Technical report RT–MAE–2007–07. Instituto de Matemática e Estatística, Universidade de São Paulo.

A product-multinomial framework for categorical data analysis with missing responses

Frederico Z. Poleto[1,†], **Julio M. Singer**[1,‡] and **Carlos Daniel Paulino**[2,§]

[1]Instituto de Matemática e Estatística, Universidade de São Paulo,
Caixa Postal 66281, São Paulo, SP, 05311-970, Brazil
[2]Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal
[†]fred@poleto.com [‡]jmsinger@ime.usp.br [§]dpaulino@math.ist.utl.pt

Abstract

We extend the multinomial modeling scenario for the analysis of categorical data with missing responses described by Paulino (1991, *Brazilian Journal of Probability and Statistics*, **5**, 1-42) to the product-multinomial setup so that the inclusion of explanatory variables is allowed. Assuming an ignorable missing data mechanism, linear and log-linear models may be fitted via maximum likelihood. Weighted least squares methodology may as well be used to fit more general functional linear models, if a missing completely at random mechanism is assumed. We also consider a hybrid approach, where any missingness process is fitted by maximum likelihood in a first step, and the estimated marginal probabilities of categorization and their covariance matrix are used in a second stage to fit the model via weighted least squares, in the spirit of functional asymptotic regression methodology. Goodness-of-fit tests are present, and the methodology is illustrated via two data sets. All the methods were computationally implemented via subroutines written in R.

Key words: Categorical and, missing or incomplete data; MAR, MCAR and MNAR; Ignorable and non-ignorable mechanism; Selection models; Linear, log-linear and functional linear models.

1 Introduction

Appropriate methods for drawing inferences in the presence of a partial classification in categorical data analyses have been proposed since, at least, the works from Blumenthal (1968) and Hocking and Oxspring (1971). The EM algorithm (Dempster, Laird and Rubin, 1977) took over an important place in the statistical inference, and it has been used in the present context to obtain maximum likelihood (ML) estimates based on expected cell frequencies in augmented tables; see, *e.g.*, Fuchs (1982) and Baker and Laird (1988) for, respectively, ignorable and non-ignorable models for the missingness mechanism. Molenberghs and Goetghebeur (1997) showed advantages

of using Newton-Raphson and Fisher scoring algorithms, and Baker (1994) suggested a combination of EM and Newton-Raphson algorithms. Although these methods embrace a part of the models described in this paper, they are still not available for ready use in the current statistical software, because need further calculations of derivatives over the assumed model, adaptations of the available computational procedures, and/or additional programming. Some exceptions are multiple imputation methods (Rubin, 1987) — available in SAS (PROC MI and MIANALYZE) and R/S-Plus (package mitools) —, and saturated and hierarchical log-linear multinomial models (Schafer, 1997) — available in R/S-Plus (package cat).

Paulino (1991) considered fitting strictly linear and log-linear multinomial models under the ignorability assumption by ML, and functional linear models under the missing completely at random assumption via weighted least squares (WLS) methodology. He also proposed a hybrid methodology, where ignorable mechanisms are fitted by ML in the first step, and the estimated marginal probabilities of categorization and their covariance matrix are used in a second stage to fit the model via WLS, in the spirit of functional asymptotic regression methodology described by Imrey, Koch, Stokes *et al.* (1981, 1982) for complete data. In most cases, this approach is computationally simpler than and asymptotically as efficient as the pure ML approach.

In this paper, we extend the theory to cover product-multinomial models and present the results in matrix formulation. We programmed the theory using R software (R Development Core Team, 2006). The required computations are automatically conducted by the designed functions when MAR or MCAR mechanisms are considered. For missing not at random mechanisms, the first step must be programmed by the user, by means of one of the built-in optimization functions in the R software, in order to obtain the ML estimates. The model formulation and usage of the functions are similar to the old GENCAT (Landis, Stanish, Freeman and Koch, 1976), later incorporated by the SAS' PROC CATMOD, but allows us to analyze complete and incomplete categorical data in a unified way. The programmed functions may be downloaded from http://www.poleto.com/missing.html.

In Section 2, we describe the problem and the notation. In Section 3, we present the probabilistic model and the missingness mechanisms. In Section 4, we show the inferential results for ML and WLS approaches without imposing constrains to the probabilities of categorization. In Section 5, we describe the ML methodoloy for linear and log-linear structural models, and the WLS approach for the functional linear ones.

2 Problem description and notation

For simplicity, we admit that the random vector $\mathbf{Y}$ of response variables can assume R values $\mathbf{y}$, corresponding to combinations of the levels of its components $Y_1, Y_2, \ldots, Y_k$. For instance, when $\mathbf{Y} = (Y_1, Y_2, Y_3)'$, and Y_1, Y_2 and Y_3 may assume, respectively, 2, 3 and 5 different values, $R = 2 \times 3 \times 5 = 30$. Likewise, we assume that the vector $\mathbf{X}$ of explanatory variables can take S values $\mathbf{x}$, corresponding to combinations of the levels of its components $X_1, X_2, \ldots, X_q$. The R response categories are indexed by r, and the S subpopulations, by s.

We assume that each one of the n_{s++} sampling units randomly selected from the s-th subpopulation can be independently classified into the r-th response category with the same probability $\theta_{r(s)}$, $r = 1, \ldots, R$, $s = 1, \ldots, S$. This indicates that the $n_{+++} = \sum_{s=1}^{S} n_{s++}$ units follow a stratified random sampling with allocation scheme along the strata in conformity with the vector $\mathbf{N}_{++} = (n_{1++}, \ldots, n_{S++})'$.

For several reasons, it may not be possible to completely observe the responses of all variables from $\mathbf{Y}$. In these cases, only part of the n_{s++} sampling units is classified into one of the R originally defined response categories, while the remaining units are associated to some type of missingness. For subpopulation s, $s = 1, \ldots, S$, we define T_s missingness patterns in the following way. The set of units with no missing data (*i.e.*, complete classification) is represented by $t = 1$, and the sets that have some degree of missingness, by $t = 2, \ldots, T_s$. We admit that the units corresponding to the t-th missingness pattern, $t = 2, \ldots, T_s$, are recorded in response classes $\mathcal{C}_{stc}$, $c = 1, \ldots, R_{st}$, formed by at least two of the R response categories, with $\mathcal{C}_{stc} \cap \mathcal{C}_{std} = \emptyset$, $c \neq d$ and $\cup_{c=1}^{R_{st}} \mathcal{C}_{stc} = \{1, \ldots, R\}$. Thus, each one of the $t = 2, \ldots, T_s$ missingness patterns form partitions $\mathcal{P}_{st} = \{\mathcal{C}_{stc}, c = 1, \ldots, R_{st}\}$ of the complete classification pattern $\mathcal{P}_{s1} = \mathcal{P}_1 = \big\{\{r\}, r = 1, \ldots, R\big\}$, and R_{st} denotes the number of response classes with the t-th missingness pattern for the s-th subpopulation. For consistence of the notation, we suppose that the complete classification pattern has classes equivalent to the R response categories, *i.e.*, $\mathcal{C}_{s1r} = \mathcal{C}_{1r} = \{r\}$, $r = 1, \ldots, R$ and $R_{s1} = R_1 = R$. We represent the number of classes with the $T_s - 1$ missingness patterns for the s-th subpopulation by $l_s = \sum_{t=2}^{T_s} R_{st}$.

For mathematical convenience, we build $R \times 1$ dimensional vectors $\mathbf{z}_{stc}$ with elements equal to 1 associated to the response categories pertaining to the class $\mathcal{C}_{stc}$, and the others elements equal to 0; $\mathbf{Z}_{st} = [\mathbf{z}_{stc}, c = 1, \ldots, R_{st}]$, an $R \times R_{st}$ matrix, encloses the indicator vectors of all classes of the t-th missingness pattern of the s-th subpopulation; and $\mathbf{Z}_s = [\mathbf{Z}_{st}, t = 1, \ldots, T_s]$, an $R \times (R + l_s)$ matrix, includes the indicator vectors from all classes of all missingness patterns of the s-th subpopulation. Note that $\mathbf{Z}_{s1} = \mathbf{I}_R$ (identity matrix of order R), $s = 1, \ldots, S$. The observable frequencies, $\{n_{stc}\}$,

indicate the units from the s-th subpopulation with the t-th missingness pattern classified into the c-th class, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$, $c = 1, \ldots, R_{st}$. The vector $\mathbf{N}_{st} = (n_{stc}, c = 1, \ldots, R_{st})'$ stacks the observable frequencies of the t-th scenario of the s-th subpopulation, $\mathbf{N}_s = (\mathbf{N}'_{st}, t = 1, \ldots, T_s)'$ encloses all the observable frequencies of the s-th subpopulation, $\mathbf{N} = (\mathbf{N}'_s, s = 1, \ldots, S)'$ includes all the observable frequencies, and $n_{st+} = \sum_{c=1}^{R_{st}} n_{stc}$ indicates the total units selected from the s-th subpopulation with the t-th missingness pattern. Note that the substitution of any subscript by "+" indicates the sum of the values over that particular subscript.

We assume that a sampling unit selected from the s-th subpopulation with the r-th response category is classified into the t-th missingness pattern with probability $\lambda_{t(rs)}$, $r = 1, \ldots, R$, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$. The $\{\lambda_{t(rs)}\}$ are the conditional probabilities of missingness, and the $\{\theta_{r(s)}\}$ are the marginal probabilities of categorization. We assume that there are no missing values in $\mathbf{X}$.

The notation will now be clarified.

Example 1 A subset of the Six Cities study (Ware *et al.*, 1984) used by Lipsitz and Fitzmaurice (1996) is presented in Table 1.

Table 1: Observed frequencies

Home city	Maternal smoking	Child's wheezing status			
		no wheeze	wheeze with cold	wheeze apart from cold	missing
Kingston-Harriman	none	167	17	19	176
	moderate	10	1	3	24
	heavy	52	10	11	121
	missing	28	10	12	
Portage	none	120	22	19	103
	moderate	8	5	1	3
	heavy	39	12	12	80
	missing	31	8	14	

We represent the subpopulation of Kingston-Harriman by the index $s = 1$, and the one of Portman by $s = 2$. The index r of the response categories follows a lexicographical order. The missingness pattern when the child's wheezing status is missing is denoted by $t = 2$, and when the maternal smoking is missing, by $t = 3$. In the complete classification pattern ($t = 1$), there are $R_{s1} = R = 9$ classes/categories, $\mathcal{P}_{s1} = \big\{\{1\}, \{2\}, \ldots, \{9\}\big\}$, $\mathbf{Z}_{s1} = \mathbf{I}_9$, $s = 1, 2$, $\mathbf{N}_{11} = (167, 17, 19, 10, 1, 3, 52, 10, 11)'$, $n_{11+} = 290$, $\mathbf{N}_{21} = (120, 22, 19, 8, 5, 1, 39, 12, 12)'$, and $n_{21+} = 238$. For both

cities, the patterns $t = 2$ have $R_{s2} = 3$ classes, $\mathcal{P}_{s2} = \big\{\{1,2,3\},\{4,5,6\},\{7,8,9\}\big\}$,

$$\mathbf{Z}_{s2} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}' = \mathbf{I}_3 \otimes \mathbf{1}_3,$$

$s = 1, 2$, where $\otimes$ denotes the Kronecker product, and $\mathbf{1}_k$ represents a $k \times 1$ vector with all elements equal to 1; $\mathbf{N}_{12} = (176, 24, 121)'$, $n_{12+} = 321$, $\mathbf{N}_{22} = (103, 3, 80)'$, $n_{22+} = 186$, and the patterns $t = 3$, $R_{s3} = 3$ classes, $\mathcal{P}_{s3} = \big\{\{1,4,7\},\{2,5,8\},\{3,6,9\}\big\}$,

$$\mathbf{Z}_{s3} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}' = \mathbf{1}_3 \otimes \mathbf{I}_3,$$

$s = 1, 2$, $\mathbf{N}_{13} = (28, 10, 12)'$, $n_{13+} = 50$, $\mathbf{N}_{23} = (31, 8, 14)'$, $n_{23+} = 53$. Additionally, $l_s = R_{s2} + R_{s3} = 6$, $\mathbf{N}_s = (\mathbf{N}'_{s1}, \mathbf{N}'_{s2}, \mathbf{N}'_{s3})'$, $\mathbf{Z}_s = [\mathbf{Z}_{s1}, \mathbf{Z}_{s2}, \mathbf{Z}_{s3}]$, $s = 1, 2$, $\mathbf{N}_{++} = (n_{1++}, n_{2++})' = (661, 477)'$, $n_{+++} = 1\,138$, and $\mathbf{N} = (\mathbf{N}'_1, \mathbf{N}'_2)'$.

Note that, in this example, the observed missingness patterns for both cities were the same. In other cases, R_{st}, $\mathcal{C}_{stc}$, $\mathcal{P}_{st}$, l_s, $\mathbf{z}_{stc}$, $\mathbf{Z}_{st}$ and $\mathbf{Z}_s$ would not necessarily be equal for $s = 1, 2$.

Example 2 The data set in Table 2 was analyzed by Soares and Paulino (2001). Differently from the usual missingness pattern caused only by incomplete classification into marginal tables, the missingness in this case is originated by embarrassment of neighbor cells.

Table 2: Observed frequencies of the risk degree to dental caries

Simplified test	Standard test		
	high	medium	low
high	7	11	2
medium	3	9	5
low	0	10	4
high / medium	8	7	3
medium / low	7	14	7

As there are no subpopulations, the index s is dropped. The index r follows the same order as in the previous example. We represent the missingness pattern without distinction between the categories high and medium (medium and low) of the simplified test by $t = 2$ ($t = 3$). In the absence of missingness pattern, $t = 1$, where the units are completely categorized into one of the $R_1 = R = 9$ classes/categories, $\mathcal{P}_1 = \big\{\{r\}, r = 1, \ldots, 9\big\}$, $\mathbf{Z}_1 = \mathbf{I}_9$, $\mathbf{N}_1 = (7, 11, 2, 3, 9, 5, 0, 10, 4)'$ and $n_{1+} = 51$. The pattern $t = 2$ can be encased in the partitions context associating the classes $\mathcal{C}_{21} = \{1, 4\}$, $\mathcal{C}_{22} = \{2, 5\}$, $\mathcal{C}_{23} = \{3, 6\}$ and $\mathcal{C}_{24} = \{7, 8, 9\}$ to the frequencies $n_{21} = 8$, $n_{22} = 7$,

$n_{23} = 3$ and $n_{24} = 0$. Note that the definition of the last class is an artifice to, jointly with the other classes, constitute a partition of the set of response categories. Hence, we have $R_2 = 4$ classes, $\mathcal{P}_2 = \big\{\{1,4\},\{2,5\},\{3,6\},\{7,8,9\}\big\}$,

$$\mathbf{Z}_2 = [\mathbf{z}_{21}, \mathbf{z}_{22}, \mathbf{z}_{23}, \mathbf{z}_{24}] = \left[\begin{array}{cccccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right]' = \left[\begin{array}{cc} \mathbf{1}_2 \otimes \mathbf{I}_3 & \mathbf{0}_6 \\ \mathbf{0}_{3,3} & \mathbf{1}_3 \end{array}\right],$$

$\mathbf{N}_2 = (8,7,3,0)'$, and $n_{2+} = 18$, where $\mathbf{0}_k$ denotes a $k \times 1$ vector with all elements equal to 0, and the $j \times k$ matrix $\mathbf{0}_{j,k}$ has all null elements. Likewise, in the pattern $t = 3$, we have $R_3 = 4$ classes, $\mathcal{P}_3 = \big\{\{1,2,3\},\{4,7\},\{5,8\},\{6,9\}\big\}$,

$$\mathbf{Z}_3 = [\mathbf{z}_{31}, \mathbf{z}_{32}, \mathbf{z}_{33}, \mathbf{z}_{34}] = \left[\begin{array}{ccc|ccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right]' = \left[\begin{array}{cc} \mathbf{1}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_6 & \mathbf{1}_2 \otimes \mathbf{I}_3 \end{array}\right],$$

$\mathbf{N}_3 = (0,7,14,7)'$, and $n_{3+} = 28$. Then, $l = R_2 + R_3 = 8$, $\mathbf{N} = (\mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$, $n_{++} = 97$, and $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3]$.

Through these examples, note that the conditions $\mathcal{C}_{stc} \cap \mathcal{C}_{std} = \emptyset$, $c \neq d$ and $\cup_{c=1}^{R_{st}} \mathcal{C}_{stc} = \{1, \ldots, R\}$ can be verified by letting $\mathbf{Z}_{st}$ has one element equal to 1 in exactly one unique column for each row, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$.

3 Probability model and missingness mechanisms

We assume that the observable frequencies $\mathbf{N}$ have a product-multinomial distribution expressed by the probability mass function

$$P\left(\mathbf{N}|\boldsymbol{\theta}, \{\lambda_{t(rs)}\}, \mathbf{N}_{++}\right) = \prod_{s=1}^{S} \frac{n_{s++}!}{\prod_{t=1}^{T_s}\prod_{c=1}^{R_{st}} n_{stc}!} \prod_{r=1}^{R} \left(\theta_{r(s)}\lambda_{1(rs)}\right)^{n_{s1r}} \prod_{t=2}^{T_s}\prod_{c=1}^{R_{st}} \left(\sum_{r \in \mathcal{C}_{stc}} \theta_{r(s)}\lambda_{t(rs)}\right)^{n_{stc}}, \quad (1)$$

where $\theta_{r(s)}$ is the probability of a sampling unit from the s-th subpopulation presenting the r-th response category; $\lambda_{t(rs)}$ is the probability of a sampling unit with the r-th response category selected from the s-th subpopulation presenting the t-th missingness pattern, $\boldsymbol{\theta} = (\boldsymbol{\theta}_s', s = 1, \ldots, S)'$, $\boldsymbol{\theta}_s = \big(\theta_{r(s)}, r = 1, \ldots, R\big)'$, $\sum_{r=1}^{R} \theta_{r(s)} = 1$, $s = 1, \ldots, S$, and $\sum_{t=1}^{T_s} \lambda_{t(rs)} = 1$, $r = 1, \ldots, R$, $s = 1, \ldots, S$. This factorization into a marginal model for the measurements, $\{\theta_{r(s)}\}$, and a conditional model for the missingness process given the measurements, $\{\lambda_{t(rs)}\}$, corresponds to the so-called selection model framework (Little and Rubin, 2002).

If it were possible to identify the response category of every observation in each of the missingness patterns, y_{str} would be the hypothetical number of sampling units from the s-th subpopulation with the t-th missingness pattern classified into the r-th response category, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$, $r = 1, \ldots, R$. Hence, $\{y_{str}\}$ denote the augmented frequencies, which are observed only under the missingness pattern $t = 1$ (no missing data), where $n_{s1r} = y_{s1r}$, $s = 1, \ldots, S$, $r = 1, \ldots, R$. Under the other patterns such frequencies are non-observable, and we know just the frequencies associated to the response classes $\mathcal{C}_{stc}$, namely

$$n_{stc} = \sum_{r \in \mathcal{C}_{stc}} y_{str}, \qquad s = 1, \ldots, S,\ t = 2, \ldots, T_s,\ c = 1, \ldots, R_{st}. \tag{2}$$

For each subpopulation, there are $R-1$ parameters $\{\theta_{r(s)}\}$, and $R(T_s-1)$ parameters $\{\lambda_{t(rs)}\}$, making up $R\,T_s - 1$ linearly independent parameters. Likewise, there are R observed frequencies in the complete classification pattern, and other l_s ones in the patterns with some missingness. As n_{s++} was previously fixed, there is a total of $R-1+l_s$ linearly independent observed frequencies in each subpopulation. Therefore, the $R\sum_{s=1}^{S} T_s - S$ linearly independent parameters $\{\theta_{r(s)}, \lambda_{t(rs)}\}$ (associated to the augmented frequencies $\{y_{str}\}$) when faced with the $S(R-1)+\sum_{s=1}^{S} l_s$ linearly independent observable frequencies $\{n_{stc}\}$ (associated to the parameters $\left\{\sum_{r \in \mathcal{C}_{stc}} \theta_{r(s)}\lambda_{t(rs)}\right\}$) highlight an overparametrization of (1) with $\sum_{s=1}^{S}[R(T_s-1) - l_s]$ non-identifiable parameters.

As the interest usually lies in $\{\theta_{r(s)}\}$, reduced structures are considered for the conditional probabilities of missingness to render the model identifiable. The most common way to overcome this problem is by assuming a non-informative missingness mechanism or, according to Rubin (1976), a missing at random (MAR) mechanism, expressed by

$$\mathbf{MAR}: \lambda_{t(rs)} = \alpha_{t(cs)}, \qquad s = 1, \ldots, S,\ t = 1, \ldots, T_s,\ c = 1, \ldots, R_{st},\ r \in \mathcal{C}_{stc}, \tag{3}$$

indicating that the conditional probabilities of missingness depend only on the observed response classes and, conditionally on these, they do not depend on the unobserved response categories. The statistical model under the MAR mechanism is saturated, and the likelihood function can be factorized as

$$L\left(\boldsymbol{\theta}, \{\alpha_{t(cs)}\}\,|\mathbf{N}; \mathrm{MAR}\right) \propto L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\right) L_2\left(\{\alpha_{t(cs)}\}\,|\mathbf{N}; \mathrm{MAR}\right), \tag{4}$$

where

$$L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\right) = \prod_{s=1}^{S}\prod_{r=1}^{R} \theta_{r(s)}^{n_{s1r}} \prod_{t=2}^{T_s}\prod_{c=1}^{R_{st}} \left(\mathbf{z}'_{stc}\boldsymbol{\theta}_s\right)^{n_{stc}}$$

and

$$L_2\left(\{\alpha_{t(cs)}\}\,|\mathbf{N}; \mathrm{MAR}\right) = \prod_{s=1}^{S}\prod_{t=1}^{T_s}\prod_{c=1}^{R_{st}} \alpha_{t(cs)}^{n_{stc}}.$$

The missing completely at random (MCAR) mechanism, a special case of the MAR mechanism, is defined by

$$\textbf{MCAR} : \lambda_{t(rs)} = \alpha_{t(s)}, \qquad s = 1, \ldots, S,\ t = 1, \ldots, T_s,\ r = 1, \ldots, R, \tag{5}$$

and implies that the conditional probabilities of missingness do not depend on the response categories, being or not partially observed. The statistical model under the MCAR mechanism has $S(R-2) + \sum_{s=1}^{S} T_s$ linearly independent parameters, since there are $T_s - 1$ parameters $\{\alpha_{t(s)}\}$ in each subpopulation, that jointly with the $R-1$ parameters $\{\theta_{r(s)}\}$, add up $R-2+T_s$ parameters. Subtracting the $S(R-2) + \sum_{s=1}^{S} T_s$ parameters from the $S(R-1) + \sum_{s=1}^{S} l_s$ observable frequencies, there are, under this missingness mechanism, $S + \sum_{s=1}^{S}(l_s - T_s)$ degrees of freedom in the likelihood function, which is expressed by

$$L\left(\boldsymbol{\theta}, \{\alpha_{t(s)}\} \,|\mathbf{N}; \text{MCAR}\right) \propto L_1\left(\boldsymbol{\theta} \,|\mathbf{N}\right) L_2\left(\{\alpha_{t(s)}\} \,|\{n_{st+}\}; \text{MCAR}\right), \tag{6}$$

where $L_1\left(\boldsymbol{\theta} \,|\mathbf{N}\right)$ has the same definition as (4) and

$$L_2\left(\{\alpha_{t(s)}\} \,|\{n_{st+}\}; \text{MCAR}\right) = \prod_{s=1}^{S} \prod_{t=1}^{T_s} \alpha_{t(s)}^{n_{st+}}.$$

This implies that the inferences about $\boldsymbol{\theta}$ can be based only on the distribution of $\mathbf{N}$ conditionally on $\{n_{st+}\}$ and the MCAR missingness mechanism can be ignored for likelihood and frequentist inferences, as is discussed by Paulino (1991) in the multinomial setting. Although, the MAR missingness mechanism is ignorable for likelihood but not for frequentist inferences on $\boldsymbol{\theta}$. See also Kenward and Molenberghs (1998) for a practical illustration that estimation of the Fisher information becomes biased when the missingness process under the MAR mechanism is ignored.

Note that the conditional probabilities of missingness defined in the MAR and the MCAR mechanisms depend upon the explanatory variables. Little (1995) believes that the term MCAR should be reserved for the case where the missingness depends neither on the response and nor on the explanatory variables, *i.e.*, when $\lambda_{t(rs)} = \alpha_t$. He also suggests to use the expression "covariate-dependent missingness" (we adapted the term covariate-dependent dropout, since the paper from Little deals only with dropout modeling) when the missingness mechanism does not depend on responses observed or missing, but depends on the explanatory variables. In this paper, we use the definition $\lambda_{t(rs)} = \alpha_{t(s)}$ for the MCAR mechanism for being the most direct generalization when we go from the multinomial distribution to the product-multinomial distribution, and because it is more general than the Little's MCAR, that can be viewed as a special case. More parsimonious structures can be proposed under both mechanisms, allowing that the conditional probabilities

of missingness do not vary for some or all subpopulations. We disregard these additional constraints, since they do not modify the estimates of $\boldsymbol{\theta}$, our main interest, due to the likelihood factorizations under the MAR and the MCAR mechanisms in a part referring to $\boldsymbol{\theta}$, denoted by $L_1(\boldsymbol{\theta}|\mathbf{N})$, and another concerning the conditional probabilities of missingness, represented by $L_2\left(\{\alpha_{t(cs)}\}|\mathbf{N};\text{MAR}\right)$ in the MAR case or $L_2\left(\{\alpha_{t(s)}\}|\{n_{st+}\};\text{MCAR}\right)$ in the MCAR case.

Missing not at random (MNAR) or informative missingness mechanisms can be formulated by assuming that at least two conditional probabilities of missingness of response categories pertaining to the same class are not equal, *i.e.*, $\{a,b\} \in \mathcal{C}_{stc}$ and $\lambda_{t(as)} \neq \lambda_{t(bs)}$. Nevertheless, it is necessary to specify at least $\sum_{s=1}^{S}[R(T_s-1)-l_s]$ parametric constraints to obtain an identifiable structure. For instance, in Example 1, we may assume that the conditional probabilities of missingness depend only on the home city and the missing result. Incorporation of the constrains $\lambda_{2(ijs)} = \lambda_{2(js)}$ and $\lambda_{3(ijs)} = \lambda_{3(is)}$ in the likelihood function of the corresponding product-multinomial distribution leads to a saturated statistical model, where, instead of the index r, we use i to represent the maternal smoking level and j to denote the child's wheezing status. MNAR mechanisms are not ignorable for likelihood or frequentist inferences on $\boldsymbol{\theta}$. As the likelihood function can not be factorized as in the MAR or the MCAR cases and the ML estimators of $\boldsymbol{\theta}$ and $\{\lambda_{t(rs)}\}$ are non-orthogonal, the covariance matrix of the ML estimator of $\boldsymbol{\theta}$ must be extracted from the corresponding component of the covariance matrix of the estimators of $\boldsymbol{\theta}$ and $\{\lambda_{t(rs)}\}$. See Molenberghs, Kenward and Goetghebeur (2001) for a discussion concerning sensitivity analysis of missingness mechanisms.

4 Saturated structural models for the marginal probabilities of categorization

As the sampling units of a total missingness pattern, represented by $\mathcal{P}_{st} = \{\mathcal{C}_{st1}\} = \{\{1,\ldots,R\}\}$, do not carry any information about $\boldsymbol{\theta}$ under the MAR and the MCAR mechanisms, we ignore these missingness patterns and redefine T_s as the number of partial missingness patterns, and n_{s++} as the number of sampling units with some type of categorization, when dealing with these mechanisms.

To simplify the presentation of the results by matrix operations, we introduce some additional notation: $\bar{\boldsymbol{\theta}}_s = [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]\ \boldsymbol{\theta}_s = \left(\theta_{r(s)}, r = 1,\ldots,R-1\right)'$ contains the $R-1$ first components of $\boldsymbol{\theta}_s$, $s = 1,\ldots,S$; $\bar{\boldsymbol{\theta}} = (\mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}])\,\boldsymbol{\theta} = \left(\bar{\boldsymbol{\theta}}'_s, s = 1,\ldots,S\right)'$; $\bar{\mathbf{Z}}_{st}$, an $(R-1)\times(R_{st}-1)$ matrix, is obtained from $\mathbf{Z}_{st}$ by deleting the last row and column, $s = 1,\ldots,S$, $t = 1,\ldots,T_s$; $\bar{\mathbf{Z}}_s = \left(\bar{\mathbf{Z}}_{st}, t = 1,\ldots,T_s\right)'$, $s = 1,\ldots,S$; $\bar{\boldsymbol{\theta}}_{st} = \bar{\mathbf{Z}}'_{st}\bar{\boldsymbol{\theta}}_s = \left(\theta_{c(st)}, c = 1,\ldots,R_{st}-1\right)'$ encloses the

parameters $\{\theta_{r(s)}\}$ associated to the first $R_{st}-1$ classes of the t-th missingness pattern of the s-th subpopulation, where $\theta_{c(st)} = \sum_{r\in\mathcal{C}_{stc}} \theta_{r(s)}$, $s = 1,\ldots,S$, $t = 1,\ldots,T_s$; $\mathbf{p}_{st} = \mathbf{N}_{st}/n_{st+} = \left(p_{c(st)}, c = 1,\ldots,R_{st}\right)'$ are the observed proportions in the t-th missingness pattern of the s-th subpopulation, $s = 1,\ldots,S$, $t = 1,\ldots,T_s$; $\mathbf{p}_s = (\mathbf{p}'_{st}, t = 1,\ldots,T_s)'$, $s = 1,\ldots,S$; $\bar{\mathbf{N}}_{st} = [\mathbf{I}_{R_{st}-1}, \mathbf{0}_{R_{st}-1}]\,\mathbf{N}_{st} = (n_{stc}, c = 1,\ldots,R_{st}-1)'$, $s = 1,\ldots,S$, $t = 1,\ldots,T_s$; $\bar{\mathbf{p}}_{st} = \bar{\mathbf{N}}_{st}/n_{st+}$, $s = 1,\ldots,S$, $t = 1,\ldots,T_s$; $\bar{\mathbf{p}}_s = (\bar{\mathbf{p}}'_{st}, t = 1,\ldots,T_s)'$, $s = 1,\ldots,S$. Whenever that it will be necessary, we obtain $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}$ from $\bar{\boldsymbol{\theta}}_s$ and $\bar{\boldsymbol{\theta}}$ by the relations

$$\boldsymbol{\theta}_s = \begin{pmatrix} \mathbf{0}_{R-1} \\ 1 \end{pmatrix} + \begin{pmatrix} \mathbf{I}_{R-1} \\ -\mathbf{1}'_{R-1} \end{pmatrix} \bar{\boldsymbol{\theta}}_s \equiv \mathbf{b}_s + \mathbf{B}_s\bar{\boldsymbol{\theta}}_s, \tag{7}$$

$$\boldsymbol{\theta} = \mathbf{1}_S \otimes \begin{pmatrix} \mathbf{0}_{R-1} \\ 1 \end{pmatrix} + \left[\mathbf{I}_S \otimes \begin{pmatrix} \mathbf{I}_{R-1} \\ -\mathbf{1}'_{R-1} \end{pmatrix}\right] \bar{\boldsymbol{\theta}} \equiv \mathbf{b} + \mathbf{B}\bar{\boldsymbol{\theta}}, \tag{8}$$

where $\mathbf{b}_s = \left(\mathbf{0}'_{R-1}, 1\right)'$, $\mathbf{B}_s = (\mathbf{I}_{R-1}, -\mathbf{1}_{R-1})'$, $\mathbf{b} = \mathbf{1}_S \otimes \left(\mathbf{0}'_{R-1}, 1\right)'$, and $\mathbf{B} = \mathbf{I}_S \otimes (\mathbf{I}_{R-1}, -\mathbf{1}_{R-1})'$.

4.1 ML inferences under MAR and MCAR assumptions

The ML estimation of $\boldsymbol{\theta}$ can be based only on the factor $L_1(\boldsymbol{\theta}\,|\mathbf{N})$ in expression (4). With the exception of the monotone missingness pattern (Rubin, 1974), $\partial \ln L_1(\boldsymbol{\theta}\,|\mathbf{N})/\partial\boldsymbol{\theta} = \mathbf{0}$ does not have, in general, an explicit solution, implying that ML estimators must be obtained through iterative methods. The EM algorithm is expressed by

$$\hat{\boldsymbol{\theta}}_s^{(i+1)} = \frac{1}{n_{s++}} \left(\mathbf{N}_{s1} + \sum_{t=2}^{T_s} \mathbf{D}_{\hat{\boldsymbol{\theta}}_s^{(i)}} \mathbf{Z}_{st} \mathbf{D}^{-1}_{\mathbf{Z}'_{st}\hat{\boldsymbol{\theta}}_s^{(i)}} \mathbf{N}_{st} \right), \quad s = 1,\ldots,S,\ i = 1,\ldots, \tag{9}$$

where $\mathbf{D}_{\hat{\boldsymbol{\theta}}_s^{(i)}}$ denotes a diagonal matrix with the elements of $\hat{\boldsymbol{\theta}}_s^{(i)}$ in the main diagonal, and $\hat{\boldsymbol{\theta}}_s^{(i)}$ is the local maximum point estimate achieved in the i-th iterate. We may initiate the iterative process with, for example, the observed proportions of sampling units completely categorized, *i.e.*, $\hat{\boldsymbol{\theta}}_s^{(0)} = \mathbf{p}_{s1} = \mathbf{N}_{s1}/n_{s1+}$. In this case, it is important to replace any frequency possibly null in the absence of missingness pattern by a small value, *e.g.*, $(R\, n_{s1+})^{-1}$ or 10^{-6}, since a null value of $\hat{\theta}_{r(s)}^{(0)}$ does not allow that information from other missingness patterns be incorporated.

The slow rate of convergence of the EM algorithm may be by-passed with Newton-Raphson or Fisher scoring algorithms, for which we need to derive the gradient vector and the hessian matrix of the log-likelihood, or the Fisher information matrix. The $S(R-1)\times 1$ score vector of $\ln L_1(\boldsymbol{\theta}\,|\mathbf{N}_s)$ is expressed by $\mathbf{S}_1\left(\bar{\boldsymbol{\theta}}\right) = (\mathbf{S}'_{1s}, s = 1,\ldots,S)'$, where

$$\mathbf{S}_{1s} = \sum_{t=1}^{T_s} \bar{\mathbf{Z}}_{st} \left[\boldsymbol{\Sigma}\left(\bar{\boldsymbol{\theta}}_{st}\right)\right]^{-1} \left(\bar{\mathbf{p}}_{st} - \bar{\boldsymbol{\theta}}_{st}\right), \quad s = 1,\ldots,S \tag{10}$$

and $\boldsymbol{\Sigma}\left(\bar{\boldsymbol{\theta}}_{st}\right) = \dfrac{1}{n_{st+}}\left(\mathbf{D}_{\bar{\boldsymbol{\theta}}_{st}} - \bar{\boldsymbol{\theta}}_{st}\bar{\boldsymbol{\theta}}'_{st}\right)$. The $S(R-1)\times S(R-1)$ hessian matrix of $\ln L_1\left(\boldsymbol{\theta}\,|\mathbf{N}_s\right)$, $\mathbf{H}_1\left(\bar{\boldsymbol{\theta}}\right)$, is a block diagonal matrix with blocks

$$\mathbf{H}_{1s} = -\sum_{t=1}^{T_s}\bar{\mathbf{Z}}_{st}\left[\mathbf{D}_{\bar{\mathbf{N}}_{st}}\mathbf{D}^{-2}_{\bar{\boldsymbol{\theta}}_{st}} + \frac{n_{stR_{st}}}{\left(1-\mathbf{1}'_{R_{st}-1}\bar{\boldsymbol{\theta}}_{st}\right)^2}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right]\bar{\mathbf{Z}}'_{st}, \quad s=1,\ldots,S, \tag{11}$$

where $\mathbf{D}^{-2}_{\bar{\boldsymbol{\theta}}_{st}} = \mathbf{D}^{-1}_{\bar{\boldsymbol{\theta}}_{st}}\mathbf{D}^{-1}_{\bar{\boldsymbol{\theta}}_{st}}$.

The Fisher scoring algorithm demand additional estimation of the conditional probabilities of missingness, since

$$E\left(n_{stc}\left|\mathbf{N}_{++},\boldsymbol{\theta},\{\boldsymbol{\alpha}^{\mathrm{MAR}}_{st}\}\right.\right) = n_{s++}\mathbf{z}'_{stc}\boldsymbol{\theta}_s\alpha_{t(cs)}, \tag{12}$$

$$E\left(n_{stc}\left|\mathbf{N}_{++},\boldsymbol{\theta},\{\boldsymbol{\alpha}^{\mathrm{MCAR}}_{st}\}\right.\right) = n_{s++}\mathbf{z}'_{stc}\boldsymbol{\theta}_s\alpha_{t(s)}, \tag{13}$$

$s=1,\ldots,S$, $t=1,\ldots,T_s$, $c=1,\ldots,R_{ts}$, where $\boldsymbol{\alpha}^{\mathrm{MAR}}_{st} = (\alpha_{t(cs)}, c=1,\ldots,R_{st})'$ and $\boldsymbol{\alpha}^{\mathrm{MCAR}}_{st} = \alpha_{t(s)}$. As the statistical model under the MAR mechanism is saturated, once the ML estimate $\{\hat{\boldsymbol{\theta}}_s\}$ of $\{\boldsymbol{\theta}_s\}$ has been obtained, by the invariance property

$$\hat{\boldsymbol{\alpha}}^{\mathrm{MAR}}_{st} = \frac{1}{n_{s++}}\mathbf{D}^{-1}_{\mathbf{Z}'_{st}\hat{\boldsymbol{\theta}}_s}\mathbf{N}_{st}, \quad s=1,\ldots,S,\ t=1,\ldots,T_s. \tag{14}$$

The factor $L_2\left(\{\alpha_{t(s)}\}\,|\{n_{st+}\};\mathrm{MCAR}\right)$ directly conduces to the ML estimators of the conditional probabilities of missingness under the MCAR mechanism

$$\hat{\boldsymbol{\alpha}}^{\mathrm{MCAR}}_{st} = \hat{\alpha}_{t(s)} = \frac{n_{st+}}{n_{s++}}, \quad s=1,\ldots,S,\ t=1,\ldots,T_s. \tag{15}$$

The Fisher information matrix referring to $\bar{\boldsymbol{\theta}}$ under the MAR mechanism, $\boldsymbol{\mathcal{I}}_1\left(\bar{\boldsymbol{\theta}},\left\{\boldsymbol{\alpha}^{\mathrm{MAR}}_{st}\right\}\right)$, is a block diagonal matrix with blocks

$$\boldsymbol{\mathcal{I}}^{\mathrm{MAR}}_{1s} = n_{s++}\sum_{t=1}^{T_s}\bar{\mathbf{Z}}_{st}\left[\mathbf{D}_{\bar{\boldsymbol{\alpha}}^{\mathrm{MAR}}_{st}}\mathbf{D}^{-1}_{\bar{\boldsymbol{\theta}}_{st}} + \frac{\alpha_{t(sR_{ts})}}{1-\mathbf{1}'_{R_{st}-1}\bar{\boldsymbol{\theta}}_{st}}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right]\bar{\mathbf{Z}}'_{st}, \quad s=1,\ldots,S, \tag{16}$$

where $\bar{\boldsymbol{\alpha}}^{\mathrm{MAR}}_{st} = [\mathbf{I}_{R_{st}-1},\mathbf{0}_{R_{st}-1}]\,\boldsymbol{\alpha}^{\mathrm{MAR}}_{st} = (\alpha_{t(cs)}, c=1,\ldots,R_{st}-1)'$, $s=1,\ldots,S$, $t=1,\ldots,T_s$. The Fisher information matrix referring to $\bar{\boldsymbol{\theta}}$ under the MCAR mechanism, $\boldsymbol{\mathcal{I}}_1\left(\bar{\boldsymbol{\theta}},\left\{\boldsymbol{\alpha}^{\mathrm{MCAR}}_{st}\right\}\right)$, is a block diagonal matrix with blocks

$$\boldsymbol{\mathcal{I}}^{\mathrm{MCAR}}_{1s} = n_{s++}\sum_{t=1}^{T_s}\alpha_{t(s)}\bar{\mathbf{Z}}_{st}\left[\mathbf{D}^{-1}_{\bar{\boldsymbol{\theta}}_{st}} + \frac{1}{1-\mathbf{1}'_{R_{st}-1}\bar{\boldsymbol{\theta}}_{st}}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right]\bar{\mathbf{Z}}'_{st}, \quad s=1,\ldots,S. \tag{17}$$

Using the Fisher information matrix as a precision measure, $\hat{\mathbf{V}}^{\mathrm{MAR}}_{\hat{\bar{\boldsymbol{\theta}}}} = \left[\boldsymbol{\mathcal{I}}_1\left(\hat{\bar{\boldsymbol{\theta}}},\left\{\hat{\boldsymbol{\alpha}}^{\mathrm{MAR}}_{st}\right\}\right)\right]^{-1}$ and $\hat{\mathbf{V}}^{\mathrm{MCAR}}_{\hat{\bar{\boldsymbol{\theta}}}} = \left[\boldsymbol{\mathcal{I}}_1\left(\hat{\bar{\boldsymbol{\theta}}},\left\{\hat{\boldsymbol{\alpha}}^{\mathrm{MCAR}}_{st}\right\}\right)\right]^{-1}$ are estimates of the approximate covariance matrices of $\hat{\bar{\boldsymbol{\theta}}}$

under the MAR and the MCAR mechanisms, respectively. Employing the observed information matrix $-\mathbf{H}_1\left(\bar{\boldsymbol{\theta}}\right)$ as a precision measure, $\left[-\mathbf{H}_1\left(\hat{\bar{\boldsymbol{\theta}}}\right)\right]^{-1}$ is an estimate of the approximate covariance matrix of $\hat{\bar{\boldsymbol{\theta}}}$ under both the MAR and the MCAR mechanisms.

Substituting $\{\hat{\alpha}_{t(cs)} = n_{stc}/(n_{s++}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s)\}$ from (14) in $\boldsymbol{\mathcal{I}}_1\left(\hat{\bar{\boldsymbol{\theta}}}, \{\hat{\boldsymbol{\alpha}}_{st}^{\text{MAR}}\}\right)$, we notice that $\boldsymbol{\mathcal{I}}_1\left(\hat{\bar{\boldsymbol{\theta}}}, \{\hat{\boldsymbol{\alpha}}_{st}^{\text{MAR}}\}\right) = -\mathbf{H}_1(\hat{\bar{\boldsymbol{\theta}}})$. Thus, effectively there are three different iterative process for the obtainment of the ML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$: (i) EM, (ii) Fisher scoring under the MCAR mechanism, and (iii) Fisher scoring under the MAR mechanism or Newton-Raphson under the MAR or the MCAR mechanisms. As the ML estimator of $\boldsymbol{\theta}$ is the same under the MAR and the MCAR mechanisms, we may use the iterative process (ii) even when assuming the MAR mechanism, since after getting $\hat{\boldsymbol{\theta}}$ we use an estimate of the approximate covariance matrix under the MAR mechanism. This may be the best choice to avoid the low speed of the EM algorithm and, at the same time, possible instability of the iterative process (iii), because as the MAR mechanism involves a saturated model, zero counts may easily generate unstable covariance matrices.

Using (8) and the *delta* method, we obtain the estimates of the approximate covariance matrices of $\hat{\boldsymbol{\theta}}$ under the MAR and the MCAR mechanisms by $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{MAR}} = \mathbf{B}\hat{\mathbf{V}}_{\hat{\bar{\boldsymbol{\theta}}}}^{\text{MAR}}\mathbf{B}'$ and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{MCAR}} = \mathbf{B}\hat{\mathbf{V}}_{\hat{\bar{\boldsymbol{\theta}}}}^{\text{MCAR}}\mathbf{B}'$.

The goodness-of-fit test of the MCAR mechanism, conditionally on the MAR assumption, can be accomplished with either the Wilks likelihood ratio statistic

$$\begin{aligned} Q_L(\text{MCAR}|\text{MAR}) &= -2\ \ln\frac{L_2\left(\{\hat{\alpha}_{t(s)}\}\,|\{n_{st+}\};\text{MCAR}\right)}{L_2\left(\{\hat{\alpha}_{t(cs)}\}\,|\mathbf{N};\text{MAR}\right)} \\ &= -2\sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} n_{stc}\left[\ln\left(\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s\right) - \ln\left(\frac{n_{stc}}{n_{st+}}\right)\right] \\ &= -2\sum_{s=1}^{S}\mathbf{N}'_s\left[\mathbf{ln}\left(\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s\right) - \mathbf{ln}\left(\mathbf{p}_s\right)\right], \end{aligned} \tag{18}$$

or the Pearson statistic, that is a score-type statistic,

$$\begin{aligned} Q_P(\text{MCAR}|\text{MAR}) &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}}\frac{\left(n_{stc} - n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s\right)^2}{n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s} \\ &= \sum_{s=1}^{S}\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s\right)'\left[\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s}\right]\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s\right) \end{aligned} \tag{19}$$

or the Neyman statistic

$$Q_N(\text{MCAR}|\text{MAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{stc} - n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s\right)^2}{n_{stc}}$$
$$= \sum_{s=1}^{S}\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s\right)'\left[\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}_{\mathbf{p}_s}^{-1}\right]\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s\right), \quad (20)$$

where $\mathbf{ln}(\mathbf{p}_s)$ denotes the vector (natural) logarithmic operator which consists of taking the natural logarithmic on each element of $\mathbf{p}_s$, and $\mathbf{N}_{s+} = \left(n_{st+} \otimes \mathbf{1}'_{R_{st}}, t = 1, \ldots, T_s\right)'$ is the vector with the same dimension as $\mathbf{N}_s$ that contains the total of the observed frequencies in each missingness pattern of the s-th subpopulation sequentially repeated according to the number of classes in each pattern (note that $\mathbf{p}_s = \mathbf{D}_{\mathbf{N}_{s+}}^{-1}\mathbf{N}_s$). All three statistics are asymptotically equivalent with null distribution $\chi^2_{(g)}$, where $g = S + \sum_{s=1}^{S}(l_s - T_s)$, and $\chi^2_{(g)}$ represents the chi-square distribution with g degrees of freedom. In spite of the general form of the expression (18), as a null observed frequency n_{stc} does not contribute to the probability mass function (1), in these cases we should use the definition $0 \times \left[\ln\left(\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s\right) - \ln\left(0/n_{st+}\right)\right] \equiv 0$ and avoid the calculation of ln(0). The Neyman statistic (20) presuppose $\{n_{stc} > 0\}$ or, equivalently, $\{p_{c(st)} > 0\}$, which does not always happen in practice. Therefore, we suggest to replace possibly null frequencies by some small value before obtaining $\mathbf{p}_s$ and calculating the inverse of $\mathbf{D}_{\mathbf{p}_s}$. In the WLS context, Koch, Imrey and Reinfurt (1972) suggest that $n_{stc} = 0$ be replaced by $(R_{st}n_{st+})^{-1}$.

The augmented expected frequencies can be estimated by

$$\hat{y}_{str}^{\text{MAR}} = \widehat{E}\left(y_{str}\middle|n_{s++}, \hat{\theta}_{r(s)}, \hat{\alpha}_{t(cs)}\right) = n_{s++}\hat{\theta}_{r(s)}\hat{\alpha}_{t(cs)}, \quad (21)$$

$s = 1, \ldots, S$, $t = 1, \ldots, T_s$, $r = 1, \ldots, R$, $\{c : r \in \mathcal{C}_{stc}\}$, under the MAR mechanism, and by

$$\hat{y}_{str}^{\text{MCAR}} = \widehat{E}\left(y_{str}\middle|n_{s++}, \hat{\theta}_{r(s)}, \hat{\alpha}_{t(s)}\right) = n_{s++}\hat{\theta}_{r(s)}\hat{\alpha}_{t(s)}, \quad (22)$$

$s = 1, \ldots, S$, $t = 1, \ldots, T_s$, $r = 1, \ldots, R$, under the MCAR mechanism.

4.2 WLS inferences under MCAR assumption

Ignorability of the missingness process under the MCAR mechanism for frequentist inferences on $\boldsymbol{\theta}$ allows us to focus on the distribution of $\mathbf{N}_s$ conditionally on $\{n_{st+}\}$, which is a product of T_s multinomial distributions for each one of the $s = 1, \ldots, S$ subpopulations

$$\mathbf{N}_{st}|n_{st+}, \bar{\boldsymbol{\theta}}_{st} \overset{\text{indep.}}{\sim} M_{R_{st}}(n_{st+}, \bar{\boldsymbol{\theta}}^0_{st}), \quad t = 1, \ldots, T_s. \quad (23)$$

The MCAR assumption implies adoption of a linear structure for the parametric vector $\bar{\boldsymbol{\theta}}^0_{s*} = \left(\bar{\boldsymbol{\theta}}^{0'}_{st}, t = 1, \ldots, T_s\right)'$, *i.e.*,

$$\text{MCAR} : \bar{\boldsymbol{\theta}}^0_{s*} = \bar{\mathbf{Z}}'_s\bar{\boldsymbol{\theta}}_s, \quad s = 1, \ldots, S, \quad (24)$$

enabling the application of the WLS methodology, also known as GSK due to Grizzle, Starmer and Koch (1969), with the difference lying in the fact that the response categories vary from one multinomial to another, as it was pointed out by Koch *et al.* (1972).

The WLS approach proceeds by minimization of the quadratic form

$$\begin{aligned} Q_N\left(\bar{\boldsymbol{\theta}}\right) &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(\bar{\mathbf{p}}_{st}-\bar{\mathbf{Z}}'_{st}\bar{\boldsymbol{\theta}}_s\right)'\left[\boldsymbol{\Sigma}\left(\bar{\mathbf{p}}_{st}\right)\right]^{-1}\left(\bar{\mathbf{p}}_{st}-\bar{\mathbf{Z}}'_{st}\bar{\boldsymbol{\theta}}_s\right) \\ &= \sum_{s=1}^{S}\left(\bar{\mathbf{p}}_{s}-\bar{\mathbf{Z}}'_{s}\bar{\boldsymbol{\theta}}_s\right)'\left[\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_{s}\right)\right]^{-1}\left(\bar{\mathbf{p}}_{s}-\bar{\mathbf{Z}}'_{s}\bar{\boldsymbol{\theta}}_s\right), \end{aligned} \tag{25}$$

where $\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)$ is a block diagonal matrix with blocks $\boldsymbol{\Sigma}\left(\bar{\mathbf{p}}_{st}\right)$, $t=1,\ldots,T_s$, that result from $\boldsymbol{\Sigma}\left(\bar{\boldsymbol{\theta}}_{st}\right)$ replacing $\bar{\boldsymbol{\theta}}_{st}$ by $\bar{\mathbf{p}}_{st}$. Under the MCAR mechanism, the WLS estimator of $\bar{\boldsymbol{\theta}}_s$ is

$$\tilde{\bar{\boldsymbol{\theta}}}_s = \left(\bar{\mathbf{Z}}_s\left[\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)\right]^{-1}\bar{\mathbf{Z}}'_s\right)^{-1}\bar{\mathbf{Z}}_s\left[\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)\right]^{-1}\bar{\mathbf{p}}_s, \tag{26}$$

and an estimate of its approximate covariance matrix is

$$\tilde{\mathbf{V}}_{\tilde{\bar{\boldsymbol{\theta}}}_s} = \left(\bar{\mathbf{Z}}_s\left[\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)\right]^{-1}\bar{\mathbf{Z}}'_s\right)^{-1}. \tag{27}$$

From (7), we obtain the WLS estimator of $\boldsymbol{\theta}_s$ by $\tilde{\boldsymbol{\theta}}_s = \mathbf{b}_s + \mathbf{B}_s\tilde{\bar{\boldsymbol{\theta}}}_s$; analogously, an estimate of the approximate covariance matrix is $\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}_s} = \mathbf{B}_s\tilde{\mathbf{V}}_{\tilde{\bar{\boldsymbol{\theta}}}_s}\mathbf{B}'_s$. An estimate of the approximate covariance matrix of $\tilde{\bar{\boldsymbol{\theta}}}$, $\tilde{\mathbf{V}}_{\tilde{\bar{\boldsymbol{\theta}}}}$, is a block diagonal matrix with blocks $\tilde{\mathbf{V}}_{\tilde{\bar{\boldsymbol{\theta}}}_s}$, $s=1,\ldots,S$, and the one from $\tilde{\boldsymbol{\theta}}$ is $\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}} = \mathbf{B}\tilde{\mathbf{V}}_{\tilde{\bar{\boldsymbol{\theta}}}}\mathbf{B}'$.

We may accomplish a goodness-of-fit test of the MCAR mechanism with the Neyman statistic

$$Q_N(\text{MCAR}) = \sum_{s=1}^{S}\left(\bar{\mathbf{p}}_s-\bar{\mathbf{Z}}'_s\tilde{\bar{\boldsymbol{\theta}}}_s\right)'\left[\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)\right]^{-1}\left(\bar{\mathbf{p}}_s-\bar{\mathbf{Z}}'_s\tilde{\bar{\boldsymbol{\theta}}}_s\right), \tag{28}$$

which has an approximate $\chi^2_{(g)}$ null distribution for large values of $\{n_{stc}\}$, where $g = S+\sum_{s=1}^{S}(l_s - T_s)$.

In the expressions (25), (26), (27) and (28), we assume that $\boldsymbol{\Sigma}_*\left(\bar{\mathbf{p}}_s\right)$ is not singular so that we can uniquely calculate its inverse. As we do not always have $\{p_{c(st)} > 0\}$ or, equivalently, $\{n_{stc} > 0\}$ in practice, we can follow the suggestion by Koch *et al.* (1972) referred to earlier.

An estimate of the augmented expected frequencies is given by

$$\tilde{y}_{str} = \widetilde{E}\left(y_{str}\middle|n_{st+},\tilde{\theta}_{r(s)}\right) = n_{st+}\tilde{\theta}_{r(s)}, \qquad s=1,\ldots,S,\ t=1,\ldots,T_s,\ r=1,\ldots,R. \tag{29}$$

5 Nonsaturated structural models for the marginal probabilities of categorization

As in the complete-data situation, in general, we want to model the probabilities of categorization by means of nonsaturated structural models with the purpose of elucidating questions of interest. We illustrate a few structures for the data sets presented. For a throughout treatment of the complete-data case, see Forthofer and Lehnen (1981) and Koch, Imrey, Singer, Atkinson and Stokes (1985).

Example 1 (cont.) We assess the association between maternal smoking and child's wheezing status given home city through log local odds ratios

$$\omega_{ij(s)} = \ln\left(\frac{\pi_{ij(s)}\pi_{i+1,j+1(s)}}{\pi_{i,j+1(s)}\pi_{i+1,j(s)}}\right), \qquad i,j,s = 1,2,$$

where $(\pi_{11(s)}, \pi_{12(s)}, \ldots, \pi_{33(s)})' = (\theta_{r(s)}, r = 1, \ldots, 9)' = \boldsymbol{\theta}_s$, $s = 1, 2$. We can perform the same type of conditional independence test considered by Lipsitz and Fitzmaurice (1996), that takes the ordering of the categories into account, by first assuming a homogeneous linear-by-linear association model with unit-spaced response scores (Agresti, 2002), namely, $\omega_{ij(s)} = \beta$, and then by testing $H : \beta = 0$. This can be formulated as a log-linear model $\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}$ with

$$\mathbf{A} = \left(\mathbf{I}_2 \otimes \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix}\right) = \mathbf{I}_2 \otimes \mathbf{E} \otimes \mathbf{E},$$

$$\mathbf{E} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{X}_L = \mathbf{1}_8, \qquad \boldsymbol{\beta} = \beta.$$

For the unordered case, also considered by the authors, the log-linear model of homogeneous association or no three-factor interaction is obtained by taking $\mathbf{X}_L = \mathbf{1}_2 \otimes \mathbf{I}_4$ and $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})'$. The independence test conditionally on this assumed model may be assessed by the hypothesis $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}_4$, where $\mathbf{C} = \mathbf{I}_4$. For instance, if the interest is not in estimating $\boldsymbol{\beta}$, but just checking the fit of the homogeneous association model, we may use the equivalent constraint formulation $\mathbf{U}_L\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_4$, where $\mathbf{U}_L = ([1, -1] \otimes \mathbf{I}_4)$. Note that the rows of the matrix $\mathbf{U}_L$ are orthogonal to the columns of the matrix $\mathbf{X}_L$, *i.e.*, $\mathbf{U}_L\mathbf{X}_L = \mathbf{0}_{4,4}$.

Example 2 (cont.) We can assess the homogeneity of marginal distributions of the risk degree to dental caries obtained under both methods through the fit of the (strictly) linear model

$\mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} [\mathbf{I}_2, \mathbf{0}_2] \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes [\mathbf{I}_2, \mathbf{0}_2] \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{1}_2 \otimes \mathbf{I}_2,$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. If there is no interest in $\boldsymbol{\beta}$, we may use the equivalent constraint formulation $\mathbf{UA}\boldsymbol{\theta} = \mathbf{0}_2$, with $\mathbf{U} = ([1, -1] \otimes \mathbf{I}_2)$. Note that the rows of the matrix $\mathbf{U}$ are orthogonal to the columns of the matrix $\mathbf{X}$, namely, $\mathbf{UX} = \mathbf{0}_{2,2}$. The degree of agreement between the methods may be measured by the weighted *kappa* index (Fleiss, Levin and Paik, 2003)

$$\kappa_w = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}\pi_{ij} - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}\pi_{i+}\pi_{+j}}, \tag{30}$$

where $(\pi_{11}, \pi_{12}, \ldots, \pi_{33})' = (\theta_r, r = 1, \ldots, 9)' = \boldsymbol{\theta}$. This measure may be written as a functional linear model $\mathbf{F} = \boldsymbol{\pi}_1 + \mathbf{exp}(\mathbf{A}_4\, \mathbf{ln}\{\mathbf{A}_3\, \mathbf{exp}[\mathbf{A}_2\, \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})]\})$ with

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{W}' \\ \mathbf{1}_9' \\ \mathbf{I}_3 \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes \mathbf{I}_3 \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2,6} \\ \mathbf{0}_{9,2} & (\mathbf{I}_3 \otimes \mathbf{1}_3, \mathbf{1}_3 \otimes \mathbf{I}_3) \end{bmatrix},$$

$$\mathbf{A}_3 = [(1, 0)', \mathbf{1}_2, -(2, 1)'\mathbf{W}'], \qquad \mathbf{A}_4 = [1, -1] \qquad \text{and} \qquad \boldsymbol{\pi}_1 = -1,$$

where $\mathbf{W} = (w_{11}, w_{12}, \ldots, w_{33})'$ is a 9×1 vector embodying the weights.

5.1 ML inferences on linear and log-linear models under MAR and MCAR assumptions

We consider (strictly) linear models written in the form of freedom equations

$$M_L : \mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}, \tag{31}$$

where $\mathbf{A}$ is an $u \times SR$ matrix defining the u linear functions of interest with rank $r(\mathbf{A}) = u \leq S(R-1)$; $\mathbf{X}$ is an $u \times p$ matrix specifying the model with rank $r(\mathbf{X}) = p \leq u$; and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector that contains the unknown parameters. This model can alternatively be fitted by using an equivalent constraint formulation

$$M_L : \mathbf{UA}\boldsymbol{\theta} = \mathbf{0}_{u-p}, \tag{32}$$

where $\mathbf{U}$ is an $(u - p) \times u$ matrix containing the $u - p$ constraints with full rank whose rows are orthogonal to the columns of $\mathbf{X}$, *i.e.*, $\mathbf{UX} = \mathbf{0}_{(u-p),p}$. We also have to include in the model specification the S natural constraints $\sum_{r=1}^{R} \theta_{r(s)} = 1$, $s = 1, \ldots, S$, matrically represented by

$$[\mathbf{I}_S \otimes \mathbf{1}_R']\, \boldsymbol{\theta} = \mathbf{1}_S. \tag{33}$$

In this regard, we assume that the rows of $\mathbf{A}$ are linearly independent from the columns of the matrix $\mathbf{I}_S \otimes \mathbf{1}_R$, *i.e.*, $r(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R) = u + S$.

To take advantage of the development in Section 4.1, in function of $\bar{\boldsymbol{\theta}}$, it is convenient to rewrite (31). The junction of (31) and (33) conducts to

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \end{pmatrix}. \tag{34}$$

Therefore, if $r(\mathbf{A}) = u = S(R-1)$, then we can obtain $\bar{\boldsymbol{\theta}}$ solely from ($\mathbf{A}$, $\mathbf{X}$ and) $\boldsymbol{\beta}$ by means of

$$\bar{\boldsymbol{\theta}}(\boldsymbol{\beta}) = (\mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]) \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \end{pmatrix}. \tag{35}$$

When $r(\mathbf{A}) = u < S(R-1)$, we need a $(S[R-1]-u) \times SR$ matrix $\mathbf{A}_0$, base of the orthocomplement matrix $(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R)'$, in order to augment the model (34) to

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \\ \mathbf{A}_0 \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \\ \boldsymbol{\beta}_0 \end{pmatrix}, \tag{36}$$

which encompasses the former, but has also $S[R-1]-u$ additional nuisance parameters included in $\boldsymbol{\beta}_0$. In this case, we obtain $\bar{\boldsymbol{\theta}}$ from $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ in agreement with

$$\bar{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = (\mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]) \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \\ \mathbf{A}_0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \\ \boldsymbol{\beta}_0 \end{pmatrix}. \tag{37}$$

For simplicity, in the following development, we suppress the $\boldsymbol{\beta}_0$ parameter, mentioning only $\boldsymbol{\beta}$. Nevertheless, always that $u < S(R-1)$, $\boldsymbol{\beta}_0$ is also used. See Koch *et al.* (1985) for more details concerning the amplification of linear models.

Incorporation of this linear structure in $\ln L_1\left(\boldsymbol{\theta}(\boldsymbol{\beta})\,|\mathbf{N}\right)$ and its differentiation with respect to $\boldsymbol{\beta}$, by means of the matrix chain rule, give rise to the score vector $\mathbf{S}_{1L}\left(\boldsymbol{\beta}\right) = \mathbf{W}'\mathbf{S}_1\left(\bar{\boldsymbol{\theta}}(\boldsymbol{\beta})\right)$, and to the Fisher information matrix under the $\mathcal{M}$ (MAR or MCAR) mechanism

$$\boldsymbol{\mathcal{I}}_{1L}\left(\boldsymbol{\beta}, \left\{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\right\}\right) = \mathbf{W}'\boldsymbol{\mathcal{I}}_1\left(\bar{\boldsymbol{\theta}}(\boldsymbol{\beta}), \left\{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\right\}\right)\mathbf{W}, \tag{38}$$

where

$$\mathbf{W} = (\mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]) \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X} \\ \mathbf{0}_{S,p} \end{pmatrix}, \tag{39}$$

if $u = S(R-1)$, or

$$\mathbf{W} = (\mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]) \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \\ \mathbf{A}_0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X} & \mathbf{0}_{u,S(R-1)-u} \\ \mathbf{0}_{S,p} & \mathbf{0}_{S,S(R-1)-u} \\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u} \end{pmatrix}, \tag{40}$$

if $u < S(R-1)$; $\mathbf{S}_1\left(\bar{\boldsymbol{\theta}}(\boldsymbol{\beta})\right)$ is presented near (10); $\boldsymbol{\mathcal{I}}_1\left(\bar{\boldsymbol{\theta}}(\boldsymbol{\beta}), \left\{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\right\}\right)$ is described near (16) for $\mathcal{M} = \text{MAR}$ and, near (17) for $\mathcal{M} = \text{MCAR}$; $\bar{\boldsymbol{\theta}}(\boldsymbol{\beta})$ is defined in (35), if $u = S(R-1)$, or in (37), if $u < S(R-1)$.

The score vector and the Fisher information matrix allows us to obtain ML estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by using the Fisher scoring algorithm. The iterative process may be initialized with the WLS estimate (special case of what we will present in Section 5.2)

$$\hat{\boldsymbol{\beta}}^{(0)} = \left\{\mathbf{X}'\left[\mathbf{A}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\mathcal{M}}\mathbf{A}'\right]^{-1}\mathbf{X}\right\}^{-1}\mathbf{X}'\left[\mathbf{A}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\mathcal{M}}\mathbf{A}'\right]^{-1}\mathbf{A}\hat{\boldsymbol{\theta}}, \tag{41}$$

if $u = S(R-1)$, or by means of an analogous expression derived from replacing $\mathbf{A}$ by $(\mathbf{A}', \mathbf{A}_0')'$ and $\mathbf{X}$ by

$$\begin{pmatrix} \mathbf{X} & \mathbf{0}_{u,S(R-1)-u} \\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u} \end{pmatrix}, \tag{42}$$

if $u < S(R-1)$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model, and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\mathcal{M}}$, an estimate of its corresponding approximate covariance matrix under the mechanism $\mathcal{M}$, obtained according to the suggestion presented in Section 4.1. An alternative iterative scheme would consist in applying the procedure by Paulino and Silva (1999), once adapted to incomplete data, by using the formulation (32) provided with (8) — see also Paulino and Singer (2006, Sect. 8.3).

An estimate of the approximate covariance matrix of $\hat{\boldsymbol{\beta}}$ under $\mathcal{M}$ is $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}L}^{\mathcal{M}} = \left[\boldsymbol{\mathcal{I}}_{1L}\left(\hat{\boldsymbol{\beta}}, \left\{\hat{\boldsymbol{\alpha}}_{st}^{\mathcal{M}}\right\}\right)\right]^{-1}$. Using (35), if $u = S(R-1)$, or (37), if $u < S(R-1)$, we obtain the ML estimate $\hat{\bar{\boldsymbol{\theta}}}(M_L)$ of $\bar{\boldsymbol{\theta}}$ under the linear model M_L and, with the *delta* method, an estimate of its corresponding approximate covariance matrix $\hat{\mathbf{V}}_{\hat{\bar{\boldsymbol{\theta}}}(M_L)}^{\mathcal{M}} = \mathbf{W}\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}L}^{\mathcal{M}}\mathbf{W}'$. By means of (8) and the *delta* method, an estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}(M_L)$ is given by $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}(M_L)}^{\mathcal{M}} = \mathbf{B}\hat{\mathbf{V}}_{\hat{\bar{\boldsymbol{\theta}}}(M_L)}^{\mathcal{M}}\mathbf{B}'$. The ML estimates of the linear functions $\mathbf{A}\boldsymbol{\theta}$ under M_L are $\mathbf{X}\hat{\boldsymbol{\beta}}$ and, using the *delta* method, an estimate of its approximate covariance matrix is obtained by $\hat{\mathbf{V}}_{\mathbf{A}\hat{\boldsymbol{\theta}}(M_L)}^{\mathcal{M}} = \mathbf{X}\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}L}^{\mathcal{M}}\mathbf{X}'$.

Another very important structure is the log-linear one, which we may express in the form

$$M_{LL} : \mathbf{ln}(\boldsymbol{\theta}_s) = \mathbf{1}_R\nu_s + \mathbf{X}_s\boldsymbol{\beta}, \quad s = 1, \ldots, S$$

or, in a condensed way, by

$$M_{LL} : \mathbf{ln}(\boldsymbol{\theta}) = [\mathbf{I}_S \otimes \mathbf{1}_R]\,\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\beta}, \tag{43}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_S)'$ is a vector with S components associated to the natural constraints, such that $\boldsymbol{\nu} = -\mathbf{ln}\left[(\mathbf{I}_S \otimes \mathbf{1}_R')\,\mathbf{exp}\,(\mathbf{X}\boldsymbol{\beta})\right]$, $\mathbf{exp}\,(\mathbf{X}\boldsymbol{\beta})$ denotes the vector exponential operator, that consists in applying the exponential function to each of the elements of $\mathbf{X}\boldsymbol{\beta}$, and $\nu_s = -\ln\left[\mathbf{1}_R'\mathbf{exp}\,(\mathbf{X}_s\boldsymbol{\beta})\right]$, $s = 1, \ldots, S$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector which embodies the $p \leq$

$S(R-1)$ unknown parameters, and $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_S')'$ is a $SR \times p$ matrix, such that each $R \times p$ submatrix $\mathbf{X}_s$ has its columns linearly independent from the vector $\mathbf{1}_R$ which defines the s-th natural constraint, $\mathbf{1}_R'\boldsymbol{\theta}_s = 1$, *i.e.*, $r(\mathbf{1}_R, \mathbf{X}_s) = 1 + r(\mathbf{X}_s)$, $s = 1, \ldots, S$, and $r(\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}) = S + p$. Rewriting (43), we may obtain $\boldsymbol{\theta}$ from $\boldsymbol{\beta}$ by

$$\boldsymbol{\theta}(\boldsymbol{\beta}) = \mathbf{D}_{\boldsymbol{\psi}}^{-1}\mathbf{exp}\left(\mathbf{X}\boldsymbol{\beta}\right), \tag{44}$$

where $\boldsymbol{\psi} = [\mathbf{I}_S \otimes (\mathbf{1}_R\mathbf{1}_R')]\,\mathbf{exp}\left(\mathbf{X}\boldsymbol{\beta}\right) = \left(\boldsymbol{\psi}_s', s = 1, \ldots, S\right)'$, $\boldsymbol{\theta}(\boldsymbol{\beta}) = \left(\boldsymbol{\theta}_s'(\boldsymbol{\beta}), s = 1, \ldots, S\right)'$, $\boldsymbol{\theta}_s(\boldsymbol{\beta}) = \mathbf{D}_{\boldsymbol{\psi}_s}^{-1}\mathbf{exp}\left(\mathbf{X}_s\boldsymbol{\beta}\right)$, and $\boldsymbol{\psi}_s = (\mathbf{1}_R\mathbf{1}_R')\,\mathbf{exp}\left(\mathbf{X}_s\boldsymbol{\beta}\right)$.

We can also consider a wider class of log-linear models, expressed by

$$M_{LL} : \mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}, \tag{45}$$

where $\mathbf{A}$ is an $u \times SR$ matrix with rank $r(\mathbf{A}) = u \leq S(R-1)$ and its rows are orthogonal to the columns of $\mathbf{I}_S \otimes \mathbf{1}_R$, *i.e.*, $\mathbf{A}\,(\mathbf{I}_S \otimes \mathbf{1}_R) = \mathbf{0}_{u,S}$. For instance, $\mathbf{A} = \mathbf{I}_S \otimes [\mathbf{I}_{R-1}, -\mathbf{1}_{R-1}]$ generates logits with the baseline category R. If $u = S(R-1)$, the $S(R-1) \times p$ matrix $\mathbf{X}_L$ has the following relations with $\mathbf{X}$: $\mathbf{X}_L = \mathbf{A}\mathbf{X}$ and $\mathbf{X} = \mathbf{A}'\left(\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{X}_L$. If $u < S(R-1)$, we need a $(S[R-1]-u) \times SR$ matrix $\mathbf{A}_0$ base of the orthocomplement $(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R)'$, such that the model to be fitted

$$M_{LL} : \begin{pmatrix} \mathbf{A} \\ \mathbf{A}_0 \end{pmatrix} \mathbf{ln}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{X}_L\boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \end{pmatrix} \tag{46}$$

can be written in the form of (43) by

$$M_{LL} : \mathbf{ln}(\boldsymbol{\theta}) = [\mathbf{I}_S \otimes \mathbf{1}_R]\,\boldsymbol{\nu} + \begin{pmatrix} \mathbf{A}'\left(\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{X}_L & , & \mathbf{A}_0'\left(\mathbf{A}_0\mathbf{A}_0'\right)^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \end{pmatrix}. \tag{47}$$

For parsimony, we suppress the $\boldsymbol{\beta}_0$ parameter in the following exposition, mentioning solely $\boldsymbol{\beta}$. However, whenever that $u < S(R-1)$, $\boldsymbol{\beta}_0$ is also used and we would take $\mathbf{X} = \left(\mathbf{A}'\left(\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{X}_L, \mathbf{A}_0'\left(\mathbf{A}_0\mathbf{A}_0'\right)^{-1}\right)$.

The freedom equations formulations (43) and (45) are equivalent, respectively, to the constraint formulations

$$\mathbf{U}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{S(R-1)-p}, \tag{48}$$

$$\mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}, \tag{49}$$

where $\mathbf{U}$ ($\mathbf{U}_L$) is a $[S\{R-1\}-p] \times SR$ ($[u-p] \times u$) matrix defining the $S[R-1]-p$ ($u-p$) constraints, with full rank and its rows are orthogonal to the columns $[\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}]$ ($\mathbf{X}_L$), *i.e.*, $\mathbf{U}[\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}] = \mathbf{0}_{(SR-p),p}$ ($\mathbf{U}_L\mathbf{X}_L = \mathbf{0}_{(u-p),p}$).

Differentiating $\ln L_1\left(\boldsymbol{\theta}(\boldsymbol{\beta})\,|\mathbf{N}\right)$ with respect to $\boldsymbol{\beta}$, we obtain the score vector

$$\mathbf{S}_{1LL}\left(\boldsymbol{\beta}\right)=\sum_{s=1}^{S}\mathbf{X}_s'\left\{\mathbf{N}_{s1}+\sum_{t=2}^{T_s}\left[\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{Z}_{st}\mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{N}_{st}\right]-n_{s++}\boldsymbol{\theta}_s(\boldsymbol{\beta})\right\}. \tag{50}$$

Further differentiation of the gradient of $L_1\left(\boldsymbol{\theta}(\boldsymbol{\beta})\,|\mathbf{N}\right)$ with respect to $\boldsymbol{\beta}'$ leads to the hessian matrix

$$\mathbf{H}_{1LL}\left(\boldsymbol{\beta}\right)=\sum_{s=1}^{S}\mathbf{X}_s'\left\{-n_{s++}\mathbf{I}_R+\sum_{t=2}^{T_s}\left[\mathbf{D}_{\mathbf{u}_{st}^{I}}-\mathbf{D}_{\mathbf{u}_{st}^{II}}\mathbf{Z}_{st}\mathbf{Z}_{st}'\right]\right\}\left\{\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}-\boldsymbol{\theta}_s(\boldsymbol{\beta})\left[\boldsymbol{\theta}_s(\boldsymbol{\beta})\right]'\right\}\mathbf{X}_s, \tag{51}$$

where $\mathbf{u}_{st}^{I}=\mathbf{Z}_{st}\mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{N}_{st}$ and $\mathbf{u}_{st}^{II}=\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{Z}_{st}\mathbf{D}^{-2}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{N}_{st}$. The Fisher information matrix under the $\mathcal{M}$ (MAR or MCAR) mechanism is expressed by

$$\boldsymbol{\mathcal{I}}_{1LL}\left(\boldsymbol{\beta},\left\{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\right\}\right)=\sum_{s=1}^{S}\mathbf{X}_s'\left\{n_{s++}\mathbf{I}_R-\sum_{t=2}^{T_s}\left[\mathbf{D}_{\mathbf{v}_{st}^{\mathcal{M}}}-\mathbf{D}_{\mathbf{w}_{st}^{\mathcal{M}}}\mathbf{Z}_{st}\mathbf{Z}_{st}'\right]\right\}\times$$

$$\left\{\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}-\boldsymbol{\theta}_s(\boldsymbol{\beta})\left[\boldsymbol{\theta}_s(\boldsymbol{\beta})\right]'\right\}\mathbf{X}_s, \tag{52}$$

where

$$\begin{aligned}\mathbf{v}_{st}^{\mathrm{MAR}}&=n_{s++}\mathbf{Z}_{st}\boldsymbol{\alpha}_{st}^{\mathrm{MAR}}, & \mathbf{w}_{st}^{\mathrm{MAR}}&=n_{s++}\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{Z}_{st}\mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\boldsymbol{\alpha}_{st}^{\mathrm{MAR}},\\ \mathbf{v}_{st}^{\mathrm{MCAR}}&=n_{s++}\boldsymbol{\alpha}_{t(s)}\mathbf{1}_R, & \mathbf{w}_{st}^{\mathrm{MCAR}}&=n_{s++}\boldsymbol{\alpha}_{t(s)}\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{Z}_{st}\mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{1}_{R_{st}}.\end{aligned}$$

These expressions allow to get ML estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by using the Newton-Raphson or Fisher scoring algorithm. The iterative processes may be initialized with the WLS estimate (special case of what we will present in Section 5.2)

$$\hat{\boldsymbol{\beta}}^{(0)}=\left[\mathbf{X}_L'\left(\mathbf{A}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\right)^{-1}\mathbf{X}_L\right]^{-1}\mathbf{X}_L'\left(\mathbf{A}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\right)^{-1}\mathbf{A}\,\ln\!\left(\hat{\boldsymbol{\theta}}\right), \tag{53}$$

if $u=S(R-1)$, or by using an analogous expression deriving from replacing $\mathbf{A}$ by $\left(\mathbf{A}',\mathbf{A}_0'\right)'$ and $\mathbf{X}_L$ by

$$\begin{pmatrix}\mathbf{X}_L & \mathbf{0}_{u,S(R-1)-u}\\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u}\end{pmatrix}, \tag{54}$$

if $u<S(R-1)$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\mathcal{M}}$, an estimate of its corresponding approximate covariance matrix under the mechanism $\mathcal{M}$, obtained according to the suggestion presented in Section 4.1.

Once obtained the ML estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, an estimate of its approximate covariance matrix is $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}LL}^{\mathcal{M}}=\left[\boldsymbol{\mathcal{I}}_{1LL}\left(\hat{\boldsymbol{\beta}},\left\{\hat{\boldsymbol{\alpha}}_{st}^{\mathcal{M}}\right\}\right)\right]^{-1}$. By means of (44), we obtain the ML estimate $\hat{\boldsymbol{\theta}}(M_{LL})$ of $\boldsymbol{\theta}$ under

M_{LL} and, using the *delta* method, an estimate of its approximate covariance matrix under the $\mathcal{M}$ mechanism is

$$\hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\theta}}(M_{LL})} = \widehat{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'}} \hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\beta}}LL} \left(\widehat{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'}} \right)' = \hat{\mathbf{V}}_{LL} \mathbf{X} \hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\beta}}LL} \mathbf{X}' \hat{\mathbf{V}}_{LL}, \tag{55}$$

where $\hat{\mathbf{V}}_{LL}$ is a block diagonal matrix with blocks equal to $\mathbf{D}_{\boldsymbol{\theta}_s(\hat{\boldsymbol{\beta}})} - \boldsymbol{\theta}_s(\hat{\boldsymbol{\beta}}) \big[\boldsymbol{\theta}_s(\hat{\boldsymbol{\beta}})\big]'$, $s = 1, \ldots, S$. The ML estimates of the log-linear functions $\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta})$ under M_{LL} are $\mathbf{X}_L \hat{\boldsymbol{\beta}}$ and, by means of the *delta* method, an estimate of its approximate covariance matrix is given by $\hat{\mathbf{V}}^{\mathcal{M}}_{\mathbf{A}\,\mathbf{ln}\left(\hat{\boldsymbol{\theta}}(M_{LL})\right)} = \mathbf{X}_L \hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\beta}}LL} \mathbf{X}'_L$.

Now, let $\mathcal{M}$ be a missingness mechanism more restrictive than MAR (*e.g.*, MCAR) and let M be a reduced model for $\boldsymbol{\theta}$ (*e.g.*, M_L or M_{LL}). The Wilks likelihood ratio test for the joint model $(M, \mathcal{M})$ conditional on the assumed MAR mechanism can be partitioned into the sum of the corresponding test statistics, separately, for M and $\mathcal{M}$, *i.e.*,

$$\begin{aligned} Q_L(M, \mathcal{M}|\mathrm{MAR}) &= -2\ \ln \frac{L_1\big(\hat{\boldsymbol{\theta}}(M)\,|\mathbf{N}; M\big) L_2\left(\{\hat{\alpha}_{t(cs)}(\mathcal{M})\}\,|\mathbf{N}; \mathcal{M}\right)}{L_1\big(\hat{\boldsymbol{\theta}}\,|\mathbf{N}\big) L_2\left(\{\hat{\alpha}_{t(cs)}\}\,|\mathbf{N}; \mathrm{MAR}\right)} \\ &= Q_L(M) + Q_L(\mathcal{M}|\mathrm{MAR}) \end{aligned} \tag{56}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model, and $\hat{\boldsymbol{\theta}}(M)$, under the model M; $\{\hat{\alpha}_{t(cs)}\}$ are the ML estimates of $\{\alpha_{t(cs)}\}$ under the MAR mechanism, and $\{\hat{\alpha}_{t(cs)}(\mathcal{M})\}$, under the mechanism $\mathcal{M}$. As noted by Williamson and Haber (1994), this partition of Q_L shows that the comparison of any pair of models for the marginal probabilities of categorization and for the conditional probabilities of missingness does not depend on, respectively, the more restrictive missingness mechanism and the more reduced model for $\boldsymbol{\theta}$ that are imposed. If the parameter of interest is just $\boldsymbol{\theta}$, the likelihood ratio statistic for the goodness-of-fit test of the model M is expressed by

$$Q_L(M|\mathcal{M}) = Q_L(M) = -2\ \ln \frac{L_1\big(\hat{\boldsymbol{\theta}}(M)\,|\mathbf{N}\big)}{L_1\big(\hat{\boldsymbol{\theta}}\,|\mathbf{N}\big)} = -2 \sum_{s=1}^{S} \mathbf{N}'_s \left\{ \mathbf{ln}\left[\mathbf{Z}'_s \hat{\boldsymbol{\theta}}_s(M)\right] - \mathbf{ln}\left[\mathbf{Z}'_s \hat{\boldsymbol{\theta}}_s\right] \right\}, \tag{57}$$

being independent of the mechanism $\mathcal{M}$ more constrained than the MAR that is assumed.

The Pearson and Neyman goodness-of-fit statistics for testing (M, MCAR) conditionally on the MAR mechanism are

$$\begin{aligned} Q_P(M, \mathrm{MCAR}|\mathrm{MAR}) &= \sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{c=1}^{R_{st}} \frac{\Big(n_{stc} - n_{st+} \mathbf{z}'_{stc} \hat{\boldsymbol{\theta}}_s(M)\Big)^2}{n_{st+} \mathbf{z}'_{stc} \hat{\boldsymbol{\theta}}_s(M)} \\ &= \sum_{s=1}^{S} \Big(\mathbf{p}_s - \mathbf{Z}'_s \hat{\boldsymbol{\theta}}_s(M)\Big)' \left[\mathbf{D}_{\mathbf{N}_{s+}} \mathbf{D}^{-1}_{\mathbf{Z}'_s \hat{\boldsymbol{\theta}}_s(M)}\right] \Big(\mathbf{p}_s - \mathbf{Z}'_s \hat{\boldsymbol{\theta}}_s(M)\Big), \end{aligned} \tag{58}$$

$$Q_N(M, \text{MCAR}|\text{MAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{stc} - n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\right)^2}{n_{stc}}$$
$$= \sum_{s=1}^{S}\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)\right)'\left[\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}_{\mathbf{p}_s}^{-1}\right]\left(\mathbf{p}_s - \mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)\right). \quad (59)$$

The corresponding statistics for testing the model M conditionally on the MAR or the MCAR mechanisms are

$$Q_P(M|\text{MAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{stc} - n_{s++}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\hat{\alpha}_{t(cs)}\right)^2}{n_{s++}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\hat{\alpha}_{t(cs)}}$$
$$= \sum_{s=1}^{S}\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right)'\left[\mathbf{D}_{\mathbf{N}_s}\mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s}\mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)}\right]\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right), \quad (60)$$

$$Q_N(M|\text{MAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{stc} - n_{s++}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\hat{\alpha}_{t(cs)}\right)^2}{n_{stc}}$$
$$= \sum_{s=1}^{S}\left(\mathbf{1}_{R+l_s} - \mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s}\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)\right)'\mathbf{D}_{\mathbf{N}_s}\left(\mathbf{1}_{R+l_s} - \mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s}\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)\right), \quad (61)$$

$$Q_P(M|\text{MCAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s - n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\right)^2}{n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)}$$
$$= \sum_{s=1}^{S}\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right)'\left[\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s(M)}\right]\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right), \quad (62)$$

$$Q_N(M|\text{MCAR}) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{\left(n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s - n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s(M)\right)^2}{n_{st+}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s}$$
$$= \sum_{s=1}^{S}\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right)'\left[\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\hat{\boldsymbol{\theta}}_s}\right]\left(\mathbf{Z}'_s\left[\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_s(M)\right]\right), \quad (63)$$

where $\mathbf{N}_{s+} = \left(n_{st+} \otimes \mathbf{1}'_{R_{st}}, t = 1, \ldots, T_s\right)'$ and $\hat{\alpha}_{t(cs)} = n_{stc}/(n_{s++}\mathbf{z}'_{stc}\hat{\boldsymbol{\theta}}_s)$. These expressions highlight that we no longer have the same advantageous result provided by the likelihood ratio statistic.

The Wald statistics for testing, respectively, M_L and M_{LL} conditionally on the missingness mechanism $\mathcal{M}$ (MAR or MCAR) are

$$Q_W(M_L|\mathcal{M}) = \left(\mathbf{U}\mathbf{A}\hat{\boldsymbol{\theta}}\right)'\left[\mathbf{U}\mathbf{A}\hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\right]^{-1}\mathbf{U}\mathbf{A}\hat{\boldsymbol{\theta}}, \quad (64)$$

$$Q_W(M_{LL}|\mathcal{M}) = \left(\mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}\big(\hat{\boldsymbol{\theta}}\big)\right)'\left[\mathbf{U}\mathbf{A}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\theta}}}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\right]^{-1}\mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}\big(\hat{\boldsymbol{\theta}}\big). \quad (65)$$

Asymptotically, under the model M and the MAR mechanism

$$Q_V(M) \overset{a}{\approx} Q_P(M|\text{MAR}) \overset{a}{\approx} Q_N(M|\text{MAR}) \overset{a}{\approx} Q_W(M|\text{MAR}) \xrightarrow{a} \chi^2_{(u-p)}$$

and, additionally under the MCAR mechanism,

$$Q_P(M|\text{MCAR}) \overset{a}{\approx} Q_N(M|\text{MCAR}) \overset{a}{\approx} Q_W(M|\text{MCAR}) \xrightarrow{a} \chi^2_{(u-p)},$$

$$Q_V(M, \text{MCAR}|\text{MAR}) \overset{a}{\approx} Q_P(M, \text{MCAR}|\text{MAR}) \overset{a}{\approx} Q_N(M, \text{MCAR}|\text{MAR}) \xrightarrow{a} \chi^2_{(u-p+g)},$$

where $g = S + \sum_{s=1}^{S}(l_s - T_s)$.

When there is interest in performing a reduction in the dimension of $\boldsymbol{\beta}$ by means of a hypothesis $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{C}_0$, where $\mathbf{C}_0$ is a $c \times 1$ vector with known elements (usually, $\mathbf{C}_0 = \mathbf{0}_c$), $\mathbf{C}$ is a $c \times p$ matrix with full rank c ($\leq p$) and its rows define the contrasts of interest, we may appeal to the Wald statistic

$$Q_W(H|M, \mathcal{M}) = \left(\mathbf{C}\hat{\boldsymbol{\beta}}(M) - \mathbf{C}_0\right)' \left[\mathbf{C}\hat{\mathbf{V}}^{\mathcal{M}}_{\hat{\boldsymbol{\beta}}(M)}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}}(M) - \mathbf{C}_0\right), \tag{66}$$

which has an asymptotically null distribution $\chi^2_{(c)}$.

5.2 WLS inferences on functional linear models under MAR, MCAR and MNAR assumptions

For the purpose of analysis of functional linear models of $\boldsymbol{\theta}$ under the validity of the MCAR mechanism, Koch *et al.* (1972) proposed the application of the WLS methodology in a second stage to the WLS estimate $\tilde{\boldsymbol{\theta}}$. In the light of the functional asymptotic regression for complete data according to Imrey *et al.* (1981, 1982) and used in different contexts by Koch, Singer and Amara (1985) and Ho and Singer (2001), it is also possible to apply the WLS approach in a second phase to the ML estimate $\hat{\boldsymbol{\theta}}$ under any missingness mechanism as Paulino (1991) suggested. Using this hybrid methodology, we hope to draw inferences about $\boldsymbol{\theta}$ more easily, mainly in the context of non-ignorable models for the missingness mechanism, by means of procedures that continue to possess good properties in large samples.

Let $\tilde{\boldsymbol{\theta}}$ denote any consistent estimator of $\boldsymbol{\theta}$ reflecting all the available data, for instance, the WLS estimator under the MCAR mechanism (Section 4.2), the ML estimator under the MAR or the MCAR mechanism (Section 4.1), or even the ML estimator under an assumed MNAR mechanism. In the same way, let $\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}$ represent an estimate deriving from a consistent estimator of the covariance matrix of $\tilde{\boldsymbol{\theta}}$ under the same missingness mechanism $\mathcal{M}$.

We consider the functional linear model in terms of freedom equations

$$M_F : \mathbf{F} \equiv \mathbf{F}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}, \tag{67}$$

where $\mathbf{F}(\boldsymbol{\theta}) = (F_i(\boldsymbol{\theta}), i = 1, \ldots, u)'$ is an $u \times 1$ vector that defines the $u \leq S(R-1)$ functions of interest, and it is such that $\mathbf{G} \equiv \mathbf{G}(\boldsymbol{\theta}) = \partial\mathbf{F}/\partial\boldsymbol{\theta}'$ and $\partial^2\mathbf{F}/(\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}')$ exist and are continuous in an open subset containing $\boldsymbol{\theta}$; $\mathbf{X}$ is an $u \times p$ matrix with rank $r(\mathbf{X}) = p \leq u$ which specifies the model, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector with unknown parameters. The constrained formulation, equivalently to (67), is expressed by

$$M_F : \mathbf{U}\,\mathbf{F}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}, \tag{68}$$

where $\mathbf{U}$ is an $(u-p) \times u$ matrix which contains the $u-p$ constrains, has full rank and its rows are orthogonal to the columns of $\mathbf{X}$, *i.e.*, $\mathbf{U}\mathbf{X} = \mathbf{0}_{(u-p),p}$.

Under the assumption that $\tilde{\boldsymbol{\theta}}$ is approximately distributed as a multivariate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}$, we have that $\tilde{\mathbf{F}} \equiv \mathbf{F}\big(\tilde{\boldsymbol{\theta}}\big) \overset{a}{\sim} N_u\left(\mathbf{F}, \tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}\right)$, where we assume that $\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}} = \tilde{\mathbf{G}}\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{G}}'$, with $\tilde{\mathbf{G}} \equiv \mathbf{G}\big(\tilde{\boldsymbol{\theta}}\big)$, is nonsingular. Therefore, the WLS estimator of $\boldsymbol{\beta}$ of (67) is given by

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\tilde{\mathbf{F}}, \tag{69}$$

and an estimate of its approximate covariance matrix can be obtained by

$$\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}} = \left(\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\mathbf{X}\right)^{-1}. \tag{70}$$

The WLS estimator of the functions $\mathbf{F}$ under M_F is obtained by $\mathbf{X}\tilde{\boldsymbol{\beta}}$ and, by means of the *delta* method, an estimate of its approximate covariance matrix is $\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}(M_F)} = \mathbf{X}\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}}\mathbf{X}'$.

The goodness-of-fit test of the model M_F conditionally to the missingness mechanism $\mathcal{M}$ can be accomplished with the Wald statistic

$$Q_W(M_F|\mathcal{M}) = \big(\mathbf{U}\tilde{\mathbf{F}}\big)'\left[\mathbf{U}\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}\mathbf{U}'\right]^{-1}\mathbf{U}\tilde{\mathbf{F}}, \tag{71}$$

which has an asymptotic null distribution $\chi^2_{(u-p)}$. Reductions in the dimension of $\boldsymbol{\beta}$ may also be assessed with a Wald test analogously to (66).

In many cases, the vector $\mathbf{F}(\boldsymbol{\theta})$ may be expressed as composition of functions: linear, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_1\boldsymbol{\theta} \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{A}_1$; logarithmic, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{ln}(\boldsymbol{\theta}) \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\boldsymbol{\theta}}^{-1}$; exponential, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{exp}(\boldsymbol{\theta}) \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\mathbf{exp}(\boldsymbol{\theta})}$; and addition of constants, $\mathbf{F}(\boldsymbol{\theta}) = \boldsymbol{\pi}_1 + \boldsymbol{\theta} \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{I}_{SR}$; where $\mathbf{A}_1$ is an $u \times SR$ matrix, with $u \leq S(R-1)$, and $\boldsymbol{\pi}_1$ is a $SR \times 1$ vector with known constants. Some examples of compounded functions and associated first derivatives are $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_1\,\mathbf{ln}(\boldsymbol{\theta}) \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{A}_1\,\mathbf{D}_{\boldsymbol{\theta}}^{-1}$

and $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{exp}[\mathbf{A}_1\ \mathbf{ln}(\boldsymbol{\theta})] \rightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\mathbf{exp}[\mathbf{A}_1\ \mathbf{ln}(\boldsymbol{\theta})]}\ \mathbf{A}_1\ \mathbf{D}_{\boldsymbol{\theta}}^{-1}$. Note that these last two matrices $\mathbf{G}(\boldsymbol{\theta})$ were obtained using the chain rule for differentiation.

When $\mathbf{F}(\boldsymbol{\theta})$ is as M_L or M_{LL} in Section 5.1, we may obtain an estimate of $\boldsymbol{\theta}$ under M_F and an estimate of its approximate covariance matrix by means of analogous developments therein presented.

Acknowledgements

This research received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Fundação Calouste Gulbenkian.

References

Agresti, A. (2002). *Categorical data analysis*. 2nd ed. New York: John Wiley & Sons.

Baker, S.G. (1994). Missing data: composite linear models for incomplete multinomial data. *Statistics in Medicine* **13**, 609-622.

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, 1232 (correction).

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association* **63**, 542-551.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **39**, 1-38.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003). *Statistical methods for rates and proportions*. 3rd ed. New York: John Wiley & Sons.

Forthofer, R.N. and Lehnen, R.G. (1981). *Public program analysis: a new categorical data approach*. Belmont: Wadsworth.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association* **77**, 270-278.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489-504.

Ho, L.L. and Singer, J.M. (2001). Generalized least squares methods for bivariate Poisson regression. *Communications in Statistics, Theory and Methods* **30**, 263-277.

Hocking, R.R. and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association* **66**, 65-70.

Imrey, P.B., Koch, G.G., Stokes, M.E. *et al.* (1981). Categorical data analysis: some reflections on the log linear model and logistic regression. Part I: historical and methodological overview. *International Statistical Review* **49**, 265-283.

Imrey, P.B., Koch, G.G., Stokes, M.E. *et al.* (1982). Categorical data analysis: some reflections on the log linear model and logistic regression. Part II: data analysis. *International Statistical Review* **50**, 35-63.

Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science* **13**, 236-247.

Koch, G.G., Imrey, P.B. and Reinfurt, D.W. (1972). Linear model analysis of categorical data with incomplete response vectors. *Biometrics* **28**, 663-692.

Koch, G.G., Imrey, P.B., Singer, J.M., Atkinson, S.S. and Stokes, M.E. (1985). *Analysis of categorical data*. Montréal: Les Presses de L'Université de Montréal.

Koch, G.G., Singer, J.M. and Amara, I.A. (1985). A two-stage procedure for the analysis of ordinal categorical data. In *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, ed. P.K. Sen. North Holland: Elsevier Science. 357-387.

Landis, J.R., Stanish, W.M., Freeman, J.L. and Koch, G.G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Computer Programs in Biomedicine* **6**, 196-231.

Lipsitz, S.R. and Fitzmaurice, G.M. (1996). The score test for independence in $R \times C$ contingency tables with missing data. *Biometrics* **52**, 751-762.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **90**, 1112-1121.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. 2nd ed. New York: John Wiley & Sons.

Molenberghs, G. and Goetghebeur, E. (1997). Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **59**, 401-414.

Molenberghs, G., Kenward, M.G. and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **50**, 15-29.

Paulino, C.D. (1991). Analysis of incomplete categorical data: a survey of the conditional maximum likelihood and weighted least squares approaches. *Brazilian Journal of Probability and Statistics* **5**, 1-42.

Paulino, C.D. and Silva, G.L. (1999). On the maximum likelihood analysis of the general linear model in categorical data. *Computational Statistics and Data Analysis* **30**, 197-204.

Paulino, C.D. and Singer; J.M. (2006). *Analysis of categorical data*. Edgard Blücher: São Paulo (in Portuguese).

R Development Core Team (2006). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69**, 467-474.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Boca Raton: Chapman & Hall / CRC.

Soares, P.J.J. and Paulino, C.D. (2001). Incomplete categorical data analysis: a bayesian perspective. *Journal of Statistical Computation and Simulation* **69**, 157-170.

Ware, J.H., Dockery, D.W., Spiro III, A., Speizer, F.E. and Ferris Jr., B.G. (1984). Passive smoking, gas cooking and respiratory health of children living in six cities. *American Review of Respiratory Diseases* **129**, 366-374.

Williamson, G.D. and Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics* **49**, 194-203.