IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Review on the boundary solutions in the analysis of an incomplete contingency table

Seongyong Kim

Assistant Professor, Division of Global Management Engineering, Hoseo University, Asan, Korea

Abstract: In the analysis of an incomplete contingency table, nonresponse models incorporating the missing mechanism are used to estimate nonresponses. There are three missing mechanisms, missing at random (MAR), not missing at random (NMAR) and missing completely at random (MCAR). The estimation results differ depending on the embedded missing mechanism in the nonresponse model. When the NMAR mechanism is assumed, it has been known that the nonresponse model has boundary solution problems. Boundary solutions are defined as the cell probabilities of certain columns are estimated to be zero, leading to distort estimation results. In this paper, boundary solution problems are reviewed with the reason, the identification of occurrence, and the method to overcome. The introduction of the analysis of an incomplete contingency table is also provided with the real data analysis.

Keywords: incomplete contingency table, boundary solutions, ML estimation, Bayesian method, NMAR model.

I. INTRODUCTION

A contingency table is a table that shows the distribution of categorical variables, and analysed to measure the association between variables, for example independence or homogeneity of variables [1]. When two variables are considered, a contingency table is called a two-way contingency table. Table 1 is an example of a two-way contingency table, and shows the distribution of newborns' birth weight and mothers' self-reported smoking [2, 3].

TABLE 1 TWO-WAY CONTINGENCY TABLE

Birth weight	Self-reported smoking	
	Yes	No
< 2500	4512	3394
≥ 2500	21009	24132

When categorical variables have nonresponses, categorical data can be summarized as an incomplete contingency table. When there are nonresponses for one variable, the incomplete contingency table has one supplementary margin. When there are nonresponses for two variables, the incomplete contingency table summarizing data has two supplemental margins. Table 2 is examples of two-way incomplete contingency table with one supplemental margin.

TABLE 2 TWO-WAY CONTINGENCY TABLE WITH ONE SUPPLEMENTAL MARGIN

Birth weight	Self-reported smoking		
	Yes	No	Missing
< 2500	4512	3394	142
≥ 2500	21009	24132	464

In the analysis of incomplete contingency tables, the purpose is to estimate nonresponses, however, it has been known that the occurrence of nonresponses (called missingness) depend on distinct patterns. These patterns are called the missing mechanism. To estimate nonresponses, the missing mechanism is embedded in a nonresponse model.

Little and Rubin [4] define three missing mechanisms: missing at random (MAR), not missing at random (NMAR) and missing completely at random (MCAR). MAR is missing mechanism that missingness depends on the observed data, NMAR is when the missingness depends on the unobserved data, and MCAR is when missingness depends on neither.
For example, when mothers do not respond to the questions about smoking because their newborn's weight is low, then the missing mechanism is MAR, however, when mothers do not respond because they were smoking, then the missing mechanism is NMAR. When nonresponses occur regardless of birth's weight or smoking, then the missing mechanism is MCAR.

DOI10.17148/IJARCCE.2017.6841

For the nonresponse model to accommodate the missing mechanism, log-linear models have been widely employed [2, 5, 6, 7, 8, 9, 10]. Depending on the assumed missing mechanism, nonresponse models produce different estimates. For convenience, we call nonresponse models incorporating MAR, NMAR, and MCAR mechanism MAR model, NMAR model and MCAR model, respectively.

When the missing mechanism is specified as NMAR, it has been known that nonresponse boundary solutions often occur in ML estimation [6]. In a two way incomplete contingency table with one incomplete response variable and an observed covariate (hereafter I × J × 2 incomplete contingency table), boundary solutions under the NMAR model take forms such that the cell probabilities concerned with nonresponse are estimated to be all zero's for certain values of the response variable [6, 9].

In this paper, it is reviewed that the analysis of an incomplete contingency table and boundary solutions under the NMAR model which we often encounter in the real data analysis. In Section 2, log-linear models depending on each missing mechanism and the estimation method by EM algorithm are introduced. In Section 3, examples of boundary solutions and their properties are reviewed. Section 4 includes conclusion.

II. LOG-LINEAR MODELS

Let X be a completely observed categorical variable with I categories, and Y be an incomplete observed categorical variable with J categories. When Y is observed, let an indicator variable of missingness denoted by R be 1, and R is 2 when Y is not observed. Then, the full array of X, Y, and R produce a $I \times J \times 2$ table with cell counts $\boldsymbol{y} = \{y_{ijk}\}$ where $i = 1, \cdots, I$, $j = 1, \cdots, J$, and $k = 1,2$. However, we can only observe $\boldsymbol{y_{obs}} = (\{y_{ij1}\}, \{y_{i+2}\})$ where the symbol "+" in the subscript indicates the summation over the corresponding subscript. Under the assumption that the observed cell counts follow a multinomial distribution with cell probabilities $\boldsymbol{\pi} = \{\pi_{ijk}\}$ and a given total count $N = \sum_i \sum_j \sum_k y_{ijk}$, a log-linear model is linked to cell probabilities. That is, under the multinomial distribution assumption, $\pi_{ijk} = m_{ijk} / \sum_{ijk} m_{ijk}$ where $m_{ijk} = E(y_{ijk})$ and m_{ijk} is modelled by a log-linear model [1]. Depending on the predetermined missing mechanism, we have three types of log-linear models as followings [3, 6, 9].

MAR model : $log\ m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k + \lambda_{XY}^{ij} + \lambda_{XR}^{ik}$
NMAR model : $log\ m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k + \lambda_{XY}^{ij} + \lambda_{YR}^{jk}$
MCAR model : $log\ m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k + \lambda_{XY}^{ij}$

In the MAR model, the inclusion of λ_{XR}^{ik} means that the missingness of Y depends on X. The opposite is true for the NMAR model while the MCAR model depends on neither. Note that $\sum_i \lambda_X^i = \sum_j \lambda_Y^j = \sum_k \lambda_R^k = \sum_i \lambda_{XY}^i = \sum_j \lambda_{XY}^j = \sum_i \lambda_{XR}^i = \sum_k \lambda_{XR}^k = \sum_j \lambda_{YR}^j = \sum_k \lambda_{YR}^k = 0$ for the identification of coefficients.

To obtain ML estimates of the coefficients in a log-linear model, EM algorithm has been widely used [6, 11]. Under the assumption that y_{ij2}s are observed, we have a following complete log-likelihood.

$$\ell_c = \sum_i \sum_j y_{ij1} log\pi_{ij1} + \sum_i \sum_j y_{ij2} log\pi_{ij2}\ .$$

For a fixed i, we also assume that y_{ij2}'s follow multinomial distribution with a given total count y_{i+2}. EM algorithm has two steps which are E-step and M-step, and these steps are repeated recursively until the stoping rule is satisfied [12]. At the t-th iteration, we have following E-step and M-step. In E-step, the expectation of the above complete log-likelihood is calculated as following:

$$E[\ell_c|\boldsymbol{\lambda}^{t-1}, y_{obs}] = \sum_i \sum_j y_{ij1} log\pi_{ij1}^{t-1} + \sum_i \sum_j E[y_{ij2}|\boldsymbol{\lambda}^{t-1}, y_{i+2}] log\pi_{ij2}^{t-1}$$

where $\boldsymbol{\lambda}^{t-1}$ is a vector of λ's from M-step at the $t-1$ th iteration, and

$$E[y_{ij2}|\lambda^{t-1}, y_{i+2}] = y_{i+2}\frac{\pi_{ij2}^{t-1}}{\sum_{ij} \pi_{ij2}^{t-1}}$$

by the multinomial assumption. In M-step, $E[\ell_c|\boldsymbol{\lambda}^{t-1}, y_{obs}]$ is maximized with respect to $\boldsymbol{\lambda}$ by optimization methods such as Newton-Raphson method, and set $\boldsymbol{\lambda}^t$ equal to the maximizer of $E[\ell_c|\boldsymbol{\lambda}^{t-1}, y_{obs}]$. These two steps are repeated until the difference of two adjacent log-likelihood values are less than an arbitrary positive value.

In the analysis of an incomplete contingency table under this framework, some ongoing issues have been arisen, the assessment of the missing mechanism, non-identification problem, and boundary solution problem under the NMAR

DOI10.17148/IJARCCE.2017.6841

model [13, 14]. Since we very often encounter the boundary solution problems practically, in this paper, boundary solution problems under the NMAR model are discussed.

III. BOUNDARY SOLUTIONS OF THE NMAR MODEL

In the analysis of an incomplete contingency table, boundary solutions are generally defined as $\hat{\pi}_{ij2} = 0$ for at least one combination of (i, j) [6]. Under the NMAR model, it has been known that boundary solutions take a following form: $\hat{\pi}_{+j2} = 0$ for at least one and at most $(J - 1)$ values of j [9]. Table 3 presents the estimation results under the NMAR model for a data provided in Table 2.

TABLE 3 BOUNDARY SOLUTIONS IN A $2 \times 2 \times 2$ CONTINGENCY TABLE

	Smoking ($R = 1$)		Smoking ($R = 2$)	
	Y = 1 (Yes)	Y = 2 (No)	Y = 1 (Yes)	Y = 2 (No)
X = 1 (< 2500)	4546	3394	108	0
X = 2 (≥ 2500)	20975	24132	498	0

As illustrated in Table 3, when the NMAR model is applied, all nonresponses at Y = 2 are estimated to be 0. That is, boundary solutions occur at Y = 2. In terms of $\boldsymbol{\lambda}$ in the NMAR model, boundary solutions can be expressed as $|\hat{\lambda}_{YR}^{jk}| = \infty$ [9]. The occurrence of boundary solutions distort the estimation results by allocating none value to certain categories while assigning all nonresponses to the other categories regardless of the true value. In addition, under the occurrence of boundary solution, ML estimates do not provide a perfect fit to observed data even though the saturated NMAR model is used.

Some researchers explained the reason boundary solutions occur [7, 8, 9]. When data lie out of the parameter space, ML estimates lie on the parameter space closest to data. Sometimes, ML estimates can lie on the edge of the parameter space, in this case, boundary solutions occur.

The practical problem about boundary solutions is whether $\hat{\pi}_{+j2} = 0$ is from boundary solutions or not. Without suffering from boundary solutions, nonresponses of specific columns also can be estimated to be 0. Thus, to judge that boundary solutions occur or not, the sufficient conditions for the occurrence of boundary solutions have been suggested. Baker and Laird [6] provided a tool for the identification of occurrence of boundary solutions for a $2 \times 2 \times 2$ contingency table. They defined response odds as $\omega_j = y_{1j1}/y_{2j1}$ for $j = 1, 2$ and nonresponse odds as $\omega = y_{1+2}/y_{2+2}$. They showed that boundary solutions occur when ω lies out of the interval between ω_1 and ω_2. By using their conditions, boundary solutions can be checked without estimation through the EM algorithm. Park et al. [15] extended the sufficient conditions proposed by Baker and Laird [6] to a $I \times I \times 2 \times 2$ contingency table. Note that, for a general shape of incomplete contingency table, denoted by a $I \times J \times 2 \times 2$ contingency table, the conditions for the occurrence of boundary solutions have not been provided.

To overcome boundary solutions under the NMAR model, various Bayesian approaches have been suggested. Foster and Smith [16], Nandram et al. [17] and Green and Park [18] proposed hierarchical Bayesian model by using MCMC. Park and Brown [19] and Choi et al. [11] proposed Bayesian methods by using EM algorithm. Park and Brown [19] used empirical priors for cell probabilities, and Choi et al. [11] used a mixture of constant and empirical priors. Park and Brown [19] imposed Dirichlet priors to cell probabilities given by

$$\prod_i \prod_j \pi_{ij1}^{\delta_{ij1}} \pi_{ij2}^{\delta_{ij2}}$$

where the δ_{ijk}s are defined as

$$\delta_{ij1} = 0 \,, \delta_{ij2} = p \frac{y_{ij1}}{y_{++1}}$$

and p is the number of parameters. Table 4 presents the estimates by using Park and Brown [19]. As expected, some values are allocated to Y = 2 unlike the result in Table 3.

TABLE 4 ESTIMATES BY USING PARK AND BROWN METHOD

	Smoking (R = 1)		Smoking (R = 2)	
	Y = 1 (Yes)	Y = 2 (No)	Y = 1 (Yes)	Y = 2 (No)
X = 1 (< 2500)	4546	3399	90	12
X = 2 (**≥ 2500**)	20975	24127	416	87

DOI10.17148/IJARCCE.2017.6841

Rather than using these Bayesian method, some researchers recommend to use other models such as the MAR model. Baker et al. [20] presented that boundary solutions under the NMAR model likely indicate the model misspecification.

IV. CONCLUSION

In this paper, first, the overview of the analysis of an incomplete contingency is introduced. To estimate nonresponses, nonresponse models incorporating missing mechanism are used which are the NMAR, MAR and MCAR models. To obtain ML estimates, EM algorithm is used. However, we often encounter boundary solution problems when the NMAR model is used.

Boundary solutions are defined as the cell probabilities of certain columns are estimated to be zero, and can distort the estimation results. To judge whether boundary solutions occur or not, the sufficient conditions were provided. By using these conditions, boundary solutions can be checked without using EM algorithm.

To overcome boundary solutions, various Bayesian methods have been suggested. In this paper, real data analysis by using ML method and Park and Brown method are performed, and show that boundary solutions occur under ML estimation, and do not under Park and Brown method. One thing noteworthy is that, when boundary solutions occur under the NMAR model, it would be appropriate to use another model such as MAR model as described by Baker et al. [20].

REFERENCES

[1] A. Agresti, Categorical Data Analysis, 2nd Edition. New York: John Wiley & Sons, Inc, 2002.

[2] S. G. Baker, W. F. Rosenberger, and R. Dersimonian, Closed-form estimates for missing counts in two-way contingency tables. Statistics in Medicine, 11, pp. 643-657, 1992.

[3] S. Kim, Y. Park, and D. Kim, On missing-at-random mechanism in two-way incomplete contingency tables. Statistics & Probability letters, 96, pp. 196-203, 2015.

[4] J. A. Little, and D. B. Rubin, Statistical Analysis with Missing Data, 2nd Edition. New York: John Wiley & Sons, Inc., 2002.

[5] R. E. Fay, Causal models for patterns of nonresponse. Journal of the American Statistical Association, 81, pp. 354-365, 1986.

[6] S. G. Baker, and N. M. Laird, Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. Journal of the American Statistical Association, 83, pp. 62-69, 1988.

[7] P. W. F. Smith, C. J. Skinner, and P.S. Clarke, Allowing for non-ignorable nonresponse in the analysis of voting intention data. Journal of the Royal Statistical Society: series C, 48, pp. 563-577, 1999.

[8] P. S. Clarke, On boundary solutions and identifiability in categorical regression with non-ignorable non-response. Biometrical Journal, 44, pp. 701-717, 2002.

[9] P. S. Clarke, and P. W. F. Smith, Interval estimation for log-linear models with one variable subject to non-ignorable non-response. Journal of the Royal Statistical Society: Series B, 66, pp. 357-368, 2004.

[10] P. S. Clarke, and P. W. F. Smith, On maximum likelihood estimation for log-linear models with non-ignorable non-responses Statistics & Probability Letters, 73, pp. 441-448, 2005.

[11] B. S. Choi, J. W. Choi, and Y. Park, Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls. Survey Methodology, 35, pp. 37-51, 2009.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society: Series B. 39, pp. 1-38, 1977.

[13] F. Z. Poleto, J. M. Singer, and C. D. Paulino, Missing data mechanisms and their implications on the analysis of categorical data. Statistics and Computing, 21, pp. 31-43, 2011.

[14] G. Molenberghs, E. Goetghebeur, S. R. Lipsitz, and M. G. Kenward, Nonrandom Missingness in Categorical Data: Strengths and Limitations. The American Statistician, 53, pp. 110-118, 1999.

[15] Y. Park, D. Kim, and Y. Kim, Identification of the occurrence of boundary solutions in a contingency table with nonignorable nonresponse. Statistics and Probability Letters, 93, pp. 34-40, 2014.

[16] J. J. Foster, and P. W. F. Smith, Model-based inference for categorical survey data subject to non-ignorable non-response. Journal of the Royal Statistical Society: Series B, 60, pp. 57-70, 1998.

[17] B. Nandram, L. H. Cox, J. W. Choi, Bayesian Analysis of Nonignorable Missing Categorical Data: An Application to Bone Mineral Density and Family Income. Survey Methodology, 31, pp. 213-225, 2005.

[18] P. E. Green, T. Park, A Bayesian hierarchical model for categorical data with nonignorable nonresponse. Biometrics, 59, pp. 886-896, 2003.

[19] T. Park, and M. B. Brown, Models for categorical data with nonignorable nonresponse. Journal of the American Statistical Association, 89, pp. 44-52, 1994.

[20] S. Baker, C. Ko, and B. I. Graubard, A sensitivity analysis for norandomly missing categorical data arising from a national health disability survey. Biostatistics, 4, pp. 41-56, 2003.

DOI10.17148/IJARCCE.2017.6841