Generalized Linear Model for Binary Data with Missing Values: An EM Algorithm Approach

Nazneen Sultana

Abstract: A procedure is derived for estimating the parameter in case of missing data. The missing data mechanism is considered as missing at random (MAR) and non-ignorable. Here we use EM algorithm for logit link approach in generalized linear model. The logit link approach shows that it can effectively estimate the value of a categorical variable when we have information on the other categorical variables. In this method the variable with missing values is considered as dependent variable. In addition a real data set for low birth weight is presented to illustrate the method proposed.

Index Terms: Binary data, EM algorithm, Generalized linear model, Logit link, Maximum likelihood estimation, Missing data, non-ignorable.

1. INTRODUCTION

DURING the process of complete data, sometimes we may not get the full observed data. This results in partially incomplete data. An inappropriate conclusion may occur when the researchers ignore, truncate, censor or collapse those data as it might contain important information. When non-response is unrelated to the missing values of the variables, the non-response is called ignorable (see Little (1992) and Little and Rubin (1987)). When non-response is related to values of missing variables, the non-response is called non-ignorable. The literature for generalized linear model with incomplete observations, however, is sparse. Ibrahim et al (1990) discussed incomplete data in generalized linear models. Ibrahim and Lipsitz (1996) proposed a method for estimating parameters in binomial regression models when the response variable is missing and the missing data mechanism is non-ignorable. Ibrahim and Lipsitz (1996) proposed a conditional model for incomplete covariates in parametric regression models. Ibrahim, Lipsitz and Chen (1999) proposed a method for estimating parameters in generalized linear models with missing covariates and a non-ignorable missing data mechanism.

In this paper, we have proposed a method for estimating parameters in generalized linear model with missing values. We used parameter estimation procedure for logit link function in generalized linear model. Here variable with missing values is considered as dependent variable. Here we considered the non-response as non-ignorable. We used EM algorithm both for categorical and continuous variable. The method proposed is computationally simple and easy.

The rest of this paper is organized as follows. In section 2, we first briefly review the previous

- *Nazneen Sultana,Lecturer, Department of Applied Statistics, East West University, Bangladesh, PH-01754756427. E-mail:nazsultana87@gmail.com*

methods of estimating parameters for missing data using EM algorithm. In section 3, we discuss the parameter estimation procedure for logit link function in generalized linear model. Then in section 4, we discuss the proposed method and in section 5, we demonstrate the methodology with example. We conclude the paper with a discussion section.

2. EM Algorithm for Missing Data

In 1977 a broadly applicable algorithm for computing the maximum likelihood estimates from incomplete data is proposed by Dempster, Laird and Rubin. They proposed an algorithm which is named as EM algorithm because it involves expectation step (E-step) and maximization step (M-step) in each iteration.

Little and Scheluchter (1985) discussed the maximum likelihood estimation procedure for mixed continuous and categorical data with missing values. The general location model of Olkin & Tate (1961) and extensions introduced by Krzanowski (1980, 1982) formed the basis for this method.

Baker and Laird (1988) developed the process of regression analysis for categorical variables with outcome subject to non-ignorable non-response. They developed a log-linear model for categorical response subject to non-ignorable non-response. To illustrate model development, they considered X as cross-classification of covariates indexed by $x = 1, \dots\dots, s, Y$ as polychotomous outcome indexed by $y = 1, \dots\dots, q,$ and R as a dichotomous response mechanism indexed by r ($r = 1 = response; r = 2 = no\ response$. Let $p_{yr|x}$ be the joint probability of outcome and response mechanism conditional on x. Let z_{xy1} be the observed counts for the completely classified data of the respondents, and let z_{x+2} denote the observed counts for the incompletely classified data of the nonrespondents. The likelihood function corresponding to their model is given by

$$L = \left[\prod_x \prod_y \left(p_{y1|x}\right)^{z_{xy1}}\right]\left[\prod_x \left(p_{+2|x}\right)^{z_{x+2}}\right]$$

Then they proceeded the E-step and M-step of EM algorithm. But this process takes a large number of iteration.

Lipsitz and Ibrahim (1996) examined that when the missing covariates are categorical, a useful technique for obtaining parameter estimates is the EM algorithm by the method of weighs proposed in Ibrahim (1990). This method requires the estimation of many nuisance parameters for the distribution of the covariates. In this paper, the distribution of K-dimensional covariate vector $x = (x_1, \dots, x_k)$ can be written through a series of one-dimensional conditional distributions, as

$$P\left(x_1,\dots, x_k|\alpha|\right) = P\left(x_k|x_1,\dots, x_{k-1}, \alpha_k\right)$$
$$P\left(x_{k-1}|x_1,\dots, x_{k-2}, \alpha_{k-1}\right)\dots P\left(x_2|x_1, \alpha_1\right)P\left(x_1|\alpha_1\right)$$

where α_k is a vector of indexing parameters for the kth conditional distribution, $\alpha = (\alpha_1, \dots, \alpha_k)$, and the α_k's are distinct. They used the EM algorithm by the method of weights where the weights are the posterior probability of the missing values. Unfortunately there are often too many probabilities to estimate in a saturated model for $P(x_1, \dots, x_k|\alpha)$ when there are many covariates with missing values. In that situation, the model becomes complicated.

Ibrahim, Lipsitz and Chen (1999) developed a method for missing covariates in Generalized Linear Model when the missing data mechanism is non-ignorable. They used a multinomial model

for the missing data indicators and proposed a joint distribution for them which can be written as a sequence of one-dimensional conditional distributions, with each one-dimensional conditional distribution consisting of a logistic regression. They allowed the covariates to be either categorical or continuous. Suppose that $(x_1, y_1), \dots, (x_n, y_n)$ are independent observations, where each y_i is the response variable and each x_i is a $p \times 1$ random vector of covariates. The missing data mechanism is defined as the distribution of the $p \times 1$ random vector r_i, whose kth component r_{ik} equals 1 if x_{ik} is observed for subject i and is 0 if x_{ik} is missing. Here $\beta = (\beta_1, \dots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, α and φ are considered as nuisance parameters. In this paper for categorical covariates,

$$p(y_i, x_i, r_i|\beta, \alpha, \varphi) = p(y_i|x_i, \beta)p(x_i|\alpha)p(r_i|y_i, x_i, \varphi)$$

which leads to the complete data log-likelihood

$$l(\gamma) = \sum_{i=1}^{n} l(\gamma, x_i, y_i, r_i)$$

$$= \sum_{i=1}^{n} \log\{p(y_i|x_i, \beta)\} + \log\{p(x_i|\alpha)\}$$

$$+ \log\{p(r_i|y_i, x_i, \varphi)\}$$

where $\gamma = (\beta, \alpha, \varphi)$ and $l(\gamma, x_i, y_i, r_i)$ is the contribution to the complete data log-likelihood for the ith observation.

Then they used the EM algorithm by the method of weights where the E-step is,

$$Q(\gamma|\gamma^{(t)})$$

$$= \sum_{i=1}^{n} \sum_{x_{mis,(j)}} w_{ij,(t)} \log[p\{y_i|x_i(j), \beta\}]$$

$$+ \sum_{i=1}^{n} \sum_{x_{mis,(j)}} w_{ij,(t)} \log[p\{x_i(j)|\alpha\}]$$

$$+ \sum_{i=1}^{n} \sum_{x_{mis,(j)}} w_{ij,(t)} \log[p\{r_i|y_i, x_i(j), \varphi\}]$$

$$= Q^{(1)}(\beta|\gamma^{(t)}) + Q^{(2)}(\alpha|\gamma^{(t)}) + Q^{(3)}(\varphi|\gamma^{(t)})$$

The weights $w_{ij,(t)}$ are the conditional probabilities corresponding to $[x_{mis,i} \,|\, x_{obs,i}, y_i, r_i, \gamma]$ and are given by

$$w_{ij,(t)}$$
$$= p\{y_i \,|\, x_{mis,i}(j), x_{obs,i}, \gamma^{(t)}\}.p\{x_{mis,i}(j), x_{obs,i} \,|\, \gamma^{(t)}\}$$
$$.p\{r_i \,|\, y_i, x_{mis,i}(j), x_{obs,i}, \gamma^{(t)}\}$$
$$\div \sum_{x_{mis,i}(j)} p\{y_i \,|\, x_i(j), \gamma^{(t)}\}.p\{x_i(j) \,|\, \gamma^{(t)}\}$$
$$.p\{r_i \,|\, y_i, x_i(j), \gamma^{(t)}\}$$

and the M-step is,

$$Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^{n} Q_i(\gamma|\gamma^{(t)})$$

$$= \sum_{i=1}^{n} \sum_{x_{mis,(j)}} w_{ij,(t)} \frac{\partial l\{\gamma_i, x_i(j), y_i, r_i\}}{\partial \gamma}$$

But this procedure is so much complicated and time consuming. That is why, an easy and simpler method is proposed in this paper.

3. Parameter Estimation: Logit Link Function in the Generalized Linear Model

For any binomial variable, Y, the probability mass function can be expressed as (McCullagh, P. and J.A. Nelder. 1989)

$$f_Y(y; \theta, \varphi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

$$= exp\left[yln\left(\frac{\pi}{1-\pi}\right) + nln(1-\pi) + ln\binom{n}{y}\right] \quad (3.1)$$

Therefore, from (3.1), for the binomial distribution

$$\theta = ln\left[\frac{\pi}{1-\pi}\right], \qquad \pi = \frac{e^{\theta}}{1+e^{\theta}},$$

$$b(\theta) = -nln(1-\pi),$$

$$a(\varphi) = 1, \qquad c(y,\varphi) = ln\binom{n}{y},$$

$$E(y) = \frac{db(\theta)}{d\theta} = \frac{db(\theta)}{d\pi}.\frac{d\pi}{d\theta}$$

where

$$\frac{d\pi}{d\theta} = \frac{e^{\theta}}{1+e^{\theta}} - \left[\frac{e^{\theta}}{1+e^{\theta}}\right]^2 = \pi(1-\pi) \quad (3.2)$$

For the exponential family, the log likelihood function corresponding to a random sample of size n is

$$l(y,\beta) = \sum_{i=1}^{n}\left[\frac{\{y_i\theta_i - b(\theta_i)\}}{a(\varphi)} + c(y_i,\varphi)\right]$$

Thus for the canonical link in the binomial case, we have

$$\eta_i = g[E(y_i)] = g(\mu_i)$$

$$= ln[\mu_i/(1-\mu_i)] = x_i\acute{}\beta = \theta_i$$

and $\mu_i = \pi_i$ where $x_i\acute{}$ is the i-th row of the X-matrix. Therefore,

$$\frac{\delta l}{\delta\beta} = \frac{\delta l}{\delta\theta_i}.\frac{\delta\theta_i}{\delta\beta}$$

$$= \frac{1}{a(\varphi)}\sum_{i=1}^{n}\left[y_i - \frac{db(\theta_i)}{d\theta_i}\right]x_i$$

$$= \frac{1}{a(\varphi)}\sum_{i-1}^{n}[y_i - \mu_i]x_i$$

Consequently, we can find the maximum likelihood estimates of the parameters by solving the system of equations

$$\frac{1}{a(\varphi)}\sum_{i-1}^{n}[y_i - \mu_i]x_i = 0 \quad (3.3)$$

For the binomial distribution $a(\varphi) = 1$, so (3.3) becomes

$$\sum_{i-1}^{n}[y_i - \mu_i]x_i = 0 \quad (3.4)$$

Thus, the maximum likelihood estimate of β is

$$\hat{\beta} = (X\acute{}V^{-1}X)^{-1}X\acute{}V^{-1}z \quad (3.5)$$

It is interesting to note similarity of (3.5) to the expression obtained in standard regression model.

4. EM Algorithm with Logit Link in GLM

In this method first we have to check out in which variable the missing values arise. Then the variable with missing data is considered as dependent variable. For example, let we have 3 variables x_{i1}, x_{i2} and x_{i3}. If missing values are arise in x_{i2} then we consider it as dependent variable and we have to calculate the conditional probabilities of x_{i2} with respect to x_{i1} and x_{i3}.

That is, we have to fine $p(x_{i2} \mid x_{i1}, x_{i3})$. After that we use the iteration procedure of EM algorithm.

Suppose that $(x_1, y_1), \ldots\ldots, (x_n, y_n)$ are independent observations. If there are missing values in y_i then y_i is considered as response variable and each x_i is a $p \times 1$ random vector of covariates. The conditional distribution of y_i given x_i is

$$p(y_i \mid x_i, \beta) = \pi^{y_i}(1-\pi)^{1-y_i}$$

where

$$\pi = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

The E-step is

$$E(y \mid x) = \hat{\pi} = \frac{exp(X\beta)}{1+exp(X\beta)} \qquad (4.1)$$

Here

$$p(y_i = 1 \mid x_i, \beta) = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

and

$$p(y_i = 0 \mid x_i, \beta) = \frac{1}{1 + exp(X\beta)}$$

From incomplete data, we calculate $E(y \mid x) = \hat{\pi}$. If $\hat{\pi} \geq 0.5$ then in missing values, we consider y = 1 and if $\hat{\pi} < 0.5$ then

y = 0. After that we get the complete data. Now the complete data likelihood function is

$$l(\beta) = \prod_{i=1}^{n} \left(\frac{exp(X\beta)}{1 + exp(X\beta)}\right)^{y_i} \cdot \left(\frac{1}{1 + exp(X\beta)}\right)^{1-y_i}$$

The log-likelihood function is

$$logl(\beta) = \sum_{i=1}^{n} [y_i X\beta - log(1 + \exp(X\beta))]$$

The M-step involves the maximization of the log-likelihood. Thus the M-step can be obtained as follows:

$$\frac{\delta logl(\beta)}{\delta \beta_j} = \sum_{i=1}^{n} \left[y_i x_j - \frac{x_j exp(X\beta)}{1 + exp(X\beta)}\right]$$

and

$$\frac{\delta^2 logl(\beta)}{\delta \beta_j \delta \beta_k} = -\sum_{i=1}^{n} \left[\frac{x_j x_k exp(X\beta)}{\left(1 + exp(X\beta)\right)^2}\right]$$

So the score vector

$$U(\beta) = \frac{\delta logl(\beta)}{\delta \beta_j} \qquad (4.2)$$

And the information matrix is

$$I(\beta) = \left[-\frac{\delta^2 logl(\beta)}{\delta \beta_j \delta \beta_k}\right] \qquad (4.3)$$

Then we get the estimate of β using Newton-Raphson method:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I\left(\hat{\beta}\right)^{-1} U\left(\hat{\beta}\right) \qquad (4.4)$$

We have to continue this repeatedly until the convergence that is

$$\left|\beta^{(t+1)} - \beta^{(t)}\right| \leq \varepsilon$$

5. Example

We consider the Low Birth Weight Data (Hosmer and Lemeshow). Consider two variables Low (low birth weight of baby) and Age (mother's age). Here we consider missing at random method. In variable Low, 20, 74, 94, 22 and 93 positions are missing. That is we have 5 missing values. So we consider Low as dependent variable and Age as independent variable. We considered age 22 as a cut-off point. Since we want to work with dichotomous data, we recode the age data into 0 and 1. Here

$$p(y_i = 1 \mid x_i, \beta) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

where $X^{'} = (1, x_i)$ is the 1×2 vector of covariates for the ith observation, including an intercept and $\beta = (\beta_0, \beta_1)'$. The initial values chosen for the regression coefficients are $(\beta_0, \beta_1) = (-0.6, -0.07)$.

Then using (4.1), we get $\hat{\pi}$. Using the condition of replacing missing values, we replace the 5 missing values by 1. Then we get the complete data. In M-step we use (4.2), (4.3) and (4.4) and then we get the estimate.

Again using this estimate we recalculate E-step and M-step. After the EM convergence we get the final estimate.

Table 5.1: Parameter estimation for categorical age

Variable	Estimated Coefficient	Standard Error	Z value	Pr(>|z|)
Intercept	-0.74194	0.22182	-3.345	0.000824
Age	-0.09546	0.31403	-0.304	0.761141

It is evident from Table 5.1 that there is no statistically significant association between age and birth weight. However we observed that all randomly chosen values for subject 20, 74, 94, 22 and 93 are matched with the estimated values. The criteria of matching are discussed in the previous section 3.

Again we considered age as continuous variable and did the whole procedure without recoding the age. Here we considered low as binary data and age as continuous data. Using these variables, we get the result for logit link approach.

Table 5.2: Parameter estimation for continuous age

Variable	Estimated Coefficient	Standard Error	Z value	Pr(>|z|)
Intercept	0.38458	0.73212	0.525	0.599
Age	-0.05115	0.03151	-1.623	0.105

From Table 5.2 we observed that we get the same result after considering age as continuous variable. That is, there is no statistically significant relationship between age and low birth weight. However we observed that all randomly chosen values are also matched with the estimated values. The percentage of matching may improve with the specification of the underlying model.

6. Discussion

We have proposed a method of estimation based on GLM. We have proposed the method for non-ignorable non-response. For the examples considered in Section 5, we observed that the missing values are fully matched with the estimated values and the estimates are matched with the estimates for complete data i.e. original data. One drawback of the EM algorithm is its slow convergence rate but the whole procedure is simple and easy. This method is easy to handle and it gives efficient result. So we can use it for real-life data.

Acknowledgments

I am grateful to Professor Dr. M. Ataharul Islam for his continuing support and encouragement. I would also like to thank the referees for their helpful comments.

References

[1] Agresti, A. (1990), "Categorical Data Analysis".*New York: Wiley.*

[2] Hosmer, D. W. and Lemeshow, S. (2000) "Applied Logistic Regression" *2nd edn. Wiley*.

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

[4] Baker, S. G. and Laird, N. M. (1988), "Regression analysis for categorical variables with outcome subject to non-ignorable non-response". *J. Am. Statist. Ass.*, **83**, 62-69.

[5] Ibrahim, J. G. (1990), "Incomplete data in generalized linear models". *J. Am. Statist. Ass.*, **85**, 765-769.

[6] Ibrahim, J. G. and Lipsitz, S. R. (1996), "Parameter estimation from incomplete data in binomial regression when the missing data mechanism is non-ignorable". *Biometrics*, **52**, 1071-1078.

[7] Lipsitz, S. R. and Ibrahim, J. G. (1996), "A conditional model for incomplete covariates in parametric regression models". *Biometrika*, **83**, 916-922.

[8] Little, R. J. A. and Rubin, D. B. (1987), "Statistical Analysis with Missing Data". *New York: Wiley*.

[9] McCullagh, P. and Nelder, J. (1989), "Generalized Linear Models", *2nd edn. New York: Chapman and Hall*.

[10] Ibrahim, J. G., Lipsitz, S. R. and Ming-Hui Chen (1999), "Missing covariates in Generalized Linear Models when the missing data mechanism is non-ignorable". *J. R. Statist. Soc.* **61**, 173-190.