Statistics and Probability Letters 93 (2014) 34–40

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Identification of the occurrence of boundary solutions in a contingency table with nonignorable nonresponse

Yousung Park[a], Daeyoung Kim[b], Seongyong Kim[c,*]

[a] Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

[b] Department of Mathematics and Statistics, Lederle Graduate Research Tower, Box 34515, University of Massachusetts Amherst, Amherst, MA, USA

[c] Institute of Economic Research, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 11 June 2014
Accepted 12 June 2014
Available online 19 June 2014

Keywords:
Boundary solutions
Incomplete contingency table
Log-linear model
Not missing at random

ABSTRACT

We establish a sufficient condition for the occurrence of nonresponse boundary solutions in a two-way square contingency table with nonignorable nonresponse. The condition depends only on the observed counts and thus it does not require computing the maximum likelihood estimates.

1. Introduction

In the analysis of incomplete contingency tables (contingency tables with nonresponses), there are three types of missing mechanisms (Little and Rubin, 2002): missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). MCAR is when missingness is independent of observed and unobserved data; MAR is when missingness depends only on observed data; and NMAR is when missingness depends on unobserved data. Nonresponses are called nonignorable nonresponses when missing mechanism is NMAR while ignorable nonresponses when missing mechanism is MAR or MCAR.

To incorporate the missing mechanism in the models for the incomplete contingency tables, log-linear models have been widely used (Fay, 1986; Baker and Laird, 1988; Baker et al., 1992; Smith et al., 1999; Clarke, 2002; Clarke and Smith, 2004, 2005). However, the log-linear models under NMAR mechanism (hereafter called NMAR models) have the problem of nonresponse boundary solutions (hereafter called boundary solutions) in the use of maximum likelihood (ML) estimation (Baker and Laird, 1988; Baker et al., 1992; Smith et al., 1999; Clarke, 2002; Clarke and Smith, 2004). For example, in a two-way incomplete contingency table with only one variable subject to missingness (denoted by $I \times J \times 2$ incomplete contingency table), the ML estimates often fall on the boundary of the parameter space such that the cell probabilities concerned with nonresponse are estimated to be all zeros for certain values of the variable (Baker and Laird, 1988; Clarke, 2002; Clarke and Smith, 2004).

Baker and Laird (1988) presented a condition for the occurrence of boundary solutions in the NMAR models for a $2 \times 2 \times 2$ incomplete contingency table. Baker et al. (1992) studied the problem of boundary solutions in a two-way contingency table

* Corresponding author. Tel.: +82 2 3290 2238.
E-mail addresses: yaba96@korea.ac.kr, skim@math.umass.edu (S. Kim).

http://dx.doi.org/10.1016/j.spl.2014.06.011
0167-7152/

Table 1
NMAR models for an $I \times J \times 2 \times 2$ contingency table.

Model	NMAR model
M1	$\log m_{ijk\ell} = \lambda^{i}_{Y_1} + \lambda^{j}_{Y_2} + \lambda^{k}_{R_1} + \lambda^{\ell}_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{k\ell}_{R_1R_2} + \lambda^{ik}_{Y_1R_1}$
M2	$\log m_{ijk\ell} = \lambda^{i}_{Y_1} + \lambda^{j}_{Y_2} + \lambda^{k}_{R_1} + \lambda^{\ell}_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{k\ell}_{R_1R_2} + \lambda^{j\ell}_{Y_2R_2}$
M3	$\log m_{ijk\ell} = \lambda^{i}_{Y_1} + \lambda^{j}_{Y_2} + \lambda^{k}_{R_1} + \lambda^{\ell}_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{k\ell}_{R_1R_2} + \lambda^{ik}_{Y_1R_1} + \lambda^{i\ell}_{Y_1R_2}$
M4	$\log m_{ijk\ell} = \lambda^{i}_{Y_1} + \lambda^{j}_{Y_2} + \lambda^{k}_{R_1} + \lambda^{\ell}_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{k\ell}_{R_1R_2} + \lambda^{jk}_{Y_2R_1} + \lambda^{j\ell}_{Y_2R_2}$
M5	$\log m_{ijk\ell} = \lambda^{i}_{Y_1} + \lambda^{j}_{Y_2} + \lambda^{k}_{R_1} + \lambda^{\ell}_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{k\ell}_{R_1R_2} + \lambda^{ik}_{Y_1R_1} + \lambda^{j\ell}_{Y_2R_2}$

with incompletely observed variables (denoted by $I \times J \times 2 \times 2$ incomplete contingency table). They showed that, in order to check the occurrence for boundary solutions, it is required to solve a system of likelihood equations with respect to the odds of the response and nonresponse cell probabilities. Smith et al. (1999) and Clarke (2002) gave a geometric description of the boundary solution problem. Clarke and Smith (2005) studied existence, uniqueness and asymptotic theory of ML estimators for the parameters in the NMAR models when boundary solutions occur.

In this paper, we establish a sufficient condition for the occurrence of boundary solutions under several identifiable NMAR models for an $I \times I \times 2 \times 2$ incomplete contingency table. The sufficient condition we present depends only on the odds of observed (joint/marginal) cell counts available in the aforementioned contingency table. Therefore, the occurrence of boundary solutions for the cell probabilities in a given data set can be easily checked, without either using EM algorithm (Dempster et al., 1977) or solving a system of likelihood equations. Note that the presented results in this paper generalize the result by Baker and Laird (1988) to the case of an $I \times I \times 2 \times 2$ incomplete contingency table.

The rest of the paper is organized as follows. In Section 2, we specify the five identifiable NMAR models widely used in an $I \times J \times 2 \times 2$ incomplete contingency table. Section 3 provides a sufficient condition for the occurrence of boundary solutions in an $I \times I \times 2 \times 2$ incomplete contingency table when the cell probabilities are modeled by the five NMAR models introduced in Section 2. An application of the results presented in Section 3 using real data is given in Section 4. Conclusion remarks are in Section 5.

2. Identifiable NMAR log-linear models

In this section, we define the likelihood for the cell probabilities in an $I \times J \times 2 \times 2$ contingency table and specify the five types of identifiable NMAR models.

Let Y_1 and Y_2 be two categorical variables with I and J categories, respectively. We also let R_1 be an index variable of missingness for Y_1 such that $R_1 = 1$ if Y_1 is observed and $R_1 = 2$ if Y_1 is not observed. Similarly, $R_2 = 1$ if Y_2 is observed and $R_2 = 2$ otherwise. For the full array of Y_1, Y_2, R_1 and R_2, we have an $I \times J \times 2 \times 2$ contingency table with cell counts $\mathbf{y} = \{y_{ijk\ell}\}$ where $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k, \ell = 1, 2$. What we observe, however, is only $\mathbf{y}_{\text{obs}} = (\{y_{ij11}\}, \{y_{i+12}\}, \{y_{+j21}\}, \{y_{++22}\})$ where the symbol "+" in the subscripts denotes summation over the corresponding subscript. The left table in Table 3 shows an example of a $2 \times 2 \times 2 \times 2$ incomplete contingency table. For observed cell counts, we assume a multinomial distribution with cell probabilities $\boldsymbol{\pi} = \{\pi_{ijk\ell}\}$ and a fixed total count $N = \sum_i \sum_j \sum_k \sum_\ell y_{ijk\ell}$. Then the log-likelihood of $\boldsymbol{\pi}$ is

$$\ell(\boldsymbol{\pi}; \mathbf{y}_{\text{obs}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij11} \log \pi_{ij11} + \sum_{i=1}^{I} y_{i+12} \log \pi_{i+12} + \sum_{j=1}^{J} y_{+j21} \log \pi_{+j21} + y_{++22} \log \pi_{++22} \tag{1}$$

where $\pi_{ijk\ell} = m_{ijk\ell} / \sum_i \sum_j \sum_k \sum_\ell m_{ijk\ell}$ and $\mathbf{m} = \{m_{ijk\ell}\}$ are the expected cell counts that are modeled by a log-linear model under an assumed missing mechanism.

Baker et al. (1992) proposed the nine identifiable nonresponse log-linear models for $\boldsymbol{\pi}$ (i.e., $\mathbf{m}$) in the log likelihood of Eq. (1), depending on missing mechanisms in the two categorical variables, Y_1 and Y_2. In this paper, we focus on the five NMAR models which have at least one variable subject to NMAR mechanism, as shown in Table 1. Note that the constraints on the parameters required for their estimation are: $\sum_i \lambda^{i}_{Y_1} = \sum_j \lambda^{j}_{Y_2} = \sum_k \lambda^{k}_{R_1} = \sum_\ell \lambda^{\ell}_{R_2} = \sum_i \lambda^{ij}_{Y_1Y_2} = \sum_j \lambda^{ij}_{Y_1Y_2} = \sum_i \lambda^{ik}_{Y_1R_1} = \sum_k \lambda^{ik}_{Y_1R_1} = \sum_i \lambda^{i\ell}_{Y_1R_2} = \sum_\ell \lambda^{i\ell}_{Y_1R_2} = \sum_j \lambda^{jk}_{Y_2R_1} = \sum_k \lambda^{jk}_{Y_2R_1} = \sum_j \lambda^{j\ell}_{Y_2R_2} = \sum_\ell \lambda^{j\ell}_{Y_2R_2} = 0$.

The models in Table 1 all handle nonignorable nonresponse in the sense that at least one of the variables has missingness which is related to the variable itself (i.e., inclusion of either $\lambda^{ik}_{Y_1R_1}$ or $\lambda^{j\ell}_{Y_2R_2}$). On the other hand, they all have different patterns of missingness in Y_1 and Y_2. Regarding the models [M1] and [M3], missingness associated with Y_1 is NMAR for both models, but missing mechanism for Y_2 is MCAR for [M1] (i.e., no interaction terms between Y_2 and missing indicators) and MAR for [M3] (i.e., inclusion of $\lambda^{i\ell}_{Y_1R_2}$). As to the models [M2] and [M4], missing mechanism for Y_2 is NMAR for both models, but missingness in Y_1 is MCAR for [M2] (i.e., no interaction terms between Y_1 and missing indicators) and MAR for [M4] (i.e., inclusion of $\lambda^{jk}_{Y_2R_1}$). Under the model [M5], NMAR missing mechanism is involved in both variables, Y_1 and Y_2, due to inclusion of $\lambda^{ik}_{Y_1R_1}$ and $\lambda^{j\ell}_{Y_2R_2}$. Note that the models [M3], [M4] and [M5] are saturated, i.e., having a degree of freedom zero.