Identification of Nonignorable Nonresponse Mechanisms for Two Way Contingency Tables

Zhi Geng and Wen-Qing Ma
Department of Probability and Statistics, Peking University, Beijing 100871, China
E-mail: zgeng@statms.stat.pku.edu.cn

Abstract

In this paper, we use decomposable graphical models to describe the mechanisms of nonignorable nonresponse in contingency tables classified by two binary variables, and we discuss identification of parameters in these models. For an unidentifiable model, we propose adding covariates which are always observed such that this model becomes identifiable.

1. Introduction

Contingency tables with nonresponses have been discussed by many authors. According to the terminology of Little and Rubin (1987), the nonresponse mechanisms can be classified into three types: missing completely at random (MCAR), missing at random (MAR) and nonignorable nonresponse mechanisms. The MCAR and MAR nonresponse mechanisms are also said to be ignorable. Fay (1986) discussed nonignorable nonresponse mechanisms and used an indicator for each variable which is subjected to nonresponse. Baker and Laird (1988) developed a log linear model for categorical variables subject to nonresponse. Baker (1992) provided a class of models of nonresponse mechanisms with closed form estimates of cell probabilities. Molenberghs et al. (1999a, b) discussed some issues on models with incomplete categorical data, including identifying of parameters and so on. Glonek(1999) presented necessary and sufficient conditions of global identifiability for simple nonignorable nonresponse models with one or two binary responses. In this paper, we use decomposable graphical models to describe the mechanisms of nonignorable nonresponse in contingency tables classified by two binary variables, and we discuss identification of parameters in these models. For an unidentifiable model, we propose adding covariates which are always observed such that this model becomes identifiable.

Section 2 introduces some notations. Section 3 discusses identification of nonresponse mechanisms for 2×2 contingency tables subject to nonresponses. Section 4 discuss estimates of the parameters.

2. Notation

Let (Y_1, Y_2) denote the response vector with $Y_t = 0$ or 1 for $t = 1, 2$. We introduce a response indicator R_t with value 1 if the response Y_t is obtained and 0 otherwise. We use decomposable graphical models to describe the association among the four variables: Y_1,

Table 1: Observed data subject to nonresponses.

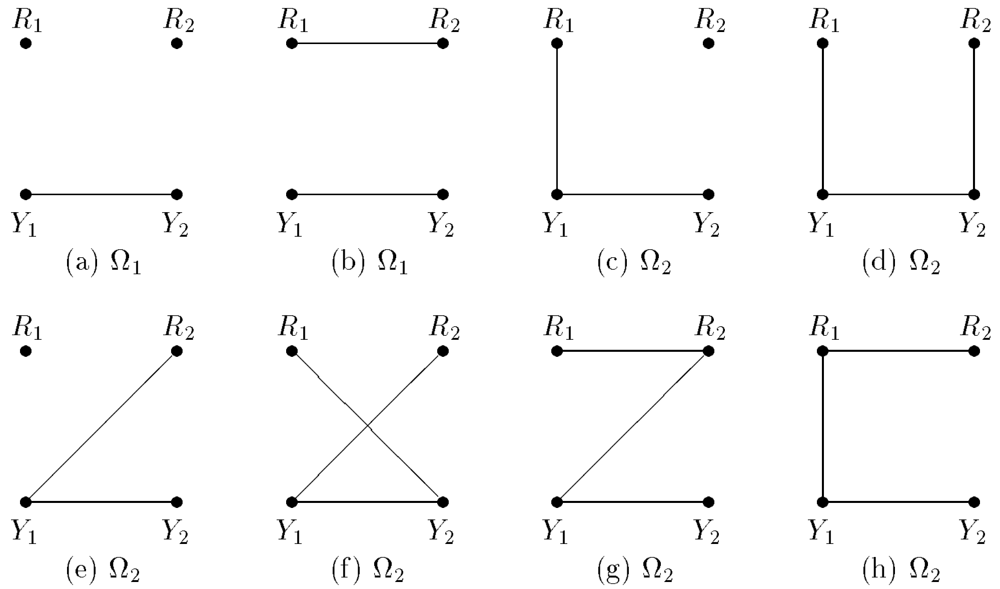
n_{0011}	n_{0111}	n_{0+10}
n_{1011}	n_{1111}	n_{1+10}
n_{+001}	n_{+101}	n_{++00}

Y_2, R_1 and R_2. Let $G = (V, E)$ denote an undirected graph where $V = \{Y_1, Y_2, R_1, R_2\}$ is the set of nodes and E are the set of undirected edges between these nodes. The absence of an edge between a pair of nodes means that the corresponding variables in this pair are independent conditionally on the other variables. A graph is decomposable if it does not has any cycle of length than or equal to 4 without a chord. A decomposable graph denotes a decomposable graphical model. Let $M(G)$ denote the set of all possible joint probabilities based on the graphical model G.

Observed data can be denoted as Table 1, where n_{ijkl} denotes the observed frequency for $Y_1 = i$, $Y_2 = j$, $R_1 = k$ and $R_2 = l$, and '+' denotes that the corresponding variable is missing. For example, n_{1+10} denotes the frequency for Y_1 is observed with value 1 but Y_2 is missing.

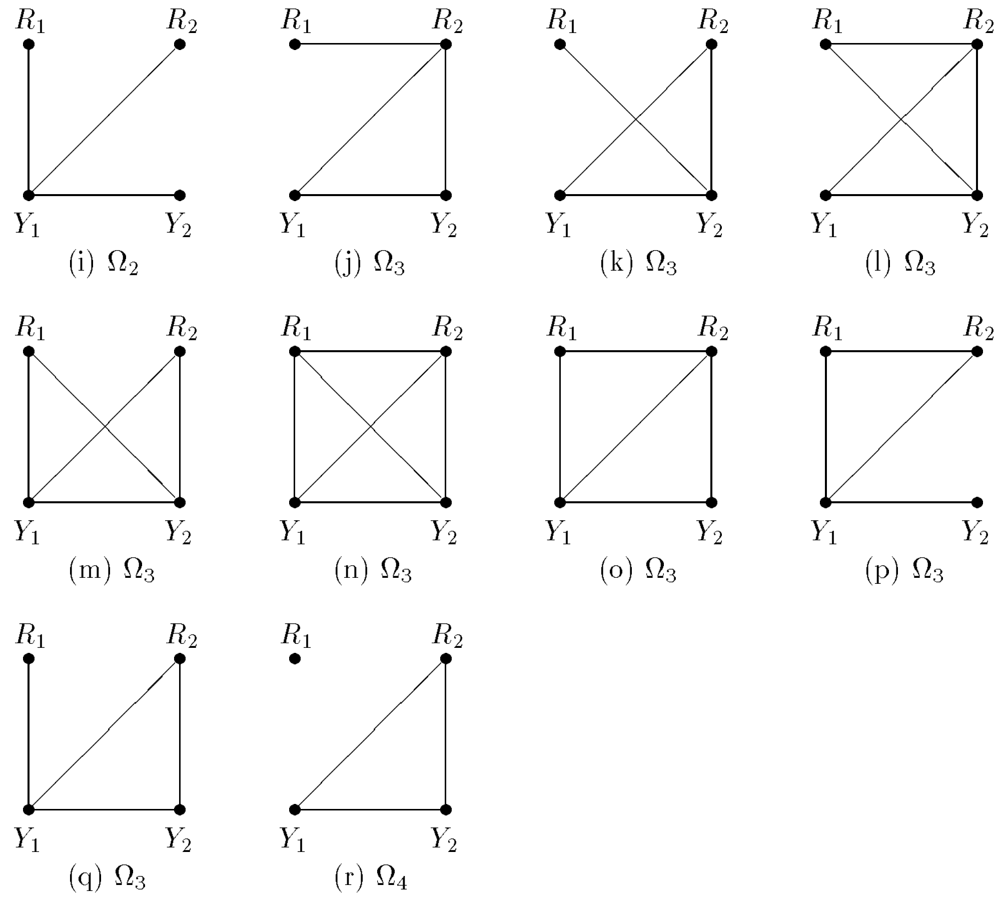
3. Identifiablity of Parameters

If any joint probability of Y_1, Y_2, R_1 and R_2 in $M(G)$ is identifiable, then we say that the model G is YR-identifable. In this paper, we only discuss decomposable graphical models. Figure (2) shows the eighteen possible decomposable graphical models.



(a) Ω_1 (b) Ω_1 (c) Ω_2 (d) Ω_2

(e) Ω_2 (f) Ω_2 (g) Ω_2 (h) Ω_2



(i) Ω_2 (j) Ω_3 (k) Ω_3 (l) Ω_3

(m) Ω_3 (n) Ω_3 (o) Ω_3 (p) Ω_3

(q) Ω_3 (r) Ω_4

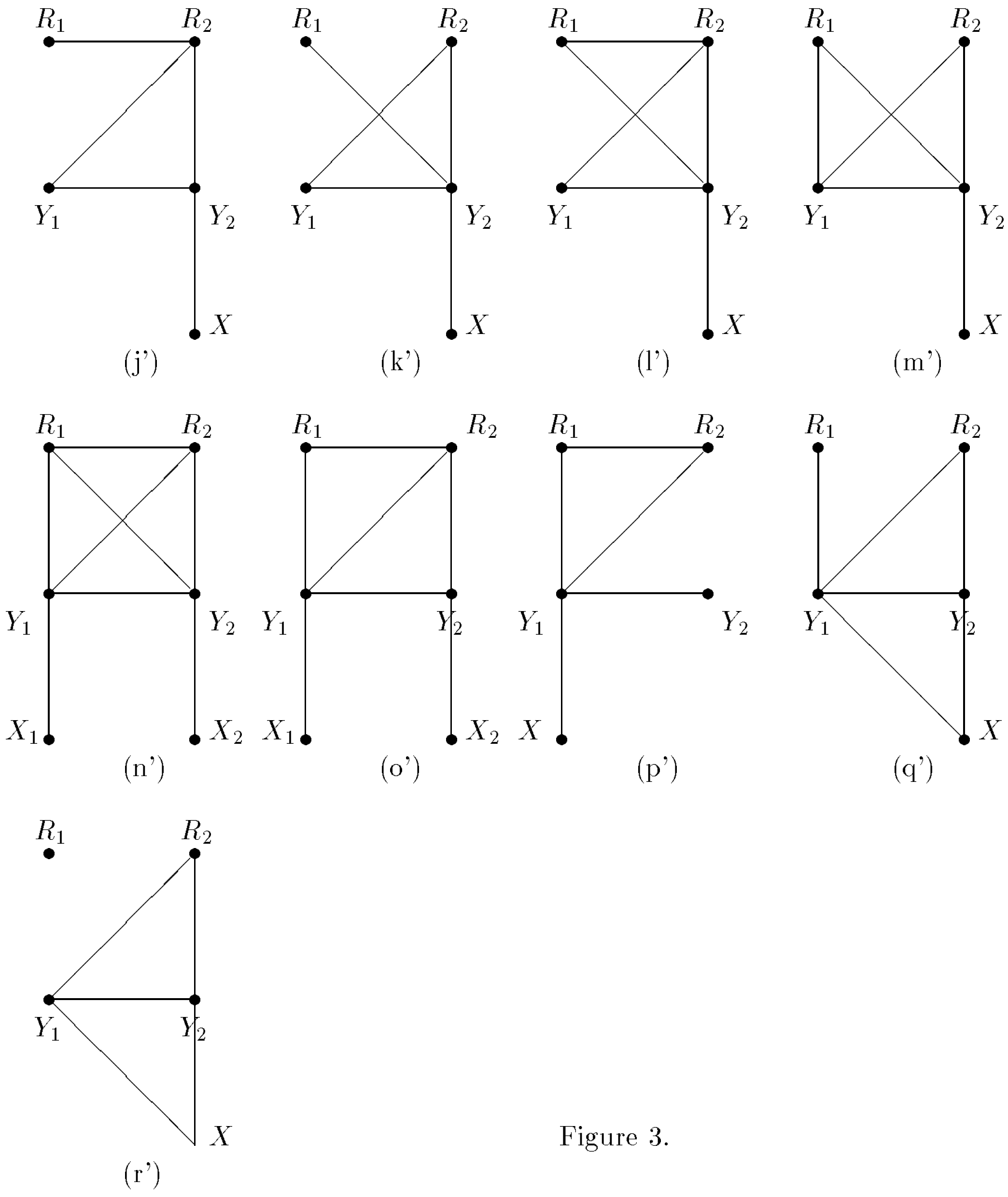
Figure 2. Eighteen composiable graphs

These eighteen graphs are classified into four types Ω_i for $i = 1, \ldots, 4$. In Ω_1, there are no edges between the response indicator R and the response variable Y. Thus the responses R_1 and R_2 does not depend on the true value of the response variables Y_1 and Y_2. The mechanism of nonresponse (a) is $MCAR$, and the mechanism (b) is MAR. Both of them are ignorable. The parameter of these two models are identifiable.

Theorem 1. For the subgraphs in the set Ω_2, we have that

1. any of graphical models (c), (d) and (h) is YR-identifiable if and only if $Y_1 \not\perp\!\!\!\perp Y_2$;
2. all graphical models (e), (f) and (g) are YR-identifiable; and
3. the graphical model (i) is YR-identifiable if and only if $Y_1 \not\perp\!\!\!\perp Y_2$ or $R_2 \not\perp\!\!\!\perp Y_1$.

There are eight observed data, but parameters in Ω_3 have more than eight degrees of freedom. Thus they are nonidentifiable. In this situation, we can introduce some covariates to identify the parameters.



(j') (k') (l') (m')

(n') (o') (p') (q')

(r')

Figure 3.

For graphs (j), (k), (l) and (m), we introduce a covariate which is always observed with two values such that $X \perp\!\!\!\perp (Y_1, R_1, R_2)|Y_2$, as shown in Figure 3 (j'), (k'), (l') and (m').

Theorem 2. If $X\not\!\perp\!\!\!\perp Y_2$, then the graphical models (j'), (k'), (l')and (m') are YRX-identifiable (*i.e.*, $P(y_1, r_1, y_2, r_2, x)$ is identifiable).

For graphs (n)and (o), we introduce a binary covariate X_1 such that $X_1 \perp\!\!\!\perp (Y_2, R_1, R_2)|Y_1$, and another binary covariate X_2 such that $X_2 \perp\!\!\!\perp (Y_1, R_1, R_2)|Y_2$, as shown in figure 3.

Theorem 3. If $X_2 \not\!\perp\!\!\!\perp Y_2$ and $X_1 \not\!\perp\!\!\!\perp Y_1$, then the graphical models (n') and (o') are YRX_1X_2- identifiable (*i.e.*, $P(y_1, r_1, y_2, r_2, x_1, x_2)$ is identifiable).

For graph (p), we introduce a always observed variable X with two values such that $X \perp\!\!\!\perp (R_1, R_2, Y_2)|Y_1$. We have graph (p') as shown in figure 3.

Theorem 4. If $X \not\!\perp\!\!\!\perp Y_1$, then the graphical model (p') is YRX-identifiable(*i.e.*, $P(y_1, r_1, y_2, r_2, x)$ is identifiable).

For the graphs (q) and (r), we introduce a binary covariate X as shown in figure 3 (q') and (r'), where X satisfy the condition that $X \not\!\perp\!\!\!\perp (R_1, R_2)|(Y_1, Y_2)$.

Theorem 5 If $X \not\!\perp\!\!\!\perp Y_2|Y_1$, then the graphical models (q') and (r') are YRX-identifiable (*i.e.*, $P(y_1, r_1, y_2, r_2, x)$ is identifiable).

An interest measure of association is the odds ratio in the margin table classified by Y_1 and Y_2 defined as

$$OR_{..} = \frac{P(Y_1 = 0, Y_2 = 0)P(Y_1 = 1, Y_2 = 1)}{P(Y_1 = 0, Y_2 = 1)P(Y_1 = 1, Y_2 = 0)}.$$

The odds ratio $OR_{11} = P(Y_1 = 0, Y_2 = 0|R_1 = 1, R_2 = 1)P(Y_1 = 1, Y_2 = 1|R_1 = 1, R_2 = 1)/[P(Y_1 = 0, Y_2 = 1|R_1 = 1, R_2 = 1)P(Y_1 = 1, Y_2 = 0|R_1 = 1, R_2 = 1)]$ describes the association in the complete response subpopulation with $R_1 = R_2 = 1$. Assume that the complete response probability $P(R_1 = R_2 = 1)$ is positive and larger than a given constant.

Theorem 6 For the graphical models in Ω_1, Ω_2 and the graph (p) in Ω_3, we have that $OR_{..} = OR_{11}$ and that $O\hat{R}_{11} = n_{0011}n_{1111}/(n_{1011}n_{0111})$ is a strong consistent estimate of $OR_{..}$, where n_{ijkl} is the observed frequency of $Y_1 = i$, $Y_2 = j$, $R_1 = k$ and $R_2 = l$.

REFERENCES

1. S.G. Baker, Closed-form estimates for missing counts in two-way contingency tables. *Statist. in Medicine.***11**1992, 643-657.

2. S.G.Baker & N.M.Laird Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. it J. Am. Statist. Assoc. **83** (1988), 62–69.

3. R. L. Chambers, and A. H. Welsh, Log-linear models for survey data with non-ignorable non-response. *J. R. Statist. Soc.* **B 55** (1993), 157–170.

4. D. R. Cox and N. Wermuth, *Multivariate Dependencies: Models, analysis and interpretation.* Chapman & Hall, London, 1996.

5. P. J. Diggle and M. G. Kenward Informative dropout in longitudinal data analysis(with discussion). *Appl. Statist.* **43** (1994), 49–93.

6. R. E. Fay, Causal models for patterns of nonresponse. *J. Am. Statist. Assoc.* **81** (1986), 354–365.

7. G. M. Fitzmaurice, N. M. Laird and G. E. P. Zahner, Multivariate logistic models for incomplete binary responses. *J. Am. Statist. Assoc.* **91** (1996), 99–108.

8. G. F. V. Glonek, On identifiability in models for incomplete binary data. *Statist. & Probabi. Letters* **41** (1999), 191–197.

9. S. L. Lauritzen, *Graphical models*, Oxford University Press, Oxford, 1996.

10. R. J. A. Little, Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bull. Int. Statist. Inst.*, **15**, no. 1 (1985), 1–15.

11. R. J. A. Little, Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Assoc.* **88** (1993), 125–134.

12. R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

13. J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, 1988.

14. A. Rotnitzky, J. M. Robins and D. O. Scharfstein, Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Statist. Assoc.* **93** (1998), 1321–1339.

15. T. J. Rothenberg, Identification in parametric models. *Econometrica* **39** (1971), 577–591.