Bayesian approach for categorical Table with Nonignorable Nonresponse

Choi, BoSeung *, Park, YouSung †

Abstract

We propose five Bayesian methods to estimate the cell expectation in an incomplete multi-way categorical table with nonignorable nonresponse mechanism. We study 3 Bayesian methods which were previously applied to one-way categorical tables. We extend them to multi-way tables and, in addition, develop 2 new Bayesian methods for multi-way categorical tables. These five methods are distinguished by different priors on the cell probabilities: two of them have the priors determined only by information of respondents; one has a constant prior; and the remaining two have priors reflecting the difference in the response mechanisms between respondent and non-respondent.We also compare the five Bayesian methods using a categorical data for a prospective study of pregnant women.

KEY WORDS: Bayesian analysis; Nonignorable nonresponse; Priors; Boundary solution; EM algorithm

1. Introduction

The problem of missing data arising from nonresponse is common in most surveys and becomes a serious issue as the nonresponse rate increases.

Nonresponse can be distinguished by three types of nonresponses (Little and Rubin 1987): missing completely at random (MCAR) which means that the probability of missing on a variable of interest is independent of all variables including itself in the survey; missing at random (MAR) in which the nonresponse depends only on observed data; non-ignorable in which nonresponse depends on the unobserved values. Any model with MCAR or MAR is called ignorable nonresponse model.

When the response mechanism obeys nonignorable nonresponse in categorical data analysis, the maximum likelihood estimation often yields boundary solutions where the probability of nonresponse is estimated to be zero in some cells of the table. The conditions that the maximum likelihood (ML) suffers from the boundary solution have been proposed in one-way categorical

*Ph.D, korea University, Institute of Statistics, 1 5-ka Anamdong, Sungbuk-ku, Seoul, Korea
†Professor, Korea University, Dept. of Statistics, 1 5-ka Anamdong, Sungbuk-ku, Seoul, Korea

table (Baker and Laird 1988, Michels and Molenbergs 1997). Baker, Rosenberger and Dersimonian (1992) presented close forms of ML estimates for incomplete two-way categorical tables using loglinear model. In particular, they provided a sufficient and necessary condition under which the ML estimates fall in the boundary solution in two-way categorical tables.

Park and Brown(1994) and Park (1998) proposed a Bayesian approach to avoid the boundary solution problem in a one-way categorical table. The prior depends only on information of respondents. However, this respondent-driven prior contradicts to the fundamental principle that the nonrespondents have different response pattern from those of respondents in the nonignorable nonresponse model. We extend Park and Brown and Park's empirical Bayesian approach (1994, 1998) not only to a two-way categorical table but also to the prior depending on information from both respondent and nonrespondent. This prior can reflect different response patterns between respondents and nonrespondents. We also present generalized expectation maximization (EM) algorithm to estimate the cell probability specified by the loglinear models.

2. Bayesian models

We describe five Bayesian approaches to accommodate nonignorable nonresponse in a two-way categorical table. Let X_1 and X_2 be response variables indexed by I and J categories, respectively. We also let $R_1 = 1$ when X_1 is observed and $R_1 = 2$ when X_1 is missing. Let y_{ijkl} be the count belonging to the ith category of X_1, the jth category of X_2, the kth value of R_1, and the lth value of R_2.

Throughout this chapter, we assume a multinomial assumption for the three types of observations to have the following log likelihood proportional to

$$\begin{aligned} l \;\propto\; & \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) \\ + \; & \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}) \end{aligned} \tag{1}$$

where $\pi_{ijkl} = Pr[X_1 = i, X_2 = j, R_1 = k, R_2 = l]$ and $N = \sum_{i,j,k,l} y_{ijkl}$ is fixed.

To avoid a boundary solution of ML in model (1), we impose the Dirichlet priors to the cell probabilities (π_{ij11}, π_{ij12}, π_{ij21}, π_{ij22}) as given by

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}}. \tag{2}$$

The multinomial distribution of (1) for observations and the prior distribution of (2) yield the

following log posterior distribution:

$$l_{pos} = \sum_i \sum_j y_{ij11} \cdot (\mathbf{m}_{ij11} \cdot \beta) + \sum_i y_{i+12} \cdot \log\Big(\sum_j \exp(\mathbf{z}_{ij12} \cdot \beta)\Big)$$
$$+ \sum_j y_{+j21} \cdot \log\Big(\sum_i \exp(\mathbf{z}_{ij21} \cdot \beta)\Big) + y_{++22} \cdot \log\Big(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \beta)\Big)$$
$$+ \sum_{i,j,k,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \beta) - (N + \delta_{++++}) \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \beta)\Big) \quad (3)$$

We maximize the posterior distribution given in (3) over parameter β by the generalized expectation maximization (GEM) algorithm (Dempster, Laird and Rubin 1977) with the following E and M steps.

E-step : Using augmented y_{ij12} given y_{i+12}, y_{ij21} given y_{+j21}, and y_{ij22} given y_{++22} for $i = 1, \cdots, I$ and $j = 1, \cdots, J$, the posterior (3) can be written as this augmented posterior distribution

$$l_{a.pos} \propto \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) + \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12})$$
$$+ \sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) + \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}). \quad (4)$$

To determine the expected augmented log posterior of (4), we average over missing counts y_{ij12}, y_{ij21}, y_{ij22} conditioning on the current parameter estimates, π^{old}_{ijkl}, and observed counts y_{i+12}, y_{+j21}, and y_{++22}. Since y_{ij12}, y_{ij21}, and y_{ij22} are multinomial random variates conditioned on marginal sum y_{i+12}, y_{+j21}, and y_{++22}, respectively, the conditional expectations are given by

$$E_{old}(y_{ijkl} | \pi^{old}_{ijkl}, y_{i+kl}) = y_{i+kl} \frac{\pi^{old}_{ijkl}}{\pi^{old}_{i+kl}} = y_{i+kl} \frac{m^{old}_{ijkl}}{m^{old}_{i+kl}}, \quad \text{kl=12, 21 and 22}$$

M-step : In this step, we maximize the expected log posterior using the pseudo observations $\tilde{y}_{ij11} = y_{ij11} + \delta_{ij11}$, $\tilde{y}_{ij12} = y_{i+12} \frac{m^{old}_{ij12}}{m^{old}_{i+12}} + \delta_{ij12}$, $\tilde{y}_{ij21} = y_{+j21} \frac{m^{old}_{ij21}}{m^{old}_{+j21}} + \delta_{ij21}$, and $\tilde{y}_{ij22} = y_{++22} \frac{m^{old}_{ij22}}{m^{old}_{++22}} + \delta_{ij22}$. We impose the constraints on these pseudo observations so that their marginal sums are the same as the corresponding marginal sums of observations: $\tilde{y}_{++11} = y_{++11}, \tilde{y}_{i+12} = y_{i+12}, \tilde{y}_{+j21} = y_{+j21}$, and $\tilde{y}_{++22} = y_{++22}$. Then, the expected log posterior function has the same form as the likelihood obtained from a four-way contingency table with fully observed cell counts y^*_{ijkl}'s. Thus, using the iterative re-weighted least squares, we obtain the maximum posterior estimator (MPE) of β.

2.1 Five Types of Bayesian Methods

To complete the EM algorithm, we need to determine the hyper-parameters δ_{ijkl}'s. We set the sum of priors $\sum_{i,j,k,l} \delta_{ijkl}$ equal to the number of parameters involved in the loglinear model, p, as Clogg et al. (1991) did. Under this constraint (i.e., $\sum_{i,j,k,l} \delta_{ijkl} = p$), we propose five types of priors as follows. We first allocate δ_{ijkl} for the MPE of m_{ijkl} to shrink toward the MLE obtained

under ignorable nonresponse. That is, we determine δ_{ijkl} depending only on the respondent counts y_{ij11}, y_{i+12}, y_{+j21}, and y_{++22}. We call these priors respondent-driven priors and classify them into two types as below.

The first type of respondent-driven priors is, for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$,

$$\delta_{ijkl} = \nabla_{kl} \frac{y_{ij11}}{y_{++11}}, \tag{5}$$

where $\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$ for $k = 1, 2$ and $l = 1, 2$.

On the other hand, the second type of respondent-driven priors gives no prior on π_{ij11}. That is, The second type of priors are the same as those of the first type except $\delta_{ij11} = 0$ for all i and j. In case of one-way contingency table (i.e., either X_1 or X_2 is fully observed without missing) and $y_{++22} = 0$, the first type is reduced to Park (1998), whereas the second type is reduced to Park and Brown (1994). These two types of respondent-driven priors may bring a controversy because the nonrespondents are usually assumed to have different response patterns from the respondents in the nonignorable model.

In order to reflect different response patterns between the respondent and nonrespondent, we propose the following third type of priors δ_{ijkl} depending on both respondent's and nonrespondent's information. So δ_{ijkl} is assigned to be proportional to expected cell frequencies, m^{old}_{ijkl}, where calculated at the previous iteration. We distinguish this third type of priors $\tilde{\delta}_{ijkl}$ from previous priors δ_{ijkl}:

$$\tilde{\delta}_{ijkl} = \begin{cases} \nabla_{kl} \cdot \left(\frac{m^{old}_{ijkl}}{m^{old}_{++kl}} \right) & \text{for } k = 1,\ l = 1 \\ \nabla_{kl} \cdot \left(\frac{m^{old}_{ijkl}}{m^{old}_{++kl}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2} & \text{for } k \neq 1 \text{ or } l \neq 2 \end{cases} \tag{6}$$

where $\nabla_{kl} = p \cdot \frac{m^{old}_{++kl}}{m^{old}_{++++}}$ for $k = 1, 2$ and $l = 1, 2$.

Therefore, these new priors depend on their respective parameters m^{old}_{ijkl} to be estimated in previous iteration. The main reason we use a weighted priors of $m^{old}_{ijkl}/m^{old}_{++kl}$ and $1/IJ$ on $\tilde{\delta}_{ij12}$, $\tilde{\delta}_{ij21}$, and $\tilde{\delta}_{ij22}$ is to prevent a boundary solution on m_{ij12}, m_{ij21}, and m_{ij22}, respectively. We also define the fourth type of priors by letting $\tilde{\delta}_{ij11} = 0$ in (6) as we obtained the second type from the first type.

The last type of priors extend the constant prior of Clogg et al. (1991) used for one-way categorical table to those for two-way categorical table as follows.

$$\tilde{\delta}_{ijkl} = \begin{cases} 0 & \text{if } k = 1,\ l = 1 \\ \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right) & \text{for } k \neq 1 \text{ or } l \neq 2. \end{cases} \tag{7}$$

These five types of priors will be compared in the subsequent two sections using empirical data and simulation studies.

Table 1: Data for the relationship between smoking and new born's weight

	Birth weight(in grams)		
Smoker	<2500	≥2500	Missing
Yes	4512	21009	1049
No	3394	24132	1135
Missing	142	464	1224

3. Case study

We compare the five Bayesian methods with the maximum likelihood estimate (ML) through a prospective study of pregnant women to access the relationship between smoking status and newborn's weight. As competitors, we also consider two other methods (i.e., another ignorable model and another nonignorable model). Table 1 provides a categorical data for a prospective study of pregnant women to assess the relationship between perinatal factors and the subsequent development and course of abnormalities in the offspring (Baker, Rosenberger and Dersimonian 1992). The categorical table cross-classifies mother's self-reported smoking status (smoker or non-smoker) with newborn's weight (<2500 grams, ≥2500 grams). The column supplement contains only data on smoking status (4 percent of the data), the row supplement contains only data on newborn's weight (1 percent of the data), and the other is the count of the number missing data on both variables (2 percent of the data).

For comparison, we consider the following one ignorable and two nonignorable nonresponse models.

$$\begin{aligned}
\text{Model 1} &: \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}, \\
\text{Model 2} &: \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}, \\
\text{Model 3} &: \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{il}_{X_1R_2} + \beta^{jk}_{X_2R_1} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}. \qquad (8)
\end{aligned}$$

Denote the ML estimates under Model 1, Model 2, and Model 3 by IG_{ML}, $NIG1_{ML}$, and $NIG2_{ML}$, respectively. We also let $NIG1_{BE1}$, $NIG1_{BE2}$, and $NIG1_{BE3}$ be the Bayesian estimates with priors δ_{ijkl} depending on parameter m_{ijkl} given by (6), with the same priors as $NIG1_{BE1}$ except $\delta_{ij11} = 0$, and with the constant priors given by (7), respectively. Finally, let $NIG1_{BE4}$ and $NIG1_{BE5}$ be the empirical Bayesian estimates with the respondent-driven priors given by (5) where $NIG1_{BE5}$ has the same priors as $NIG1_{BE4}$ except $\delta_{ij11} = 0$. All of these $NIG1_{BE1}$, $NIG1_{BE2}$, $NIG1_{BE3}$, $NIG1_{BE4}$, and $NIG1_{BE5}$ are obtained under Model 2 given in (8).

Table 2 summarizes the four conditional probabilities from each of the eight estimation methods. The $NIG2_{ML}$ is actually the model that Baker, Rosenberger and Dersimonian (1992) selected. From the third and sixth rows, we can observe that there is no difference between IG_{ML} and $NIG2_{ML}$, implying little advantage of the nonignorable Model 3 over the ignorable Model

Table 2: Conditional probabilities in the relationship between smoking and new born's weight (W:weight, S:smoking, NS:non-smoking)

	$NIG1_{ML}$	$NIG1_{BE1}$	$NIG1_{BE2}$	$NIG1_{BE3}$	$NIG1_{BE4}$	$NIG1_{BE5}$	IG_{ML}	$NIG2_{ML}$
$P(W < 2500\|S)(1)$	.1774	.1781	.1834	.1850	.1774	.1775	.1779	.1799
$P(W < 2500\|NS)(2)$	.1231	.1233	.1256	.1268	.1231	.1231	.1241	.1256
$\frac{(1)}{(2)}$	1.441	1.444	1.460	1.459	1.441	1.442	1.434	1.432
$P(S\|W < 2500)(3)$	.5883	.5885	.5892	.5862	.5879	.5842	.5707	.5707
$P(S\|W \geq 2500)(4)$	.4817	.4814	.4784	.4754	.4813	.4777	.4654	.4654
$\frac{(3)}{(4)}$	1.221	1.222	1.232	1.233	1.221	1.223	1.226	1.226

1. Compared to IG_{ML}, the $NIG1_{BE1}$, $NIG1_{BE2}$, and $NIG1_{BE3}$ produce larger conditional probabilities P(weight $<$ 2500|smoking) and P(weight $<$ 2500|non-smoking) with the exception of $NIG1_{BE1}$ for P(weight $<$ 2500| non-smoking). This is completely reversed for $NIG1_{BE4}$ and NIG_{BE5}. Thus, $NIG1_{BE1}$, $NIG1_{BE2}$, and $NIG1_{BE3}$ more allocate the column supplements into the category "weight $<$ 2500" than IG_{ML}, but $NIG1_{BE4}$ and NIG_{BE5} less allocate than IG_{ML}. Since IG_{ML} can ignore the supplement information in inference, we may conclude that $NIG1_{BE1}$, $NIG1_{BE2}$, and $NIG1_{BE3}$ are more reasonable for the nonignorable model, Model 2.

4. Concluding Remarks

We investigated Bayesian analysis for incomplete two-way categorical tables with nonignorable nonresponse under which the maximum likelihood estimates often fall in the boundary solution, causing the ML estimates unstable. To avoid the boundary solution problem, we proposed the five types of Bayesian methods. These Bayesian methods include the previous Bayesian models as special cases. The two among the five Bayesian models were proposed to reflect different response patterns between respondents and nonrespondents.

Data analysis showed that these new Bayesian methods were more reasonable in the sense that nonignorable nonrespose mechanisms are more reflected and close to the actual results.

References

Baker, S. G. and Laird, N. M. (1988), "Regression analysis for categorical variables with outcome subject to nonignorable nonresponse," *Journal of the American Statistical Association*, **83**, 62-69.

Baker, S. G., Rosenberger, W. F., and Dersimonian, R. (1992), "Closed-form estimates for missing counts in two-way contingency tables," *Statistics in Medicine*, **11**, 643-657.

Clogg, C. C., Rubin, D. B., Schenker, N., and Schultz, B. (1991), "Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression,"

Journal of the American Statistical Association, **86**, 68-78.

Little, J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, second edition. Wiley, New York.

Park, T. and Brown, M. B. (1994), "Models for categorical data with nonignorable nonresponse," *Journal of the American Statistical Association*, **89**, 44-52.

Park, T. (1998), "An approach to categorical data with nonignorable nonresponse," *Biometrics*, **54**, 1579-1690.