cases we deal with missing values.

In this study, it is assumed that the missingness is at random and independent of observed values (MAR). For example, in analysis of effective factors on Goiter disease, the variables such as sex, age, place of residence and the iodine consumption may be of interest and due to reasons just stated, some questions are not answered and this missingness is not affected by sex, age and place of residence.

There are different approaches for analysis of such data. The simplest method is to eliminate the cases with missing values and do the analysis based on the complete cases. This approach causes the loss of information and in some cases introduces bias to the estimates (Little and Rubin, 2002). This method is implemented as the default setting in most statistical softwares such as SAS, SPSS and SPLUS (Gao and Hui, 1997). The second approach is to impute the missing values and the analysis is done. This approach has two important problems when there is a considerable number of missing values in the data. First, this kind of imputation changes the distribution of the missing-valued covariate and secondly the mean and standard error of the sample statistics is changed.

In this study, the inference is done by considering the missing values in likelihood function. The maximum likelihood estimation for both completed data and missing imputed data are the same other than that the likelihood function for missing data has some changes. Several authors have introduced their methods of dealing with missing covariates for logistic regression models. Some have used the Expectation Maximization (EM) algorithm to estimate the model parameters with discrete covariates or a combination of discrete and continuous covariates with missing values (Fuch et al., 1982; Little and Schluchter, 1985). The EM algorithm generally needs iterations. When a covariate is continuous and follows a normal distribution, the maximum likelihood method using EM algorithm does not need iterations. In a study, three

*Corresponding author. E-mail: jandaghi@ut.ac.ir.

methods of data analysis (using complete cases, imputation of missing values and maximum likelihood method when one of the two covariates has missing values) were compare with the aid of Monte Carlo method and it is concluded that the third method did better than the two other methods (Blackhurst and Schluchter, 1989).

Other researchers expanded the method to use the alternative covariates for finding information about the missing-valued covariates (Sattan and Kupper 1993a; Satten and Kupper, 1993b). There has been also some improvements in analysis of matched case-control studies when there is missing values in covariates (Paik and Sacco, 2000). Some researchers considered some distribution for missing-valued covariate and with some modifications to the likelihood function for conditional and unconditional logistic regression models, improved the estimates (Satten and Carol, 2000). In addition, a new class of estimators was established for modeling the distributions of covariates and the type of missingness (Rathouz et al., 2003). In all above studies the response variable was dichotomous. This paper introduces a method of dealing with missing values when the response variable is polychotomous.

The model

In this section, we present the logistic regression model with polychotomous response and missing values in some covariate X and show how to estimate the maximum likelihood of its parameters.

Let Y_i be a response variable taking three values 0, 1 and 2. Suppose X and Z be two fully observed covariates. In general, in saturated logistic models with polychotomous response, the conditional probabilities of response values on covariates is defined as follows: (Hosmer and Lemeshow Jr. (1999); Kleinbaum and Klein, 2002).

$$\pi_0(x)=P(Y=0|\,X=x,Z=z)=\frac{1}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)} \quad (1)$$

$$\pi_1(x)=P(Y=1|\,X=x,Z=z)=\frac{\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)} \quad (2)$$

$$\pi_2(x)=P(Y=2|\,X=x,Z=z)=\frac{\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)} \quad (3)$$

Because Y_i is considered to have three values, therefore the two odds and two odds ratios are defined as:

$$\theta_1(x,z)=\frac{P(Y=1\,|X=x,Z=z)}{P(Y=0\,|X=x,Z=z)}=\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz) \quad (4)$$

$$\theta_2(x,z)=\frac{P(Y=2\,|X=x,Z=z)}{P(Y=0\,|X=x,Z=z)}=\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz) \quad (5)$$

and

$$\psi_1(x,z,x',z')=\frac{\theta_1(x,z)}{\theta_1(x',z')} \quad (6)$$

$$\psi_2(x,z,x',z')=\frac{\theta_2(x,z)}{\theta_2(x',z')} \quad (7)$$

The purpose of any logistic regression is estimation of model parameters (here β_{10}, β_{11}, β_{12}, β_{13}, β_{20}, β_{21}, β_{22} and β_{23}) to describe the relationship between the dependent variable Y and a set of covariates (Armitage, 1997). In present paper we have considered two covariates X and Z. When these two variables are fully observed, we use standard methods to estimate the parameters. Now suppose some X values are not observed, in other words we have some missing values for X. In this case we define the indicator variable Δ_i as follows:

If the value X_i is not observed then $\Delta_i=0$ and when it is observed then $\Delta_i=1$. So, the odds and odds ratio in absence of the variable X in the model will be:

$$\tilde{\theta}_1(Z)=\frac{P(Y=1\,|Z=z)}{P(Y=0\,|Z=z)} \quad (8)$$

$$\tilde{\theta}_2(Z)=\frac{P(Y=2\,|Z=z)}{P(Y=0\,|Z=z)} \quad (9)$$

$$\tilde{\psi}_1(z,z')=\frac{\tilde{\theta}_1(z)}{\tilde{\theta}_1(z')} \quad (10)$$

$$\tilde{\psi}_2(z,z')=\frac{\tilde{\theta}_2(z)}{\tilde{\theta}_2(z')} \quad (11)$$

In addition we make the following definitions:

$$\rho_0(X\,|Z)=P(X=x\,|Y=0,Z=z) \quad (12)$$

$$\rho_1(X \mid Z) = P(X = x \mid Y = 1, Z = z) \quad (13)$$

$$\rho_2(X \mid Z) = P(X = x \mid Y = 2, Z = z) \quad (14)$$

As can be seen, the probability functions $\rho_0(X \mid Z)$, $\rho_1(X \mid Z)$ and $\rho_2(X \mid Z)$ are the probability distributions of X given $Y = 0$, $Y = 1$ and $Y = 2$ respectively. Using the Bayes theorem and formulas 12, 13 and 14, the formulas 8 and 9 will change to:

$$\tilde{\theta}_1(z) = \sum_x \theta_1(x, z).\rho_0(x \mid z) \quad (15)$$

$$\tilde{\theta}_2(z) = \sum_x \theta_2(x, z).\rho_0(x \mid z) \quad (16)$$

Where the summation is over all possible values of X and we also have

$$\rho_1(X|Z) = \frac{\theta_1(X, Z)\rho_0(X|Z)}{\sum_x \theta_1(x, Z)\rho_0(x|Z)} \quad (17)$$

$$\rho_2(X|Z) = \frac{\theta_1(X, Z)\rho_0(X|Z)}{\sum_x \theta_1(x, Z)\rho_0(x|Z)} \quad (18)$$

Now we can construct the likelihood function.

Likelihood function incorporating missing values in covariates

The likelihood function for logistic regression with polychotomous response variable Y and fully observed covariates X and Z is

$$L(\beta) = \prod_{i=1}^{n} \left\{ [\pi_0(x_i)]^{y_{0i}} [\pi_1(x_i)]^{y_{1i}} [\pi_2(x_i)]^{y_{2i}} \right\} \quad (19)$$

Where $\sum_{i=1}^{n} Y_{ji} = 1$ for every *i*(Hosmer and Lemeshow Jr, 1989).

If some covariate contains missing values, the likelihood function will become as follows(Little and Rubin, 2002):

$$P(Y, X, \Delta|Z) = P(Y|Z).\ P(\Delta \mid Y, Z)P(X \mid Y, Z, D) \quad (20)$$

Using equations (15) to (19), the likelihood function with missing values takes the following shape:

$$L(\beta) = \prod_{i=1}^{n} \left\{ [\pi_0(z_i)]^{y_{0i}} [\pi_1(z_i)]^{y_{1i}} [\pi_2(z_i)]^{y_{2i}} \right\}$$

$$[\rho_0(X_i \mid Z_i)]^{\Delta_i y_{0i}} [\rho_1(X_i \mid Z_i)]^{\Delta_i y_{1i}} [\rho_2(X_i \mid Z_i)]^{\Delta_i y_{2i}} \quad (21)$$

By noticing that $\sum_{i=1}^{n} Y_{ji} = 1$ and from equations (15) to (18) the likelihood function becomes:

$$L(\beta) = \prod_{i=1}^{n} \left\{ [P_0(X_i \mid Z_i)]^{\Delta_i y_{0i}} [P_1(X_i \mid Z_i)]^{\Delta_i y_{1i}} \left[\sum_x \rho_0(x_i|z_i)\theta_1(x_i, z_i) \right]^{(1-\Delta_i) y_{1i}} \times [\rho_0(x_i|z_i)\theta_1(x_i, z_i)]^{\Delta_i y_{2i}} \left[\sum_x \rho_0(x_i|z_i)\theta_1(x, z_i) \right]^{(1-\Delta_i) y_{2i}} \right\} \Big/ \left\{ 1 + \sum_x \rho_0(x_i|z_i)\theta_1(x, z_i) + \sum_x \rho_0(x_i|z_i)\theta_2(x, z_i) \right\} \quad (22)$$

On the other side the distribution $\rho_0(X \mid Z)$ is unknown. When X and Z have limited number of values, we can consider a distribution for $\rho_0(X \mid Z)$ from the exponential family like

$$\rho_0(x|z) = \frac{e^{\gamma_{xz}}}{\sum_{x'} e^{\gamma_{x'z}}} = \frac{e^{\gamma_0 + \gamma_1 x + \gamma_2 z + \gamma_3 xz}}{\sum_{x'} e^{\gamma_0 + \gamma_1 x + \gamma_2 z + \gamma_3 xz}} = \frac{e^{\gamma_1 x + \gamma_3 x z}}{\sum_{x'} e^{\gamma_1 x' + \gamma_3 x' z}} \quad (23)$$

Which gives interesting results (Satten and Carrol, 2000). With the aid of equations (15) to (18) and (23) and rewriting (22) we will have a function of parameters β_{10}, β_{11}, β_{12}, β_{13}, β_{20}, β_{21}, β_{22}, β_{23}, γ_1 and γ_3.

After taking logarithm of likelihood function and taking its first partial derivatives with respect to each parameter to solve the, we have a system of 10 equations with 10 unknowns. Because of nonlinearity of the equations, we need some numerical method to estimate the parameters. In the following example we show how this approach works.

Example: The data of this example has taken from the National Health Survey in Iran (NHS) in 2001(Noorbala AA, Mohammad K 2001)

The data of this example has taken from the National Health Survey in Iran (NHS) in 2002. In this study the data on thyroid diseases shows that Qazvin province has the first rank with a prevalence rate of 11.4%. The pro-

Table 1. Comparison of MLE estimates for logistic model parameters based on the proposed and standard methods in both complete data and data with 35% missing in X varible (data has taken from tiroid situation in $Qazvin$ province).

			Full data		Data with 35% missing in "Area"	
Logit	**Variables**	**Parameters**	**New model**	**Standard model**	**New model**	**Standard model**
1	Intercept	β_{10}	-0.993 (0.368)	-0.993 (0.370)	-1.035 (0.386)	-0.693 (0.423)
	Sex (Z)	β_{11}	1.463 (0.617)	1.463 (0.670)	1.502 (0.658)	1.019 (0.777)
	Area(X)	β_{12}	1.190 (0.542)	1.194 (0.582)	1.278 (0.565)	1.281 (0.706)
	Area*Sex(X*Z)	β_{13}	-2.263 (0.835)	-2.270 (0.960)	-2.409 (0.911)	-2.308 (1.136)
2	Intercept	β_{20}	-1.910 (0.534)	-1.910 (0.536)	-1.810 (0.582)	-1.386 (0.559)
	Sex (Z)	β_{21}	2.497 (0.719)	2.497 (0.773)	2.90 (0.783)	1.792 (0.854)
	Area(X)	β_{22}	2.106 (0666)	2.110 (0699)	2.067 (0.719)	1.856 (0.798)
	Area*Sex(X*Z)	β_{23}	-2.891 (0881)	-2.899 (1.001)	-2.813 (0.981)	-2.549 (1.160)

i. Data from Iran National Health Study.

ii. Logit 1 and 2 means $\ln \theta_1(x, z)$ and $\ln \theta_2(x, z)$ respectively

iii. The numbers in parantheses are standard errors.

vinces Kurdestan and Yazd have second and third places with prevalence rates 10.8 and 9.6% respectively. The prevalence rate in rural area of Qazvin is 6.7 while in its countryside is around 17.8%. Men with 9.9% have lower rate than women with 12.6% (Noorbala and Mohammad, 2003). In previous NHS study whose results was published in 1992, the percent of observable thyroid information in Qazvin showed high value of prevalence (Zali et al., 1992). This prevalence was chosen for its missingness in some variables to see the performance of the likelihood function discussed in this paper. After analyzing the data by logistic regression model, it turned out that the variables sex and place of residence showed a significant relationship with the response variable, thyroid disease which had three categories (healthy, 1A, 1B and higher) (Zali et al., 1995). In this province, the studied samples were 758 people from whom 60% had thyroid disease (Noorbala and Mohammad, 2002). Here the value Y = 0, Y = 1 and Y = 2 represent healthy, 1A and 1B respectively. Although this variable is an ordinal variable, we consider Y as a nominal polychotomous variable. Variables sex (Z) and place of residence(X) were fully observed. We took a random sample of 120 and made some values missing and then used the likelihood function to estimate the parameters. The results are shown in next section.

RESULTS

In this section, to reach the study goals, we did several stages of analysis. The purpose was only to evaluate the new likelihood and our new program. So, the significant relationship between covariates and the response variable was not of our priority.

As stated, the response variable had three categories, Y = 0 (healthy), Y = 1(1A) and Y = 2(1B) and two covariates sex (Z = 0 for male and Z = 1 for female) and place of residence(X = 1 for rural and X = 0 for countryside). First a logistic model was fitted to the data and both variables remained in the model at level $\alpha = 0.05$. The results are shown in Table 1. The numbers in the Table 1 are maximum likelihood estimation of the model para-

meters and their standard errors. Columns 1 and 2 show the estimates for full data. Comparison of standard estimates given by SPSS software with those of our new Splus program shows the same values for model parameters. After elimination of 35% of the variable X (missing at random) and repeating the analysis by both SPSS and our S-plus program, the outputs which are placed in columns 3 and 4 in Table 1 shows the estimates from our program is closer to the full data estimates than those of SPSS software. To evaluate our approach more carefully, we repeated the analysis 10 times with 20, 25, 30 and 35% of missingness in variable X. To confirm our results we did a two-way analysis of variance to see the effects of percent of missingness and type of the model on estimates. The analysis of variance showed that only the model type was significant at $\alpha = 0.001$ meaning that the estimates are significantly different for two types of models and the estimates of our new model are closer to those of full data model.

DISCUSSION AND CONCLUSION

Satten and Carrol compared the parameter estimates using this approach for a binary response variable and concluded that this approach is more efficient than those methods that ignore the missing cases. As can be seen from Table 1, using the new approach for data in which the response variable has three categories, resulted in more accurate estimates with lower variances. Additionally, A Kruskal-Wallis test showed that the variances of the estimates in new approach were significantly different from the standard model.

REFERENCES

Armitage P, Colton T (1997). Encyclopedia of biostatistics , John Wiley ,New York.

Blackhurst DW, Schluchter MD (1989). Logistic regression with a partially observed covariate , Comm. Statist. Simul, 18(1):163-177.

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

Gao S, Hui SL(1997). Logistic regression models with missing covariate value for complex survey data , Stat. Med. 16: 2419-2428.

Hosmer DW, Lemeshow Jr. S (1989). Applied logistic regression, John Willey & Sons

Kleinbaum DG, Klein M (2002). Logistic Regression A Self – Learning Text ,Second Edition Springer.

Little RJA, Rubin DB (2002). Statistical analysis with missing data , John Wiley & Sons, Second Edition, New York.

Little RJA, Schluchter MD (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values , Biometrika, 72: 497- 512.

Noorbala AA, Mohammad K (2001). Health Survey in Iran, National Center of Medical Researches.

Noorbala AA, Mohammad K (2002). Health Survey in Iran, National Center of Medical Researches.

Paik MC, Sacco RL (2000). Matched case – control data analyses with missing covariates, Appl. Stat. 49, 146-156.

Rathouz PJ, Satten GA, Carrol RJ(2003). Semiparametric inference in matched case – control studies with missing covariate data , Biometrika .

Satten GA, Kupper L (1993a). Inferences about exposure – disease associations using probability of exposure information, J. Amer. Statist. Assoc , 88: 200-208.

Satten GA, Carroll RJ (2000). Conditional and unconditional categorical regression models with missing covariates , Biometrics, 56: 384-388.

Satten GA, Kupper L (1993b). 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: 429-440.

Stuart RL, Michael P, Marium E (1998). Inference using conditional logistic regression with missing covariates, Biometrics, 54: 295-303.