Statistics and Probability Letters 119 (2016) 63–75

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

On the occurrence of boundary solutions in multidimensional incomplete tables

S. Ghosh *, P. Vellaisamy

Department of Mathematics, Indian Institute of Technology Bombay, India

ARTICLE INFO

Article history:
Received 19 February 2016
Received in revised form 12 July 2016
Accepted 13 July 2016
Available online 21 July 2016

Keywords:
Incomplete tables
Boundary solutions
Log-linear models
NMAR models

ABSTRACT

We provide sufficient conditions for the occurrence of boundary solutions under nonignorable nonresponse models in arbitrary three-way and n-dimensional incomplete tables with one or more variables missing. These conditions involve only the observed counts, and avoid solving likelihood equations.

1. Introduction

Incomplete contingency tables consist of (i) fully observed counts and (ii) partially classified margins (nonresponses). According to Little and Rubin (2002), three types of missing data mechanisms have been considered in the literature: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). A mechanism is said to be MCAR when the probability (of an observation being missing) is independent of both observed and unobserved data, MAR if conditional on the observed data, the probability is independent of unobserved data, and NMAR if the probability depends only on unobserved data. For likelihood inference, the missing data mechanism is classified as ignorable if it is MAR or MCAR and nonignorable if it is NMAR. The nonresponses under the mechanism are said to be ignorable or nonignorable accordingly.

Log-linear models have generally been used to analyze missing data mechanisms in incomplete contingency tables (see Baker and Laird, 1988; Baker et al., 1992; Smith et al., 1999; Clarke, 2002; Clarke and Smith, 2004, 2005). In this paper, we extend the log-linear parametrization introduced by Baker et al. (1992) for an $I \times J \times 2 \times 2$ table to three-dimensional and n-dimensional incomplete tables in general. We consider all possible cases when data on one or more of the variables are missing. Note that the problem of boundary solutions occurs in such models under the NMAR mechanism while using maximum likelihood (ML) estimation. For a $2 \times 2 \times 2$ incomplete contingency table, Baker and Laird (1988) proposed a condition for the occurrence of boundary solutions in NMAR models. For an $I \times J \times 2 \times 2$ table, Baker et al. (1992) showed that boundary solutions under NMAR models may occur when certain systems of likelihood equations are solved to obtain parameter estimates. The boundary solution problem was geometrically described by Smith et al. (1999) and Clarke (2002). Clarke and Smith (2005) described various properties of ML estimators for NMAR models when boundary solutions occur. Recently, Park et al. (2014) provided a sufficient condition for the occurrence of boundary solutions in NMAR models when data on both variables are missing in square two-way incomplete tables ($I \times I \times 2 \times 2$ tables). This condition involves only the observed counts, and thus does not require computing the maximum likelihood estimates.

* Corresponding author.
E-mail address: sayang@math.iitb.ac.in (S. Ghosh).

http://dx.doi.org/10.1016/j.spl.2016.07.011
0167-7152/

Table 1
$2 \times 2 \times 2 \times 2$ incomplete table.

			$Y_3 = 1$	$Y_3 = 2$
$R = 1$	$Y_1 = 1$	$Y_2 = 1$	y_{1111}	y_{1121}
		$Y_2 = 2$	y_{1211}	y_{1221}
	$Y_1 = 2$	$Y_2 = 1$	y_{2111}	y_{2121}
		$Y_2 = 2$	y_{2211}	y_{2221}
$R = 2$	Missing	$Y_2 = 1$	y_{+112}	y_{+122}
		$Y_2 = 2$	y_{+212}	y_{+222}

In this paper, we extend the results of Park et al. (2014) and establish sufficient conditions under several identifiable NMAR models for $I \times J \times K \times 2$, $I \times J \times K \times 2 \times 2$, $I \times J \times K \times 2 \times 2 \times 2$ and arbitrary n-dimensional incomplete tables. These conditions require only the fully observed and partially classified cell counts. Hence, they are easily verifiable and eliminate the need for using the EM algorithm (see Dempster et al., 1977) or solving likelihood equations. In Section 2, we provide log-linear parametrization for $I \times J \times K \times 2$, $I \times J \times K \times 2 \times 2$, $I \times J \times K \times 2 \times 2 \times 2$ incomplete tables and consider various identifiable NMAR models for such tables. Section 3 deals with boundary solutions, their various forms and sufficient conditions for their occurrence under the NMAR models in each of the above tables. Section 4 extends the discussions and results in Sections 2 and 3 to arbitrary n-dimensional incomplete tables. A real-life example is analyzed in Section 5 to illustrate the results in Sections 2 and 3. Section 6 provides some concluding remarks.

2. Log-linear parametrization for 3-dimensional incomplete tables

Baker et al. (1992) suggested nine identifiable log-linear models for studying missing data mechanisms in an $I \times J \times 2 \times 2$ incomplete table. In this section, we propose such hierarchical log-linear models for three-way contingency tables where data on one, two or all variables may be missing. Since boundary solutions may occur in nonignorable nonresponse models only, we consider models in which the missing mechanism is NMAR for at least one of the variables. It is assumed that partially classified (supplementary) margins are positive. Also, null fully observed counts do not appear in the likelihood function and hence in the likelihood ratio statistic. Now suppose Y_1, Y_2 and Y_3 are three categorical variables with I, J and K levels respectively. Then we have the following cases.

2.1. One of the variables is missing

Let Y_1 be missing and R denote the missing indicator for Y_1 such that $R = 1$ if Y_1 is observed and $R = 2$ otherwise. Then for Y_1, Y_2, Y_3 and R, we have an $I \times J \times K \times 2$ table with cell counts $\mathbf{y} = \{y_{ijkr}\}$, where $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$ and $r = 1, 2$. The vector of observed counts is $\mathbf{y}_{\text{obs}} = (\{y_{ijk1}\}, \{y_{+jk2}\})$ where $\{y_{ijk1}\}$ are the fully observed counts and $\{y_{+jk2}\}$ are the supplementary margins. Note that '+' denotes summation over levels of the corresponding variable. Let $\pi = \{\pi_{ijkr}\}$ be the vector of cell probabilities, $\mu = \{\mu_{ijkr}\}$ be the vector of expected counts and $N = \sum_{i,j,k,r} y_{ijkr}$ be the total cell count. For $I = J = K = 2$, we have the $2 \times 2 \times 2 \times 2$ incomplete table (Table 1).
The log-linear model (with no three-way or four-way interactions) is then given by

$$\log \mu_{ijkr} = \lambda + \lambda_{Y_1}(i) + \lambda_{Y_2}(j) + \lambda_{Y_3}(k) + \lambda_R(r) + \lambda_{Y_1Y_2}(i,j) + \lambda_{Y_1Y_3}(i,k) + \lambda_{Y_2Y_3}(j,k) + \lambda_{Y_1R}(i,r) + \lambda_{Y_2R}(j,r) + \lambda_{Y_3R}(k,r). \tag{2.1}$$

Each log-linear parameter in (2.1) satisfies the constraint that the sum over each of its arguments is 0, for example, $\sum_i \lambda_{Y_1Y_3}(i,k) = \sum_k \lambda_{Y_1Y_3}(i,k) = 0$. Define $a_{ijk} = \frac{P(R=2|Y_1=i,Y_2=j,Y_3=k)}{P(R=1|Y_1=i,Y_2=j,Y_3=k)} = \frac{\pi_{ijk2}}{\pi_{ijk1}} = \frac{\mu_{ijk2}}{\mu_{ijk1}}$, which describes the missing data mechanism of Y_1. It is the odds of Y_1 being missing. Denote a_{ijk} by $\alpha_{i..}$ or $\alpha_{.j.}$ or $\alpha_{..k}$ or $\alpha_{...}$ if it depends on i or j or k or none, respectively.

Definition 2.1. The missing mechanism of Y_1 under (2.1) is NMAR if $a_{ijk} = \alpha_{i..}$, MAR if $a_{ijk} = \alpha_{.j.}$ or $\alpha_{..k}$ and MCAR if $a_{ijk} = \alpha_{...}$.

The identifiable NMAR model in this case is given by $\alpha_{i..}$ (NMAR for Y_1).

2.2. Two of the variables are missing

Suppose Y_1 and Y_2 are missing and for $i = 1, 2$, let R_i denote the missing indicator for Y_i such that $R_i = 1$ if Y_i is observed and $R_i = 2$ otherwise. Then for Y_1, Y_2, Y_3, R_1 and R_2, we have an $I \times J \times K \times 2 \times 2$ table with cell counts $\mathbf{y} = \{y_{ijkrs}\}$, where $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$ and $r, s = 1, 2$. The vector of observed counts is $\mathbf{y}_{\text{obs}} = (\{y_{ijk11}\}, \{y_{+jk21}\}, \{y_{i+k12}\}, \{y_{++k22}\})$. Let $\pi = \{\pi_{ijkrs}\}$ be the vector of cell probabilities, $\mu = \{\mu_{ijkrs}\}$ be the vector of expected counts and N be the total cell count. For $I = J = K = 2$, we have the $2 \times 2 \times 2 \times 2 \times 2$ incomplete table (Table 2).