Sampling estimating equation for logistic regression with multi-dimensional missing binary covariates: application to a divorce data

Renwen Luo1, Yihe Yang, Bing Guo, Yingting Luo and Jianxin Pan

College of Mathematics, Sichuan University, Chengdu, 610063, China

1Corresponding author’s e-mail: rwluo@stu.scu.edu.cn

Abstract. This paper focuses on analysing a divorce data collected in Sichuan Province, China. The responses of this divorce data are binary variables. Among 17 covariates, there are 12 binary variables with missing values. Thus, the vectors of covariates with missing values are multi-dimensional and have binary components. Existing approaches usually focus on one-dimensional binary covariates and therefore cannot deal with this divorce data. To figure out multi-dimensional binary covariates, a sampling estimating equation based on the logistic regression and sampling the missing binary covariates are proposed. Statistical inference based on this sampling estimating equation is presented. This paper has two main contributions. First, the proposed sampling estimating equation is able to obtain reasonable parameter estimators under multi-dimensional missing binary covariates setting, which cannot be handled by existing approaches. Second, for two-category problem, subjects with multi-dimensional missing binary covariates can be classified efficiently, which is intractable for existing approaches. Real data analysis of the divorce data reveals profound results for the divorce cases in Sichuan Province. The proposed sampling estimating equation performs well in terms of classification and prediction on this divorce data.

1. Introduction

Missing data may occur in real application for various reasons. Sometimes data is difficult to observe due to physical location or machine restriction. Sometimes they are missing due to mistakes made by data collectors. Missing values often lead to biased and unreasonable statistical inference (Little and Rubin [1]). Thus it is necessary to study the underlying mechanism of missing data. Little and Rubin [1] proposed a conception called “missing mechanism”. The three main mechanisms are missing completely at random, missing at random, not missing at random. Among these missing mechanisms, missing completely at random is what this paper considers. We analyze a divorce data consisting of 231 legal divorce cases in Sichuan Province, China. Each case includes 1 binary 0-1 coded response variable, the result of judgement: divorce (1) or not divorce (0); and 17 corresponding covariates, in which there are 5 continuous variables and 12 binary 0-1 coded variables. Binary covariates of the last 116 cases are subject to missing values due to the mistakes made by data collectors. Therefore, the missing mechanism is missing completely at random and this data is subject to multi-dimensional binary covariates. Although the high rate of missing values in binary covariates, all continuous covariates are precisely observed so that it is not appropriate to ignore the information in divorce cases with missing values and approach to deal with multi-dimensional binary covariates should be developed. The interests in this divorce data are due to the fact that the result of a divorce judgement
seems to depend on variable factors (e.g. 17 corresponding covariates). Analysing this divorce data may provide constructive and scientific suggestions for judges to help them to make more justified divorce judgement.

In the literature, existing methods often deal with continuous missing data, attention to binary missing data is relatively rare. Chen and Little [2] considered missing binary covariate in Cox proportional model, Heinze [3], Scheppner et al. [4] and White and Royston [5] used imputation to deal with missing binary covariate and Hsu et al. [6] considered logistic regression with missing binary covariates. However, these works only focuses on 1-dimensional binary covariates, and therefore they are not appropriate to deal with the aforementioned divorce data, where the missing components in 12-dimensional binary covariates should be addressed. K-nearest neighbor imputation (KNNI) is considered by Jonsson and Wohlin [7] and KNNI can be applied to the divorce data by imputing missing binary covariates. However, as pointed out by Vach and Martin [8], imputation methods is based on implicit modelling of the conditional distributions for the incompletely observed subjects, which may lead to efficiency loss.

To overcome the drawbacks of aforementioned works, we propose a new estimating equation called sampling estimating equation based on the logistic model (Kleinbaum and Klein [8]) and sampling the missing binary variables. The sample estimating equation can be easily extended to the generalized linear model (Nelder and Wedderburn [9]) setting, though in this paper we consider the logistic model setting. This is due to the responses in the divorce data are binary variables. The reason why we choose the logistic model is that we find the logistic model performs well on the first 115 cases of the divorce data, where there is no missing value. Instead of imputing missing binary covariates, the sampling estimating equations sample the binary variables. Compared to existing methods focusing on 1-dimensional missing binary covariate, sampling estimating equation has two main advantages. First, reasonable parameter estimator can be obtained under missing multi-dimensional binary covariates setting, which cannot be handled by 1-dimensional methods. Second, sampling estimating equation can be applied to two-category classification (fitting) for subjects with multi-dimensional missing covariates, which is intractable for 1-dimensional approaches. Compared to KNNI, sampling estimating equation is approximation to an unbiased estimating equation so that the resulting estimator is expected to more efficient than estimator obtained by KNNI, which is confirmed by divorce data analysis in section 4. We also propose new statistical inference procedure based on this sampling estimating equation. For illustration, the divorce data is analyzed to reveal the underlying legal basis for divorce judgements of Sichuan Province.

This paper is organized as follows, in Section 2, we present mathematical notation and model setting. Section 3 focuses on methodology in terms of the construction of the sampling estimating equation and corresponding algorithm. Real data analysis of the divorce data are given in Section 4. This paper is ended by main findings and future interest summarized in Section 5.

2. Notation and model setting

2.1. Response model

For subject $i$, let $y_i$ be the $i$th binary response variable. Let $x_i$ be the $p_1 \times 1$ vector of covariates with intercept and continuous components, $b_i$ be the $p_2 \times 1$ vector of binary covariates, where binary means each component of $b_i$ is binary scalar. Let $b_{ij}$ be the $j$th components of $b_i$, we assume $b_{ij}$ is independent of $b_{ik}$ for $j < k$.

We are interested in the relationship between the response variable $y_i$ and $p \times 1$ covariate $z_i = (x_i', b_i')'$, where $p = p_1 + p_2$. In particular, we assume the response $y_i$ are linked to $z_i$ through the logistic function (Kleinbaum and Klein [4])

$$\log \left( \frac{\mu_i}{1-\mu_i} \right) = z_i' \beta,$$

where $\mu_i = \mathbb{E}(y_i | z_i)$ is the mean of $y_i$, and $\beta$ is $p \times 1$ vector of the parameters to be estimated. It is noted that model (1) is a special case of generalized linear model (GLM) (Nelder and Wedderburn [9]).
and the construction of the sampling estimating equation in Section 3 can be easily extended to GLM setting. It is noteworthy that we choose the logistic model with linear predictors because we find that model (1) performs well on the former 115 divorce cases, which is not subject to missing values.

2.2. Multi-dimensional binary missing covariates

Denote $b^*_ij$ by binary missing component occurring at the $j$th component of $b_i$ for $i = n_1 + 1, \ldots, n$ and $1 \leq j \leq p_2$, where $1 < n_1 < n$. Note for $b^*_i$, there may be at most $p_2$ missing components. Suppose all the components of $x_i$ can be observed for all $n$ subjects. Therefore, we completely observe $x_i$ for $i = 1, \ldots, n$ and $b_i$ for $i = 1, \ldots, n_1$.

For covariates with measurement errors or misclassification, statistical models for measurement errors and misclassification often inspired by the information from completely observed covariates, see Yi el al. [10] for an example. Although such insight is proposed in the field of modelling multi-dimensional binary covariates. In spirit of Yi et al. [11], we propose to model the binary covariates, in which the missing values occur, based on the information from completely observed covariates $x_i$. Specifically, we study the relationship between $x_i$ and $b_i$ based on the first $n_1$ subjects, of which the covariates are completely observed. Suppose $\eta_{ij} = E(b_{ij}|x_i)$ is the mean of $b_{ij}$ given $x_i$, we assume $\eta_{ij}$ and $x_i$ are linked through the logistic function

$$\log\left(\frac{\eta_{ij}}{1-\eta_{ij}}\right) = x_i'y_j$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p_2$, where $y_j$ is a $p_1 \times 1$ vector to be estimated.

It is noted that model (2) implies that $y_j$ can be estimated using the logistic regression of binary variables $b^*_ij$ on $x_i$ for $i = 1, \ldots, n_1$. After that, consistent estimator of $\eta_{ij}$ can be obtained based on consistent estimator of $y_j$. Specifically, denote $\hat{\eta}_{ij}$ and $\hat{y}_j$ by the estimators of $\eta_{ij}$ and $y_j$, respectively. Then $\hat{\eta}_{ij} = \exp(\hat{y}_j)/(1 + \exp(\hat{y}_j))$ is consistent provided that $\hat{y}_j$ is consistent. With such consistent $\hat{\eta}_{ij}$, we can sample missing $b^*_ij$ for $i = n_1 + 1, \ldots, n$, $1 \leq j \leq p_2$, and then construct consistent estimating equation of $\beta$.

3. Sampling estimating equation

Denote $B^*$ be a vector consisting of all the missing $b^*_ij$, let $Y = (y_1, \ldots, y_n)'$ and $Z = (z_1, \ldots, z_n)'$. Consider the score function induced by model (1),

$$S(\beta, B^*) = \sum_{i=1}^n z_i(y_i - \mu_i(z_i'\beta)),$$

where components of $B^*$ occurs in $z_i$ for $i = n_1 + 1, \ldots, n$. The sampling estimating equation is denoted by

$$\frac{1}{R} \sum_{r=1}^R S(\beta, B^*) = 0,$$

where $B^{(1)}, \ldots, B^{(R)}$ are $R$ independent and identically distributed samples generated from the distribution of $B^*$ given $X = (x_1, \ldots, x_n)'$.

The left side of equation (4) is actually a numerical approximation to a consistent estimating equation $E_{B^*}(S(\beta, B^*)|Y)$. Note that we have

$$E_{Y|Z}(E_{B^*}(S(\beta, B^*)|Y)) = E_{Y|Z,B^*}(S(\beta, B^*)) = E_{B^*}(E_{Y|Z}(S(\beta, B^*)|B^*)) = 0,$$

since $E_{Y|Z}(S(\beta, B^*)|B^*) = E_Y(S(\beta, B^*)|Z) = 0$, where $E_Y$ means taking expectation on $Y$ and $E_{Y|Z}$ means taking expectation on $Y$ given $Z$. $E_{B^*}$ and $E_{Y|Z,B^*}$ are defined similarly.

It is easy to see from (5) that $E_{B^*}(S(\beta, B^*)|Y) = 0$ is a consistent estimating equation for $\beta$. However, the explicit expression of $E_{B^*}(S(\beta, B^*)|Y)$ is analytically complicated. If the dimension of $B^*$ is $D$, then $E_{B^*}(S(\beta, B^*)|Y)$ is a sum of $2^D$ terms since $B^*$ consists of $D$ binary components, where $D$ depends on $n$. Although the mathematical form of $E_{B^*}(S(\beta, B^*)|Y)$ is not intractable, the realization in computer is almost impossible.

A natural idea is to find a numerical approximation of $E_{B^*}(S(\beta, B^*)|Y)$. Inspired by the law of large number, we have
\[
\lim_{R \to \infty} \frac{1}{R} \sum_{r=1}^{R} S(\beta, B^*(r)) \to E_{\beta^*}(S(\beta, B^*)|Y),
\]  

(6)
given \(B^*(r)\) is generated from the distribution of \(B\) given \(Y\). However, an issue occurs that the distribution of \(B^*(r)\) given \(Y\) is analytically intractable, and we address this issue by approximating the distribution of \(B^*(r)\) given \(Y\) by the distribution of \(B^r\) given \(X\), where \(X = (x_1', \ldots, x_n')\). Such approximation may lead to information loss, but it utilizes information as much as possible in our model setting. After the approximation, \(B^*(r)\) can be generated based on \(\hat{\eta}_{ij}\), the estimated mean of \(b_{ij}^r\).

Once having constructed equation (4), \(\hat{\beta}\) can be solved by quasi-Fisher algorithm. Denote \(z_i^*(r)\) by \((x_i', b_i^*(r))'\) for \(i = n_1 + 1, \ldots, n\), where \(b_i^*(r)\) is defined by completing the missing components \(b_{ij}^r\) of \(b_i\) by \(b_{ij}^*(r)\). Then sampling estimating equation (3) can be rewritten as
\[
\frac{1}{R} \sum_{r=1}^{R} S(\beta, B^*(r)) = \sum_{i=n_1+1}^{n} z_i (y_i - \mu_i(z_i')\mu_i(z_i'\beta) + \sum_{i=n_1+1}^{n} \frac{1}{R} \sum_{r=1}^{R} z_i^*(r) (y_i - \mu_i(z_i^*(r)\beta)) = 0,
\]  

(7)
so that the Fisher information of \(\beta\), which is the expectation of derivative of the sampling estimating equation (4) with respect to \(\beta\), is
\[
H(\beta) = \sum_{i=n_1+1}^{n} d_i(z_i')\mu_i(z_i') + \sum_{i=n_1+1}^{n} \frac{1}{R} \sum_{r=1}^{R} d_i(z_i^*(r)\beta)z_i^*(r)z_i^*(r)\beta,
\]  

(8)
where \(d_i(z_i')\beta = \exp(z_i'/\beta)/(1 + \exp(z_i'/\beta))^2\) and \(d_i(z_i^*(r)\beta)\) is similarly defined. Then \(\beta\) can be updated by
\[
\beta^{(1)} = \beta^{(0)} + [H(\beta^{(0)})]^{-1} \left[\frac{1}{R} \sum_{r=1}^{R} S(\beta^{(0)}, B^*(r))\right].
\]  

(9)
Once having obtained \(\hat{\beta}\), the mean \(\mu_i\) of \(Y_i\) can then be estimated by \(\hat{\mu}_i = \exp(z_i'/\hat{\beta})/(1 + \exp(z_i'/\hat{\beta}))\) for \(i = 1, \ldots, n_1\) and by \((1/R) \sum_{r=1}^{R} \exp(z_i^*(r)\hat{\beta})/(1 + \exp(z_i^*(r)\hat{\beta}))\) for \(i = n_1 + 1, \ldots, n\). If \(\hat{\mu}_i < 0.5\), \(\hat{y}_i\) is classified to 0, otherwise \(\hat{y}_i\) is classified to 1. The covariance of \(\hat{\beta}\) can be estimated by \(H(\hat{\beta})^{-1}\).

4. Real data analysis: application to a divorce data

In this section, we analyse the divorce data mentioned in introduction. The divorce data consists of 231 legal divorce cases in Sichuan Province, China. Each case includes 1 response variable, the result of judgement: divorce or not divorce; and 17 corresponding covariates: the ages of a couple (denoted by Age1 for age of wife, Age2 for age of husband); the time between the first meet and the marriage of the couple (Time Lag); whether the couple has children (Children); the number of the children (Nchild); the number of the past lawsuits for divorce of the couple (Nlawsuit); whether the couple have planned to the divorce (Plan); whether the husband or wife has an affair (Affair); whether the domestic violence occurs (Violence), whether one of the couple takes bet; whether the couple is lack of relationship (LR); whether economic problem occurs (Economy); whether parental teaching conflict occurs (PT); whether problem of family relationship arises (FR); whether the defendant approves to divorce (Approval); whether the couple continues to live separately for one year after the last time the court rejected divorce lawsuit (Sep1); whether the couple continues to live separately for two years (Sep2).

It is clear that the response variable and some of covariates are binary variables, which coincides with our model setting. There are missing binary covariates occurring in the last 116 of 231 cases for 12 binary covariates. 5 continuous covariates are completely observed in all cases. The appearance of missing values is due to the mistakes made by data collectors. So, the missing mechanism is missing completely at random. In addition, after computing the sample covariance matrix of the 12 binary covariates for 115 cases with completely observed covariates, we find the non-diagonal elements of this sample covariance are almost zeros. Therefore, it is reasonable to regard each of the 12 binary covariates is independent of the others.

We employ the proposed sampling estimating equation (3) to estimate the corresponding coefficients of 17 covariates. The results are summarized in Table 1, where the sample number \(R\) is set as 500.
Table 1. Estimation results for the divorce data by sampling estimating equation (SE means the standard error estimated by square roots of diagonal elements of $H(\beta)^{-1}$).

| Parameter  | Age1 | Age2 | Time lag | Nchid | Nlawsuit | Children | Plan     |
|------------|------|------|----------|-------|----------|----------|----------|
| SE         | 0.0779 | -0.0770 | 0.0107 | 0.4347 | -0.2055 | -0.5077 | 0.2039   |
| p-value    | 0.9662 | 0.8621 | 0.9603 | 0.1399 | 0.7414  | 0.1068  | 0.7722   |

| Parameter  | Affair | Violence | Bet | LR | Economy | PT | FR |
|------------|--------|----------|-----|----|---------|----|----|
| SE         | 9.5912 | -0.9976 | -1.9393 | 1.7718 | 0.0726 | -0.1301 | 0.7405 |
| p-value    | 0.0018 | 0.1832 | 0.2016 | 0.0079 | 0.9184 | 0.1068  | 0.7722 |

| Parameter  | Approval | Sep1 | Sep2 |
|------------|----------|-----|-----|
| SE         | 3.3688   | 1.4642 | 1.0568 |
| p-value    | <0.0001 | 0.0321 | 0.0223 |

Table 1 shows that only 5 variables are statistically significant at level 0.05, among the 17 covariates, including Affair, LR, Approval, Sep1 and Sep2. It is not difficult to understand the fact that one of a couple has an affair or the couple is lack of relationship would lead to divorce. The explanation for the statistical significance of Approval, Sep2 and Sep1 comes from the Marriage Law of the People's Republic of China. In fact, marriage law of the China regards whether the defendant approves to the divorce, whether the couple continues to live separately for one year after the last time the court rejected divorce lawsuit and whether the couple continues to live separately for two years as three main fundamental legal basis for divorce judgements. In addition, it is notable that Violence and Bet are not significant, this is probably due to most of cases of this divorce data are collected in rural areas of Sichuan Province, domestic violence and bet abuse are still common and frequent in rural areas, so that the court does not base its judgement on the domestic violence and bet abuse.

We also assess performance of the proposed sample estimating equation (3) in terms of classification and prediction on this divorce data. Compared to real data, we find that only 8 cases among 231 are incorrectly classified. As for prediction, we predict the judgement of $i$th case by first using the data of remaining $n-1$ cases to estimate covariates coefficients, then estimating the mean of $i$th response. If the estimated mean is less than 0.5, the $i$th response is predicted as 0, otherwise it is predicted as 1. We only predict responses of 115 cases with completely observed covariates as prediction cannot be conducted with missing values in covariates. We find 18 incorrectly predicted cases among 115. Overall, the sampling estimating equation performs well in terms of not only classification but also prediction on this divorce data.

Table 2. Estimation results for the divorce data by KNNI (SE means the standard error estimated using imputed Fisher information matrix).

| Parameter  | Age1 | Age2 | Time lag | Nchid | Nlawsuit | Children | Plan     |
|------------|------|------|----------|-------|----------|----------|----------|
| SE         | -0.0448 | -0.0340 | 0.1189 | 0.5703 | -0.3544 | -0.6291 | -0.2752 |
| p-value    | 0.7541 | 0.3812 | 0.2424 | 0.3343 | 0.7128  | 0.3572  | 0.8512  |
| SE         | 0.9526 | 0.9289 | 0.6238 | 0.0800 | 0.6191  | 0.0782  | 0.7465  |
| p-value    | 13.5384 | -0.7590 | -1.6852 | 2.0490 | -0.0538 | 0.0223  | 0.8469  |
| SE         | 3.1607 | 0.8619 | 1.7981 | 0.8449 | 0.9455  | 1.0994  | 0.5938  |
| p-value    | <0.0001 | 0.3785 | 0.3486 | 0.0153 | 0.9546  | 0.9838  | 0.1538  |

| Parameter  | Approval | Sep1 | Sep2 |
|------------|----------|-----|-----|
| SE         | 6.3567   | 1.6471 | 0.9944 |
| p-value    | 1.1785   | 0.7367 | 0.5276 |
| p-value    | <0.0001  | 0.0254 | 0.0595 |
For comparison, we apply the KNNI method to this divorce data. Data from the first 115 cases, in which there is no missing value, are used to impute the binary missing covariates of last 116 cases. After imputation, where the optimal number $k$ of neighbors is selected using mean square errors, we use usual logistic regression and the estimation results are shown in Table 2, where SE means the standard errors estimated by the square roots of the diagonal elements of inverse of imputed Fisher information matrix. We can see from Table 2 that only 4 variables are significant, including Affair, LR, Approval, Sep1. Compared to the results in Table 1, Sep 2 is regarded as insignificant. While according to the Marriage Law of China, Sep 2 should be included in the model. This implies that the proposed sampling estimating equation can obtain more efficient estimator than KNNI method. However, we find that KNNI method outperforms sampling estimating equation in terms of classification and prediction. Specifically, we find only 8 cases among 231 are incorrectly classified and only 4 cases among 115 are incorrectly classified.

5. Conclusion and discussion
In this paper, we propose a new estimating equation called sampling estimating equation to deal with the situation that multi-dimensional binary covariates are subject to missingness. The construction of this sampling estimating equation is based on the logistic model and sampling the missing components of binary covariates. The sampling estimating equation can be easily extended to the generalized linear model setting. Compared to existing methods focusing on 1-dimensional missing binary covariate, the main advantages of the sampling estimating equation are: reasonable parameter estimator can be obtained under missing multi-dimensional binary covariates, which cannot be handled by 1-dimensional approaches; sampling estimating equation can be applied to two-category classification for subjects with multi-dimensional missing covariates, which is intractable for existing 1-dimensional approaches. Compared to KNNI method, the estimator of sampling estimating equation is more efficient, which is confirmed by divorce data analysis. However, KNNI method outperform sampling estimating equation in terms of classification and prediction. The statistical inference based on the sampling estimating equation is also proposed. Specifically, the standard errors of the parameter estimates can be estimated by the square roots of the diagonal elements of $H(\hat{\beta})^{-1}$.

The divorce data is analysed to reveal the underlying legal basis for the divorce judgements in Sichuan Province. We conclude that whether one of the couple has an affair, whether the couple is lack of relationship, whether the defendant approves to divorce, whether the couple continues to live separately for one year after the last divorce lawsuit and whether the couple continues to live separately for two years are 5 main legal basis for the divorce judgements in Sichuan Province.

Note that in our model setting, we assume $b_{ij}$ is independent of $b_{ik}$. This assumption may be violated in practice. The relaxation of this assumption is the future interest. In addition, we use the distribution of $\mathcal{B}^{(r)}$ given $X$ to approximate the distribution of $\mathcal{B}^{(r)}$ given $Y$, which may lead to information loss. This information loss is expected to be figured out in the future. Finally, we are also interested in whether the sampling estimating equation can improve prediction performance under generalized linear model setting.

Acknowledgments
This work was supported by the National Key R & D Program of China (2018YFC0831900), Sichuan University Post-Doctor Research Project, the National Natural Science Foundation of China (11871357), and SCU LAIW.

References
[1] Chen H Y and Little R J A 1999 Proportional hazards regression with missing covariates *Journal of the American Statistical Association* 94 896–908
[2] Heinzle G 1998 Missing binary covariate data and imputation in regression models *Advances in Methodology, Data analysis and Statistics* 14 91-99
[3] Hsu H, Chan W, Goff D, Ortiz C, Chan M and Nichaman M 1995 Analysis of logistic
regression with missing binary covariates American Journal of Epidemiology 141 S46–S46

[4] Kleinbaum D G and Klein M 2002 Logistic Regression (A Self-Learning Text) Springer

[5] Little R J A and Rubin D B 2002 Statistical analysis with missing data Technometrics 45 364–5

[6] Nelder J A and Wedderburn R W M 1972 Generalized linear models Journal of the Royal Statistical Society 135 370–84

[7] Jonsson P and Wohlin C 2004 An evaluation of k-nearest neighbour imputation using Likert data 10th International Symposium on Software Metrics 2004 Proceeding 10.1109/Metric.2004.1357898 108-118

[8] Vach W and Martin S 1993 Logistic regression with incompletely observed categorical covariates Biometrika 80 353-362

[9] Schemper M and Heinze G 1997 Probability imputation revisited for prognostic factor studies Statistics in Medicine 16 73–80

[10] White I R and Royston P 2010 Imputing missing covariate values for the cox model Statistics in Medicine 28 1982–98

[11] Yi G Y, Ma Y, Spiegelman D and Carroll R J 2015 Functional and structural methods with mixed measurement error and misclassification in covariates Journal of the American Statistical Association 110 681–96