Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 1, 109–139
DOI: 10.1214/12-BJPS198

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

Frederico Z. Poleto[a], Julio M. Singer[a] and Carlos Daniel Paulino[b]

[a]*Universidade de São Paulo*
[b]*Universidade Técnica de Lisboa (and CEAUL-FCUL)*

Abstract. With the objective of analysing categorical data with missing responses, we extend the multinomial modelling scenario described by Paulino (*Braz. J. Probab. Stat.* **5** (1991) 1–42) to a product-multinomial framework that allows the inclusion of explanatory variables. We consider maximum likelihood (ML) and weighted least squares (WLS) as well as a hybrid ML/WLS approach to fit linear, log-linear and more general functional linear models under ignorable and nonignorable missing data mechanisms. We express the results in an unified matrix notation that may be easily used for their computational implementation and develop such a set of subroutines in R. We illustrate the procedures with the analysis of two data sets, and perform simulations to assess the properties of the estimators.

1 Introduction

A common pattern of missing (partially or incompletely classified) responses observed in the collection of categorical data may be illustrated with the following examples.

Example 1. The data in Table 1 were extracted from Lipsitz and Fitzmaurice (1996) and relate to the evaluation of the association between wheezing status in children and maternal smoking habits.

Example 2. The data in Table 2 contain information on the obesity status (yes or no) of 5 to 15 years old children (in 1977) of both genders which participated in one or more surveys in 1977, 1979 and 1981. The objective is to estimate the probability of obesity as a function of gender and age. See Woolson and Clarke (1984) for more details.

The genesis of the missingness pattern in Example 1 lies in the incomplete classification of subjects with respect to one of the intervening variables. In Example 2, the missingness is related to the lack of observation of the response in one or more instants of the longitudinal study.

Key words and phrases. EM algorithm, incomplete data, missing data, missingness mechanism, selection models.

Received January 2012; accepted May 2012.

Table 1 *Observed frequencies of maternal smoking cross-classified by child's wheezing status and home city*

		Child's wheezing status			
Home city	Maternal smoking	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	

Methods for drawing inferences from partially or incompletely classified categorical data have been widely considered in the literature. Early accounts may be found in Blumenthal (1968) and Hocking and Oxspring (1971), for example. Analyses of categorical data assuming missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) mechanisms have been proposed by many authors under different approaches; among them, we mention Koch et al. (1972), Chen and Fienberg (1974) and Molenberghs et al. (1999). For details on the taxonomy for missing data the reader is referred to Rubin (1976) and Little and Rubin (2002).

In particular, Paulino (1991) considered fitting strictly linear and log-linear multinomial models to data generated by MAR and MCAR mechanisms via maximum likelihood (ML) methods and more general functional linear models to data generated by MCAR mechanisms via weighted least squares (WLS) methodology. In the spirit of functional asymptotic regression methodology described by Imrey et al. (1981, 1982) for complete data, he also proposed a hybrid methodology, where simple models are fitted via ML to the data under ignorable (MAR or MCAR) or nonignorable (MNAR) mechanisms in the first stage and the estimated marginal probabilities of categorization and their covariance matrix are used in a second stage to fit more general functional linear models via WLS. In most cases, this approach is computationally simpler than and asymptotically as efficient as the pure ML approach. Paulino and Soares (2003) extended the results to cover situations where the data follow product-Poisson distributions. We review such methods placing them in a more general setup where the underlying distribution is product-multinomial. We also extend the ML methodology considered in Paulino (1991) to more general classes of linear and log-linear models.

In general, the ML approach to the analysis of categorical data requires iterative procedures to compute the estimates of the pertinent parameters. The EM

Table 2 *Observed frequencies of children classified by gender, age (years) and obesity status*

Obesity status† in			Gender:	boy					girl				
1977	1979	1981	Age in 1977:	5–7	7–9	9–11	11–13	13–15	5–7	7–9	9–11	11–13	13–15
n	n	n		90	150	152	119	101	75	154	148	129	91
n	n	o		9	15	11	7	4	8	14	6	8	9
n	o	n		3	8	8	8	2	2	13	10	7	5
n	o	o		7	8	10	3	7	4	19	8	9	3
o	n	n		0	8	7	13	8	2	2	12	6	6
o	n	o		1	9	7	4	0	2	6	0	2	0
o	o	n		1	7	9	11	6	1	6	8	7	6
o	o	o		8	20	25	16	15	8	21	27	14	15
n	n	m		16	38	48	42	82	20	25	36	36	83
n	o	m		5	3	6	4	9	0	3	0	9	15
o	n	m		0	1	2	4	8	0	1	7	4	6
o	o	m		0	11	14	13	12	4	11	17	13	23
n	m	n		9	16	13	14	6	7	16	8	31	5
n	m	o		3	6	5	2	1	2	3	1	4	0
o	m	n		0	1	0	1	0	0	0	1	2	0
o	m	o		0	3	3	4	1	1	4	4	6	1
m	n	n		129	42	36	18	13	109	47	39	19	11
m	n	o		18	2	5	3	1	22	4	6	1	1
m	o	n		6	3	4	3	2	7	1	7	2	2
m	o	o		13	13	3	1	2	24	8	13	2	3
n	m	m		32	45	59	82	95	23	47	53	58	89
o	m	m		5	7	17	24	23	5	7	16	37	32
m	n	m		33	33	31	23	34	27	23	25	21	43
m	o	m		11	4	9	6	12	5	5	9	1	15
m	m	n		70	55	40	37	15	65	39	23	23	14
m	m	o		24	14	9	14	3	19	13	8	10	5

†n indicates not obese, o, obese, and m, missing.

algorithm (Dempster et al., 1977) has been used to obtain ML estimates based on expected cell frequencies in augmented tables; see, for example, Fuchs (1982) and Baker and Laird (1988) for, respectively, ignorable and nonignorable models for the missingness mechanism. Molenberghs and Goetghebeur (1997) considered the advantages of using Newton–Raphson and Fisher's scoring algorithms, and Baker (1994) suggested a combination of EM and Newton–Raphson algorithms for such purposes. Although these methods embrace a part of the models described in this paper, they are not yet available in many of the current commercial statistical software, either because they need further input of the derivatives, adaptation of the available computational procedures, and/or additional programming. Some excep-

tions are multiple imputation methods (Rubin, 1987)—available in SAS (PROC MI and MIANALYZE) and R/S-Plus (mitools package)—and saturated and hierarchical log-linear multinomial models (Schafer, 1997)—available in R/S-Plus (cat package).

The matrix approach we adopt allows a unified and general formulation of models and inferential procedures that may be easily employed for their computational implementation. We developed subroutines written in R (R Development Core Team, 2012) for such purposes. To obtain the ML estimates for the first-stage models under nonignorable mechanisms, built-in optimization functions in R are required. Model formulation and use of the functions are similar to those considered in GENCAT (Landis et al., 1976) or SAS' PROC CATMOD. The distinctive feature of the proposed functions is that they allow the analysis of complete and incomplete categorical data in a unified way. Hopefully, the package ACD will be available in CRAN soon; meanwhile, ACD as well as the commands to reproduce the analyses presented in this paper may be downloaded from http://www.poleto.com/missing.html.

In order to present an overview of the pertinent statistical methods, we first introduce the problem and the notation in Section 2 and describe the probabilistic model along with the missing data generating mechanisms in Section 3. In Section 4, we present the ML and WLS approaches to obtain inferential results for saturated models, that is, where no structural constraints are imposed on the probabilities of categorization. In Section 5, we describe the ML methodology for fitting linear and log-linear models as well as the WLS and the hybrid ML/WLS approaches for fitting more general functional linear models. We apply the methods to the couple of aforementioned examples in Section 6. In Section 7, we conduct simulations to assess the properties of some of the estimators in small to moderate sized samples. Some concluding remarks are presented in Section 8.

2 Problem description and notation

For simplicity, we admit that the random vector $\mathbf{Y} = (Y_1, \ldots, Y_k)'$ of response variables can assume R values, corresponding to combinations of the levels of its components. 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, we have $R = 2 \times 3 \times 5 = 30$. Likewise, we assume that the vector $\mathbf{X} = (X_1, \ldots, X_q)'$ of explanatory variables can take S values (each of which defining a subpopulation of interest), corresponding to combinations of the levels of its components. The R response categories are indexed by r, and the S subpopulations, by s.

We assume that each of the n_{s++} sampling units randomly selected from the sth subpopulation can be independently classified into the rth response category with the same probability $\theta_{r(s)}$, $r = 1, \ldots, R$, $s = 1, \ldots, S$. This implies that the

$n_{+++} = \sum_{s=1}^{S} n_{s++}$ units are (at least conceptually) obtained according to a stratified random sampling scheme with sample sizes for the S strata given by the elements of 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 in $\mathbf{Y}$, so only part of the n_{s++} units in the sth stratum is completely classified into one of the R originally defined response categories, while the remaining units are associated to some type of missingness. For subpopulation s, we define T_s missingness patterns as follows. The set of units with no missing data (i.e., with complete classification) is indexed by $t = 1$ and the sets that have some degree of missingness, by $t = 2, \ldots, T_s$. We assume that each unit with the tth missingness pattern is recorded in one of R_{st} response classes $\mathcal{C}_{stc}$, $c = 1, \ldots, R_{st}$, defined by at least two of the R response categories and such that $\mathcal{C}_{stc} \cap \mathcal{C}_{std} = \varnothing$, $c \neq d$ and $\bigcup_{c=1}^{R_{st}} \mathcal{C}_{stc} = \{1, \ldots, R\}$. Thus, the response classes for the units with the tth pattern form a partition $\mathcal{P}_{st} = \{\mathcal{C}_{stc}, c = 1, \ldots, R_{st}\}$ of the set of response categories for units with complete classification, that is, $\mathcal{P}_{s1} = \mathcal{P}_1 = \{\{r\}, r = 1, \ldots, R\}$. For notational simplicity, we let $\mathcal{C}_{s1r} = \mathcal{C}_{1r} = \{r\}$ and $R_{s1} = R_1 = R$. We represent the total number of response classes for units with some missingness pattern in the sth subpopulation by $l_s = \sum_{t=2}^{T_s} R_{st}$.

For mathematical convenience, we consider $(R \times 1)$-dimensional indicator vectors $\mathbf{z}_{stc}$, the elements of which are equal to 1 for the positions corresponding to the response categories in the class $\mathcal{C}_{stc}$ and to 0 otherwise. We also let $\mathbf{Z}_{st} = (\mathbf{z}_{stc}, c = 1, \ldots, R_{st})$ denote an $R \times R_{st}$ matrix having as columns the indicator vectors $\mathbf{z}_{stc}$ corresponding to all response classes for units with the tth missingness pattern in the sth subpopulation. Finally, we let $\mathbf{Z}_s = (\mathbf{Z}_{st}, t = 1, \ldots, T_s)$ denote an $R \times (R + l_s)$ matrix combining, columnwise, the indicator vectors $\mathbf{z}_{stc}$ corresponding to all response classes for units with all missingness patterns in the sth subpopulation. Note that $\mathbf{Z}_{s1} = \mathbf{I}_R$ (an identity matrix of order R). The observed frequencies $\{n_{stc}\}$ indicate the units in the sth subpopulation with the tth missingness pattern classified into the cth response class. The vector $\mathbf{N}_{st} = (n_{stc}, c = 1, \ldots, R_{st})'$ stacks the observed frequencies of units with the tth missingness pattern in the sth subpopulation, and $\mathbf{N}_s = (\mathbf{N}'_{st}, t = 1, \ldots, T_s)'$ encloses all the observed frequencies corresponding to the sth subpopulation. Additionally, $\mathbf{N} = (\mathbf{N}'_s, s = 1, \ldots, S)'$ includes all the observed frequencies, and $n_{st+} = \sum_{c=1}^{R_{st}} n_{stc}$ indicates the total units with the tth missingness pattern in the sth subpopulation. Replacement of any subscript by "+" indicates the sum of the values over that particular subscript.

We assume that units randomly selected from the sth subpopulation and that should be classified in the rth response category are actually considered as pertaining to the tth missingness pattern (i.e., classification in the only set with this pattern that includes the rth category) with probability $\lambda_{t(rs)}$. The $\{\lambda_{t(rs)}\}$ are the conditional probabilities of missingness, and the marginal probabilities of categorization are represented by $\{\theta_{r(s)}\}$. Underlying this simplified notation is the assumption of no misclassification. We assume throughout that there are no missing

values in $\mathbf{X}$. We illustrate the notation with the two examples described previously.

Example 1. We index the subpopulation of Kingston–Harriman by $s = 1$, and that of Portman by $s = 2$. The index r is used to order the 9 response categories (with $r = 1$ corresponding to Maternal smoking = none, Child's wheezing status = no wheeze) lexicographically. We index the missingness pattern associated to child's wheezing status by $t = 2$ and the missingness pattern associated to maternal smoking by $t = 3$. For the complete classification pattern ($t = 1$), there are $R_{s1} = R = 9$ classes/categories, so that $\mathcal{P}_{s1} = \{\{1\}, \{2\}, \ldots, \{9\}\}$, $\mathbf{Z}_{s1} = \mathbf{I}_9$, $s = 1, 2$, $\mathbf{N}_{11} = (167, 17, 19, 10, 1, 3, 52, 10, 11)'$, $n_{11+} = 290$, $\mathbf{N}_{21} = (120, 22, 19, 8, 5, 1, 39, 12, 12)'$, and $n_{21+} = 238$. For either city, the missingness pattern $t = 2$ has $R_{s2} = 3$ classes, so that $\mathcal{P}_{s2} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}$,

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

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

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

$s = 1, 2$, $\mathbf{N}_{13} = (28, 10, 12)'$, $n_{13+} = 50$, $\mathbf{N}_{23} = (31, 8, 14)'$, $n_{23+} = 53$. Additionally, $l_s = R_{s2} + R_{s3} = 6$, $\mathbf{N}_s = (\mathbf{N}'_{s1}, \mathbf{N}'_{s2}, \mathbf{N}'_{s3})'$, $\mathbf{Z}_s = (\mathbf{Z}_{s1}, \mathbf{Z}_{s2}, \mathbf{Z}_{s3})$, $s = 1, 2$, $\mathbf{N}_{++} = (n_{1++}, n_{2++})' = (661, 477)'$, $n_{+++} = 1138$ and $\mathbf{N} = (\mathbf{N}'_1, \mathbf{N}'_2)'$.

Example 2. There are $S = 10$ subpopulations, defined from the combinations of the levels of gender and age, and $R = 8$ response categories, obtained from the three longitudinal binary responses. We index the subpopulations by s and the response categories by r lexicographically following the display in Table 2. As the missingness patterns are equal for all the subpopulations, we present the partitions $\mathcal{P}_{st}$ and matrices of response indicators $\mathbf{Z}_{st}$ for a general subpopulation, and we illustrate the vectors of frequencies $\mathbf{N}_{st}$ for $s = 1$ (boys, 5 to 7 years old in 1977). The index $t = 1$ corresponds to the $R = 8$ response classes/categories with complete classification of units, so that $\mathcal{P}_{s1} = \{\{r\}, r = 1, \ldots, 8\}$, $\mathbf{Z}_{s1} = \mathbf{I}_8$, and $\mathbf{N}_{11} = (90, 9, 3, 7, 0, 1, 1, 8)'$. When only the response in 1981, 1979 or 1977 is missing, the missingness patterns are indexed by $t = 2, 3, 4$, respectively, resulting in $R_{s2} = R_{s3} = R_{s4} = 4$ response classes, so

that

$$\begin{aligned}
&\mathcal{P}_{s2} = \big\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\big\},\\
&\mathbf{Z}_{s2} = \mathbf{I}_4 \otimes \mathbf{1}_2, \qquad \mathbf{N}_{12} = (16,5,0,0)',\\
&\mathcal{P}_{s3} = \big\{\{1,3\},\{2,4\},\{5,7\},\{6,8\}\big\},\\
&\mathbf{Z}_{s3} = \mathbf{I}_2 \otimes \mathbf{1}_2 \otimes \mathbf{I}_2, \qquad \mathbf{N}_{13} = (9,3,0,0)',\\
&\mathcal{P}_{s4} = \big\{\{1,5\},\{2,6\},\{3,7\},\{4,8\}\big\},\\
&\mathbf{Z}_{s4} = \mathbf{1}_2 \otimes \mathbf{I}_4, \qquad \mathbf{N}_{14} = (129,18,6,13)'.
\end{aligned}$$

The indices $t = 5, 6, 7$ correspond to the missingness patterns where only the response in 1977, 1979 or 1981 is observed, yielding $R_{s5} = R_{s6} = R_{s7} = 2$ response classes; thus

$$\begin{aligned}
&\mathcal{P}_{s5} = \{\{1,2,3,4\},\{5,6,7,8\}\}, \qquad \mathbf{Z}_{s5} = \mathbf{I}_2 \otimes \mathbf{1}_4, \qquad \mathbf{N}_{15} = (32,5)',\\
&\mathcal{P}_{s6} = \{\{1,2,5,6\},\{3,4,7,8\}\}, \qquad \mathbf{Z}_{s6} = \mathbf{1}_2 \otimes \mathbf{I}_2 \otimes \mathbf{1}_2, \qquad \mathbf{N}_{16} = (33,11)',\\
&\mathcal{P}_{s7} = \{\{1,3,5,7\},\{2,4,6,8\}\}, \qquad \mathbf{Z}_{s7} = \mathbf{1}_4 \otimes \mathbf{I}_2, \qquad \mathbf{N}_{17} = (70,24)'.
\end{aligned}$$

The conditions $\mathcal{C}_{stc} \cap \mathcal{C}_{std} = \varnothing$, $c \neq d$ and $\bigcup_{c=1}^{R_{st}} \mathcal{C}_{stc} = \{1,\ldots,R\}$ are guaranteed by letting $\mathbf{Z}_{st}$ have one element equal to 1 in exactly one column for each row. In both examples, the observed missingness patterns for each subpopulation are the same. More generally, R_{st}, $\mathcal{C}_{stc}$, $\mathcal{P}_{st}$, l_s, $\mathbf{z}_{stc}$, $\mathbf{Z}_{st}$ and $\mathbf{Z}_s$ need not be necessarily equal for $s = 1,\ldots,S$.

3 Probability model and missingness mechanisms

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

$$\begin{aligned}
P\big(\mathbf{N}|\boldsymbol{\theta},\{\lambda_{t(rs)}\},\mathbf{N}_{++}\big) = \prod_{s=1}^{S} \frac{n_{s++}!}{\prod_{t=1}^{T_s}\prod_{c=1}^{R_{st}} n_{stc}!} \prod_{r=1}^{R} (\theta_{r(s)}\lambda_{1(rs)})^{n_{s1r}}\\
\times \prod_{t=2}^{T_s}\prod_{c=1}^{R_{st}} \Big(\sum_{r\in\mathcal{C}_{stc}} \theta_{r(s)}\lambda_{t(rs)}\Big)^{n_{stc}},
\end{aligned} \tag{3.1}$$

where $\theta_{r(s)}$ is the marginal probability that a unit selected at random from the sth subpopulation is classified in the rth response category, $\lambda_{t(rs)}$ is the conditional probability that a unit randomly selected from the sth subpopulation and that should be classified in the rth response category is associated to the tth missingness pattern, $\boldsymbol{\theta} = (\boldsymbol{\theta}_s', s = 1,\ldots,S)'$, $\boldsymbol{\theta}_s = (\theta_{r(s)}, r = 1,\ldots,R)'$, $\sum_{r=1}^{R}\theta_{r(s)} = 1$, $s = 1,\ldots,S$ 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 and a conditional model for the missingness process given the measurements corresponds to the so-called selection model framework described in Little and Rubin (2002).

If it were possible to identify the response category associated to every unit in each of the missingness patterns, y_{str} would be the number of sampling units, selected from the sth subpopulation and with the tth missingness pattern, classified into the rth response category. Hence, $\{y_{str}\}$ denote nonobservable augmented frequencies, except for the missingness pattern $t = 1$ (no missing data), where $y_{s1r} = n_{s1r}$. Under the other patterns, we only know the frequencies associated to the response classes $\mathcal{C}_{stc}$, namely $n_{stc} = \sum_{r \in \mathcal{C}_{stc}} y_{str}$.

For each subpopulation, there are $R - 1$ parameters $\{\theta_{r(s)}\}$ and $R(T_s - 1)$ parameters $\{\lambda_{t(rs)}\}$ not functionally related to the former, totalizing $RT_s - 1$ parameters. Likewise, there are R observed frequencies for the complete classification pattern and other l_s ones for the patterns with some missingness. As n_{s++} is fixed, there is a total of $R - 1 + l_s$ not functionally related observed frequencies in each subpopulation. Therefore, the mismatch between the $R\sum_{s=1}^{S} T_s - S$ parameters $\{\theta_{r(s)}, \lambda_{t(rs)}\}$ associated to the augmented frequencies $\{y_{str}\}$ and the $S(R - 1) + \sum_{s=1}^{S} l_s$ observed frequencies $\{n_{stc}\}$ associated to the parameters $\{\sum_{r \in \mathcal{C}_{stc}} \theta_{r(s)}\lambda_{t(rs)}\}$ clearly point towards an overparameterization of (3.1) with $\sum_{s=1}^{S}[R(T_s - 1) - l_s]$ nonidentifiable parameters.

As the interest usually lies in $\{\theta_{r(s)}\}$, we consider reduced structures for $\{\lambda_{t(rs)}\}$ to identify the model. One of them corresponds to a noninformative missingness mechanism or, according to Rubin (1976), a missing at random (MAR) mechanism, expressed by

$$\begin{aligned} \text{MAR:}\quad & \lambda_{t(rs)} = \alpha_{t(cs)}, \\ & s = 1, \ldots, S, t = 1, \ldots, T_s, c = 1, \ldots, R_{st}, r \in \mathcal{C}_{stc}. \end{aligned} \tag{3.2}$$

This indicates 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 factored as

$$L(\boldsymbol{\theta}, \{\alpha_{t(cs)}\}|\mathbf{N}; \text{MAR}) \propto L_1(\boldsymbol{\theta}|\mathbf{N}) L_2(\{\alpha_{t(cs)}\}|\mathbf{N}; \text{MAR}), \tag{3.3}$$

where

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

and

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

A special case known as the missing completely at random (MCAR) mechanism, namely

$$\text{MCAR:}\quad \lambda_{t(rs)} = \alpha_{t(s)}, \qquad s = 1, \ldots, S, t = 1, \ldots, T_s, r = 1, \ldots, R, \tag{3.4}$$

implies that the conditional probabilities of missingness do not depend on the response categories, irrespectively of whether they are partially observed or not. There are, under this missingness mechanism, $S + \sum_{s=1}^{S}(l_s - T_s)$ degrees of freedom in the likelihood function

$$L\big(\boldsymbol{\theta}, \{\alpha_{t(s)}\}|\mathbf{N}; \text{MCAR}\big) \propto L_1\big(\boldsymbol{\theta}|\mathbf{N}\big) L_2\big(\{\alpha_{t(s)}\}|\{n_{st+}\}; \text{MCAR}\big), \tag{3.5}$$

where $L_1\big(\boldsymbol{\theta}|\mathbf{N}\big)$ has the same definition as (3.3) and

$$L_2\big(\{\alpha_{t(s)}\}|\{n_{st+}\}; \text{MCAR}\big) = \prod_{s=1}^{S}\prod_{t=1}^{T_s} \alpha_{t(s)}^{n_{st+}}. \tag{3.6}$$

This implies that inferences about $\boldsymbol{\theta}$ can be based only on the distribution of $\mathbf{N}$ conditionally on $\{n_{st+}\}$, the kernel of which, under (3.4), is L_1. Then, the MCAR missingness mechanism can be ignored for both likelihood- and frequentist-based inferences, as discussed by Paulino (1991) in the multinomial setting. The MAR missingness mechanism, on the other hand, is ignorable for likelihood-based but not for frequentist-based inferences on $\boldsymbol{\theta}$. Kenward and Molenberghs (1998) present a practical illustration where the estimation of the Fisher information becomes biased when the missingness process under the MAR mechanism is ignored.

Alternative models that allow the conditional probabilities of missingness not to vary for some or all subpopulations may be considered. Since under either the MCAR or the MAR mechanisms the likelihood function factors as in (3.3) and (3.5), such alternative models for $\{\lambda_{t(rs)}\}$ have no effect on the ML estimates of $\boldsymbol{\theta}$ and are not developed further.

Missing not at random (MNAR) or informative missingness mechanisms can be formulated by assuming that at least two conditional probabilities of missing response categories pertaining to the same class are not equal, that is, $\lambda_{t(as)} \neq \lambda_{t(bs)}$, $\{a, b\} \in \mathcal{C}_{stc}$. For such models, it is necessary to specify at least $\sum_{s=1}^{S}[R(T_s - 1) - l_s]$ parametric constraints to obtain an identifiable structure. In Example 1, for instance, we may assume that the conditional probabilities of missingness depend only on the home city and on the missing result. Substituting the index r by two indices, i to represent the maternal smoking level and j to indicate child wheezing status, and incorporating the constrains $\lambda_{2(ijs)} = \alpha_{2(js)}$ and $\lambda_{3(ijs)} = \alpha_{3(is)}$ in the likelihood function, we obtain a saturated statistical model under a MNAR structure. MNAR mechanisms are not ignorable for likelihood- or frequentist-based inferences on $\boldsymbol{\theta}$ since the likelihood function cannot be factored as in the MAR or the MCAR cases; here the ML estimators of $\boldsymbol{\theta}$ and $\{\lambda_{t(rs)}\}$ are not orthogonal and the term corresponding to the covariance matrix of $\boldsymbol{\theta}$ must be ex-

tracted from the joint covariance matrix. In Section 8, we address some additional details regarding these mechanisms.

4 Fitting saturated models for the marginal probabilities of categorization

Estimates of the probabilities of the response categories obtained under saturated models using all the available data may be used as input in the process of fitting unsaturated models as we show in Section 5.

As the units with missing responses in all variables, that is, with $\mathcal{P}_{st} = \{\mathcal{C}_{st1}\} = \{\{1, \ldots, R\}\}$, do not carry any information about $\boldsymbol{\theta}$ under either the MAR or the MCAR mechanisms, we ignore such missingness pattern and redefine T_s as the number of partial missingness patterns; in this context, we also let n_{s++} be the number of units with some type of categorization.

To avoid technical problems related to the natural restriction on the probabilities of categorization in each subpopulation, we let $\overline{\boldsymbol{\theta}}_s = (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})\boldsymbol{\theta}_s = (\theta_{r(s)}, r = 1, \ldots, R-1)'$ contain the $R-1$ first components of $\boldsymbol{\theta}_s$ and $\overline{\boldsymbol{\theta}} = [\mathbf{I}_S \otimes (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})]\boldsymbol{\theta} = (\overline{\boldsymbol{\theta}}_s', s = 1, \ldots, S)'$, where $\mathbf{0}_k$ denotes a $k \times 1$ vector with all elements equal to 0. We also let $\overline{\mathbf{Z}}_{st}$ denote an $(R-1) \times (R_{st}-1)$ matrix obtained from $\mathbf{Z}_{st}$ by deleting the last row and column and $\overline{\mathbf{Z}}_s = (\overline{\mathbf{Z}}_{st}, t = 1, \ldots, T_s)$. Then, $\overline{\boldsymbol{\theta}}_{st} = \overline{\mathbf{Z}}_{st}'\overline{\boldsymbol{\theta}}_s = (\theta_{c(st)}, c = 1, \ldots, R_{st}-1)'$ encloses the parameters $\{\theta_{r(s)}\}$ related to the first $R_{st}-1$ classes associated to the tth missingness pattern of the sth subpopulation, where $\theta_{c(st)} = \sum_{r \in \mathcal{C}_{stc}} \theta_{r(s)} = \mathbf{z}_{stc}'\boldsymbol{\theta}_s$. Similarly, we let $\mathbf{p}_{st} = \mathbf{N}_{st}/n_{st+} = (p_{c(st)}, c = 1, \ldots, R_{st})'$ be the observed proportions of units in the classes associated with the tth missingness pattern in the sth subpopulation and $\mathbf{p}_s = (\mathbf{p}_{st}', t = 1, \ldots, T_s)'$. Finally, we let $\overline{\mathbf{N}}_{st} = (\mathbf{I}_{R_{st}-1}, \mathbf{0}_{R_{st}-1})\mathbf{N}_{st} = (n_{stc}, c = 1, \ldots, R_{st}-1)'$, $\overline{\mathbf{p}}_{st} = \overline{\mathbf{N}}_{st}/n_{st+}$ and $\overline{\mathbf{p}}_s = (\overline{\mathbf{p}}_{st}', t = 1, \ldots, T_s)'$. We may obtain $\boldsymbol{\theta}_s$ from $\overline{\boldsymbol{\theta}}_s$ and $\boldsymbol{\theta}$ from $\overline{\boldsymbol{\theta}}$, respectively, from the relations

$$\boldsymbol{\theta}_s = \begin{pmatrix} \mathbf{0}_{R-1} \\ 1 \end{pmatrix} + \begin{pmatrix} \mathbf{I}_{R-1} \\ -\mathbf{1}_{R-1}' \end{pmatrix} \overline{\boldsymbol{\theta}}_s = \mathbf{b}_s + \mathbf{B}_s\overline{\boldsymbol{\theta}}_s, \tag{4.1}$$

$$\boldsymbol{\theta} = \mathbf{1}_S \otimes \begin{pmatrix} \mathbf{0}_{R-1} \\ 1 \end{pmatrix} + \left[\mathbf{I}_S \otimes \begin{pmatrix} \mathbf{I}_{R-1} \\ -\mathbf{1}_{R-1}' \end{pmatrix}\right] \overline{\boldsymbol{\theta}} = \mathbf{b} + \mathbf{B}\overline{\boldsymbol{\theta}}, \tag{4.2}$$

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

4.1 ML inferences under MAR and MCAR assumptions

Maximum likelihood estimation of $\boldsymbol{\theta}$ can be based only on the factor $L_1(\boldsymbol{\theta}|\mathbf{N})$ in (3.3) and in general, must be carried out through iterative methods. Among the available alternatives, the EM algorithm has the advantage of not requiring

derivatives of the log-likelihood function. For both MCAR and MAR mechanisms, the EM algorithm is specified by

$$\widehat{\boldsymbol{\theta}}_s^{(i+1)} = \frac{1}{n_{s++}}\left(\mathbf{N}_{s1} + \sum_{t=2}^{T_s} \mathbf{D}_{\widehat{\boldsymbol{\theta}}_s^{(i)}} \mathbf{Z}_{st} \mathbf{D}^{-1}_{\mathbf{Z}'_{st}\widehat{\boldsymbol{\theta}}_s^{(i)}} \mathbf{N}_{st}\right), \quad s = 1, \ldots, S, i = 0, 1, \ldots, \tag{4.3}$$

where $\mathbf{D}_{\widehat{\boldsymbol{\theta}}_s^{(i)}}$ denotes a diagonal matrix with the elements of $\widehat{\boldsymbol{\theta}}_s^{(i)}$ along the main diagonal. We may start the iterative process by letting $\widehat{\boldsymbol{\theta}}_s^{(0)} = \mathbf{p}_{s1} = \mathbf{N}_{s1}/n_{s1+}$. It is important to replace any null frequencies by a small value, for example, $(Rn_{s1+})^{-1}$ or 10^{-6}, since null values for $\widehat{\theta}_{r(s)}^{(0)}$ do not allow information from other missingness patterns to be incorporated. Alternatively, one may consider Newton–Raphson or Fisher's scoring algorithms. The $[S(R-1)\times 1]$-dimensional score function associated to $\ln L_1(\boldsymbol{\theta}|\mathbf{N}_s)$ is $\mathbf{S}(\overline{\boldsymbol{\theta}}) = (\mathbf{S}'_s, s = 1, \ldots, S)'$, where

$$\mathbf{S}_s(\overline{\boldsymbol{\theta}}_s) = \sum_{t=1}^{T_s} \overline{\mathbf{Z}}_{st}\left[\boldsymbol{\Sigma}(\overline{\boldsymbol{\theta}}_{st})\right]^{-1}(\overline{\mathbf{p}}_{st} - \overline{\boldsymbol{\theta}}_{st}), \qquad s = 1, \ldots, S, \tag{4.4}$$

and $\boldsymbol{\Sigma}(\overline{\boldsymbol{\theta}}_{st}) = \frac{1}{n_{st+}}(\mathbf{D}_{\overline{\boldsymbol{\theta}}_{st}} - \overline{\boldsymbol{\theta}}_{st}\overline{\boldsymbol{\theta}}'_{st})$. The corresponding $S(R-1)\times S(R-1)$ hessian matrix, $\mathbf{H}(\overline{\boldsymbol{\theta}})$, is a block diagonal matrix with blocks

$$\mathbf{H}_s(\overline{\boldsymbol{\theta}}_s) = -\sum_{t=1}^{T_s} \overline{\mathbf{Z}}_{st}\left[\mathbf{D}_{\overline{\mathbf{N}}_{st}}\mathbf{D}^{-2}_{\overline{\boldsymbol{\theta}}_{st}} + \frac{n_{st}R_{st}}{(1 - \mathbf{1}'_{R_{st}-1}\overline{\boldsymbol{\theta}}_{st})^2}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right]\overline{\mathbf{Z}}'_{st}, \tag{4.5}$$

$s = 1, \ldots, S$, where $\mathbf{D}^{-2}_{\overline{\boldsymbol{\theta}}_{st}} = \mathbf{D}^{-1}_{\overline{\boldsymbol{\theta}}_{st}}\mathbf{D}^{-1}_{\overline{\boldsymbol{\theta}}_{st}}$.

For both MCAR and MAR mechanisms, the Newton–Raphson algorithm is then specified by

$$\widehat{\overline{\boldsymbol{\theta}}}_s^{(i+1)} = \widehat{\overline{\boldsymbol{\theta}}}_s^{(i)} + \left[-\mathbf{H}_s\left(\widehat{\overline{\boldsymbol{\theta}}}_s^{(i)}\right)\right]^{-1}\mathbf{S}_s\left(\widehat{\overline{\boldsymbol{\theta}}}_s^{(i)}\right), \qquad s = 1, \ldots, S, i = 0, 1, \ldots. \tag{4.6}$$

Fisher's scoring algorithm requires additional estimation of the conditional probabilities of missingness, since

$$E\left(n_{stc}|\mathbf{N}_{++}, \boldsymbol{\theta}, \{\boldsymbol{\alpha}_{st}^{\mathrm{MAR}}\}\right) = n_{s++}\mathbf{z}'_{stc}\boldsymbol{\theta}_s\alpha_{t(cs)}, \tag{4.7}$$

$$E\left(n_{stc}|\mathbf{N}_{++}, \boldsymbol{\theta}, \{\boldsymbol{\alpha}_{st}^{\mathrm{MCAR}}\}\right) = n_{s++}\mathbf{z}'_{stc}\boldsymbol{\theta}_s\alpha_{t(s)}, \tag{4.8}$$

where $\boldsymbol{\alpha}_{st}^{\mathrm{MAR}} = (\alpha_{t(cs)}, c = 1, \ldots, R_{st})'$ and $\boldsymbol{\alpha}_{st}^{\mathrm{MCAR}} = \alpha_{t(s)}$. For the MAR mechanism, estimators of these additional parameters may be obtained from the ML estimator $\{\widehat{\boldsymbol{\theta}}_s\}$ in view of the invariance property and are given by

$$\widehat{\boldsymbol{\alpha}}_{st}^{\mathrm{MAR}} = \frac{1}{n_{s++}}\mathbf{D}^{-1}_{\mathbf{Z}'_{st}\widehat{\boldsymbol{\theta}}_s}\mathbf{N}_{st}, \qquad s = 1, \ldots, S, t = 1, \ldots, T_s. \tag{4.9}$$

Under the MCAR mechanism, on the other hand, the factor (3.6) leads directly to the ML estimators of the conditional probabilities of missingness that are given by

$$\widehat{\boldsymbol{\alpha}}_{st}^{\text{MCAR}} = \widehat{\alpha}_{t(s)} = \frac{n_{st+}}{n_{s++}}, \qquad s = 1, \ldots, S, t = 1, \ldots, T_s. \tag{4.10}$$

The Fisher information matrix $\mathcal{I}(\overline{\boldsymbol{\theta}}, \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\})$ corresponding to $\overline{\boldsymbol{\theta}}$ under the mechanism $\mathcal{M}$ (= MAR or MCAR) may be obtained from the above results and is detailed in Appendix A.1. Estimators of the asymptotic covariance matrix $\widehat{\mathbf{V}}_{\widehat{\overline{\boldsymbol{\theta}}}}^{\mathcal{M}}$ of $\widehat{\overline{\boldsymbol{\theta}}}$ may be obtained either as the inverse of the observed information matrix, $-\mathbf{H}(\overline{\boldsymbol{\theta}})$, computed at $\widehat{\overline{\boldsymbol{\theta}}}$, or as the inverse of the Fisher information matrix computed at $(\widehat{\overline{\boldsymbol{\theta}}}, \{\widehat{\boldsymbol{\alpha}}_{st}^{\mathcal{M}}\})$. From (4.2), we may estimate the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ as $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{\mathcal{M}} = \mathbf{B}\widehat{\mathbf{V}}_{\widehat{\overline{\boldsymbol{\theta}}}}^{\mathcal{M}}\mathbf{B}'$.

Substituting $\{\widehat{\alpha}_{t(cs)} = n_{stc}/(n_{s++}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s)\}$ from (4.9) in the expression for the Fisher information matrix corresponding to the MAR mechanism, we obtain $\mathcal{I}(\widehat{\overline{\boldsymbol{\theta}}}, \{\widehat{\boldsymbol{\alpha}}_{st}^{\text{MAR}}\}) = -\mathbf{H}(\widehat{\overline{\boldsymbol{\theta}}})$ so that essentially three strategies may be employed to obtain the ML estimate of $\boldsymbol{\theta}$. The first relies exclusively on (4.3) for both MAR or MCAR mechanisms; the second relies on (4.6) for both cases too, and the third relies on (4.6) with the observed information matrix replaced by the Fisher information matrix under the MCAR mechanism. We may use this iterative process even when assuming the MAR mechanism, if after obtaining $\widehat{\boldsymbol{\theta}}$ we consider an estimate of its asymptotic covariance matrix under the MAR mechanism. In fact, this may be the best choice to avoid the low speed of the EM algorithm and, at the same time, the possible instability of the iterative process based on the MAR mechanism where zero counts may easily generate unstable covariance matrices.

Goodness-of-fit tests for the MCAR mechanism, conditionally on the MAR assumption, can be obtained either from Wilks' likelihood ratio statistic

$$\begin{aligned} Q_L(\text{MCAR}|\text{MAR}) &= -2\ln\frac{L_2(\{\widehat{\alpha}_{t(s)}\}|\{n_{st+}\}; \text{MCAR})}{L_2(\{\widehat{\alpha}_{t(cs)}\}|\mathbf{N}; \text{MAR})} \\ &= -2\sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} n_{stc}\left[\ln(\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s) - \ln\left(\frac{n_{stc}}{n_{st+}}\right)\right] \\ &= -2\sum_{s=1}^{S}\mathbf{N}'_s[\mathbf{ln}(\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s) - \mathbf{ln}(\mathbf{p}_s)], \end{aligned} \tag{4.11}$$

where $\mathbf{ln}(\mathbf{p}_s)$ denotes a vector containing the logarithms of the elements of $\mathbf{p}_s$, or from the Pearson (Q_P) and the Neyman (Q_N) statistics, which are given in Appendix B.1. Under the MCAR hypothesis, all three statistics follow an asymptotic $\chi^2_{(g)}$ distribution, with $g = S + \sum_{s=1}^{S}(l_s - T_s)$ degrees of freedom. Since null observed frequencies n_{stc} do not contribute to the probability mass function (3.1),

we use the definition $0 \times [\ln(\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s) - \ln(0/n_{st+})] \equiv 0$ in (4.11) to avoid inconsistencies in the computation of the logarithm. The Neyman statistic [(B.2), in Appendix B.1] requires $\{n_{stc} > 0\}$ 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 computing the inverse of $\mathbf{D}_{\mathbf{p}_s}$. In the WLS context, Koch et al. (1972) suggest replacing $n_{stc} = 0$ by $(R_{st}n_{st+})^{-1}$.

The expected augmented frequencies under the MAR and the MCAR mechanisms can be estimated by $\{\widehat{y}^{\text{MAR}}_{str} = n_{s++}\widehat{\theta}_{r(s)}\widehat{\alpha}_{t(cs)}\}$ and $\{\widehat{y}^{\text{MCAR}}_{str} = n_{s++}\widehat{\theta}_{r(s)}\widehat{\alpha}_{t(s)}\}$, respectively.

4.2 WLS inferences under MCAR assumption

The 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 with parameters $\{\overline{\boldsymbol{\theta}}^0_{st}\}$ for each subpopulation. The MCAR assumption implies the linear structure $(\overline{\boldsymbol{\theta}}^{0'}_{st}, t = 1, \ldots, T_s)' = \overline{\mathbf{Z}}'_s\overline{\boldsymbol{\theta}}_s$, $s = 1, \ldots, S$, so that the WLS methodology proposed by Grizzle et al. (1969) may be considered for analysis. Here the response categories vary from one missingness pattern to the other, as pointed out by Koch et al. (1972).

The WLS approach involves the minimization of the quadratic form

$$Q_N(\{\overline{\boldsymbol{\theta}}_s\}) = \sum_{s=1}^{S}(\overline{\mathbf{p}}_s - \overline{\mathbf{Z}}'_s\overline{\boldsymbol{\theta}}_s)'\big[\boldsymbol{\Sigma}_*(\overline{\mathbf{p}}_s)\big]^{-1}(\overline{\mathbf{p}}_s - \overline{\mathbf{Z}}'_s\overline{\boldsymbol{\theta}}_s),$$

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

$$\widetilde{\overline{\boldsymbol{\theta}}}_s = \big(\overline{\mathbf{Z}}_s\big[\boldsymbol{\Sigma}_*(\overline{\mathbf{p}}_s)\big]^{-1}\overline{\mathbf{Z}}'_s\big)^{-1}\overline{\mathbf{Z}}_s\big[\boldsymbol{\Sigma}_*(\overline{\mathbf{p}}_s)\big]^{-1}\overline{\mathbf{p}}_s, \tag{4.12}$$

and an estimate of its asymptotic covariance matrix is

$$\widetilde{\mathbf{V}}_{\widetilde{\overline{\boldsymbol{\theta}}}_s} = \big(\overline{\mathbf{Z}}_s\big[\boldsymbol{\Sigma}_*(\overline{\mathbf{p}}_s)\big]^{-1}\overline{\mathbf{Z}}'_s\big)^{-1}.$$

From (4.1), we obtain the WLS estimator of $\boldsymbol{\theta}_s$ as $\widetilde{\boldsymbol{\theta}}_s = \mathbf{b}_s + \mathbf{B}_s\widetilde{\overline{\boldsymbol{\theta}}}_s$; analogously, an estimate of the corresponding asymptotic covariance matrix is $\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}_s} = \mathbf{B}_s\widetilde{\mathbf{V}}_{\widetilde{\overline{\boldsymbol{\theta}}}_s}\mathbf{B}'_s$. An estimate of the asymptotic covariance matrix of $\widetilde{\overline{\boldsymbol{\theta}}}$, denoted by $\widetilde{\mathbf{V}}_{\widetilde{\overline{\boldsymbol{\theta}}}}$, is a block diagonal matrix with blocks $\widetilde{\mathbf{V}}_{\widetilde{\overline{\boldsymbol{\theta}}}_s}$, $s = 1, \ldots, S$; similarly, for $\widetilde{\boldsymbol{\theta}}$, we have $\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}} = \mathbf{B}\widetilde{\mathbf{V}}_{\widetilde{\overline{\boldsymbol{\theta}}}}\mathbf{B}'$.

Goodness-of-fit for the MCAR mechanism may be tested with the Neyman statistic, $Q_N(\{\overline{\boldsymbol{\theta}}_s\})$ computed at $\{\widetilde{\overline{\boldsymbol{\theta}}}_s\}$, which asymptotically follows a $\chi^2_{(g)}$ distribution under the null hypothesis. Since we assume that $\boldsymbol{\Sigma}_*(\overline{\mathbf{p}}_s)$ is nonsingular and in practice we do not always have $\{p_{c(st)} > 0\}$ or, equivalently, $\{n_{stc} > 0\}$, we replace such values with small quantities as suggested earlier. Estimates of the expected augmented frequencies are given by $\{\widetilde{y}_{str} = n_{st+}\widetilde{\theta}_{r(s)}\}$.

5 Fitting unsaturated models for the marginal probabilities of categorization

In general, questions of interest are related to a reduction of the number of parameters obtained by considering models based on functions of the marginal probabilities of categorization. In this context, we examine (strictly) linear and log-linear models under the ML approach and more general functional linear models under a hybrid ML/WLS approach.

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

We first focus on (strictly) linear models expressed as

$$M_L : \mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}, \tag{5.1}$$

where $\mathbf{A}$ is a $u \times SR$ matrix defining the u linear functions of interest with rank $r(\mathbf{A}) = u \leq S(R-1)$, $\mathbf{X}$ is a $u \times p$ model specification matrix 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 also be expressed under the equivalent constraint formulation $\mathbf{UA}\boldsymbol{\theta} = \mathbf{0}_{u-p}$, where $\mathbf{U}$ is a $(u-p) \times u$ full rank matrix such that $\mathbf{UX} = \mathbf{0}_{(u-p),p}$. We also have to include in the model specification the S natural constraints

$$(\mathbf{I}_S \otimes \mathbf{1}'_R)\boldsymbol{\theta} = \mathbf{1}_S. \tag{5.2}$$

We assume that $r(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R) = u + S$.

To take advantage of the expressions of the models in terms of $\overline{\boldsymbol{\theta}}$ considered in Section 4.1, it is convenient to consider (5.1) and (5.2) simultaneously, by setting

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \end{pmatrix}. \tag{5.3}$$

Therefore, if $u = S(R-1)$, we can obtain $\overline{\boldsymbol{\theta}}$ solely from $\mathbf{A}$, $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}) = \left[\mathbf{I}_S \otimes (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})\right] \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \end{pmatrix}. \tag{5.4}$$

Incorporating this linear structure in $\ln L_1(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{N})$ and differentiating the resulting expression with respect to $\boldsymbol{\beta}$, we obtain the score vector $\mathbf{S}_L(\boldsymbol{\beta}) = \mathbf{W}'\mathbf{S}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}))$, where

$$\mathbf{W} = \left[\mathbf{I}_S \otimes (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})\right] \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X} \\ \mathbf{0}_{S,p} \end{pmatrix}, \tag{5.5}$$

and $\mathbf{S}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}))$ is defined in (4.4). We also obtain the hessian matrix $\mathbf{H}_L(\boldsymbol{\beta}) = \mathbf{W}'\mathbf{H}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}))\mathbf{W}$, where $\mathbf{H}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}))$ is given in (4.5), and the Fisher information matrix under the $\mathcal{M}$ (= MAR or MCAR) mechanism, namely $\mathcal{I}_L(\boldsymbol{\beta}, \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\}) =$

$\mathbf{W}'\mathcal{I}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}), \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\})\mathbf{W}$, where $\mathcal{I}(\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}), \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\})$ is defined in Appendix A.1. The iterative process for the Newton–Raphson or Fisher' scoring algorithms may be initialized with the WLS estimate

$$\widehat{\boldsymbol{\beta}}^{(0)} = [\mathbf{X}'(\mathbf{A}\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{\mathcal{M}}\mathbf{A}')^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{A}\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{\mathcal{M}}\mathbf{A}')^{-1}\mathbf{A}\widehat{\boldsymbol{\theta}}; \tag{5.6}$$

where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model, and $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{\mathcal{M}}$ is an estimate of its corresponding asymptotic covariance matrix under the mechanism $\mathcal{M}$, obtained according to the suggestion presented in Section 4.1. We might as well use the alternative iterative scheme proposed by Paulino and Silva (1999) adapted to incomplete data, by using the constraint formulation expressed in terms of (4.2). When $u < S(R-1)$, we need to augment the model (5.3) and, consequently, perform suitable changes in (5.4)–(5.6) as detailed in Appendix C.1.

Estimators of the asymptotic covariance matrix $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}L}^{\mathcal{M}}$ of $\widehat{\boldsymbol{\beta}}$ under $\mathcal{M}$ may be obtained along similar lines as in Section 4.1. We obtain the ML estimate $\widehat{\overline{\boldsymbol{\theta}}}(M_L)$ of $\overline{\boldsymbol{\theta}}$ under the linear model M_L computing (5.4) at $\widehat{\boldsymbol{\beta}}$ and, given its linear structure, we derive an estimate of its corresponding asymptotic covariance matrix $\widehat{\mathbf{V}}_{\widehat{\overline{\boldsymbol{\theta}}}(M_L)}^{\mathcal{M}} = \mathbf{W}\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}L}^{\mathcal{M}}\mathbf{W}'$. Analogously, ML estimates of the linear functions $\mathbf{A}\boldsymbol{\theta}$ under M_L are $\mathbf{X}\widehat{\boldsymbol{\beta}}$ and an estimate of its asymptotic covariance matrix is $\widehat{\mathbf{V}}_{\mathbf{A}\widehat{\boldsymbol{\theta}}(M_L)}^{\mathcal{M}} = \mathbf{X}\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}L}^{\mathcal{M}}\mathbf{X}'$.

Now, we consider log-linear models

$$M_{LL} : \mathbf{ln}(\boldsymbol{\theta}) = (\mathbf{I}_S \otimes \mathbf{1}_R)\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\beta}, \tag{5.7}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_S)' = -\mathbf{ln}[(\mathbf{I}_S \otimes \mathbf{1}_R')\,\mathbf{exp}(\mathbf{X}\boldsymbol{\beta})]$, $\mathbf{exp}(\mathbf{X}\boldsymbol{\beta})$ is a vector, the elements of which are exponentials of those of $\mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ 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 an $SR \times p$ matrix with each $R \times p$ submatrix $\mathbf{X}_s$ such that $r(\mathbf{1}_R, \mathbf{X}_s) = 1 + r(\mathbf{X}_s)$, $s = 1, \ldots, S$, and $r(\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}) = S + p$. Rewriting (5.7), we may obtain $\boldsymbol{\theta}$ from $\boldsymbol{\beta}$ by

$$\boldsymbol{\theta}(\boldsymbol{\beta}) = \mathbf{D}_{\boldsymbol{\psi}}^{-1}\,\mathbf{exp}(\mathbf{X}\boldsymbol{\beta}), \tag{5.8}$$

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

We can also consider a larger class of (generalized) log-linear models, expressed by

$$M_{LL} : \mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}, \tag{5.9}$$

where $\mathbf{A}$ is a $u \times SR$ matrix with rank $r(\mathbf{A}) = u \leq S(R-1)$ such that $\mathbf{A}(\mathbf{I}_S \otimes \mathbf{1}_R) = \mathbf{0}_{u,S}$. Taking $\mathbf{A} = \mathbf{I}_S \otimes (\mathbf{I}_{R-1}, -\mathbf{1}_{R-1})$, for instance, generates logits with the baseline category R. When $u = S(R-1)$, the $S(R-1) \times p$ matrix $\mathbf{X}_L$ is related to $\mathbf{X}$ via $\mathbf{X}_L = \mathbf{A}\mathbf{X}$ and $\mathbf{X} = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{X}_L$. When $u < S(R-1)$, we need to augment the model (5.9) before re-expressing it in the form of (5.7) as detailed in Appendix C.2.

The freedom equation formulations (5.7) and (5.9) are respectively equivalent to the constraint formulations $\mathbf{U}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{S(R-1)-p}$ and $\mathbf{U}_L\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}$, where $\mathbf{U}$ is full rank $[S(R-1)-p] \times SR$ matrix defining the $S(R-1)-p$ constraints such that $\mathbf{U}(\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}) = \mathbf{0}_{(SR-p),p}$, and $\mathbf{U}_L$ is full rank $(u-p) \times u$ matrix defining the $u-p$ constraints such that $\mathbf{U}_L\mathbf{X}_L = \mathbf{0}_{(u-p),p}$.

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

$$\mathbf{S}_{LL}(\boldsymbol{\beta}) = \sum_{s=1}^{S} \mathbf{X}_s' \left[\mathbf{N}_{s1} + \sum_{t=2}^{T_s} \left(\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})} \mathbf{Z}_{st} \mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})} \mathbf{N}_{st} \right) - n_{s++}\boldsymbol{\theta}_s(\boldsymbol{\beta}) \right]. \tag{5.10}$$

Further differentiation with respect to $\boldsymbol{\beta}'$ leads to the hessian matrix

$$\begin{aligned} \mathbf{H}_{LL}(\boldsymbol{\beta}) = \sum_{s=1}^{S} \mathbf{X}_s' \left[-n_{s++}\mathbf{I}_R + \sum_{t=2}^{T_s} \left(\mathbf{D}_{\mathbf{u}_{st}^I} - \mathbf{D}_{\mathbf{u}_{st}^{II}} \mathbf{Z}_{st} \mathbf{Z}_{st}' \right) \right] \\ \times \left\{ \mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})} - \boldsymbol{\theta}_s(\boldsymbol{\beta}) \left[\boldsymbol{\theta}_s(\boldsymbol{\beta}) \right]' \right\} \mathbf{X}_s, \end{aligned} \tag{5.11}$$

where $\mathbf{u}_{st}^I = \mathbf{Z}_{st}\mathbf{D}^{-1}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{N}_{st}$ and $\mathbf{u}_{st}^{II} = \mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{Z}_{st}\mathbf{D}^{-2}_{\mathbf{Z}_{st}'\boldsymbol{\theta}_s(\boldsymbol{\beta})}\mathbf{N}_{st}$. Under the $\mathcal{M}$ (= MAR or MCAR) mechanism, the Fisher information matrix $\mathcal{I}_{LL}(\boldsymbol{\beta}, \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\})$ is given in Appendix A.2.

These expressions allow us to get ML estimates $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ via either Newton–Raphson or Fisher's scoring algorithms. When $u = S(R-1)$, the iterative processes may be initialized with the WLS estimate

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(0)} = \left[\mathbf{X}_L' \left(\mathbf{A}\mathbf{D}^{-1}_{\widehat{\boldsymbol{\theta}}} \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}} \mathbf{D}^{-1}_{\widehat{\boldsymbol{\theta}}} \mathbf{A}' \right)^{-1} \mathbf{X}_L \right]^{-1} \\ \times \mathbf{X}_L' \left(\mathbf{A}\mathbf{D}^{-1}_{\widehat{\boldsymbol{\theta}}} \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}} \mathbf{D}^{-1}_{\widehat{\boldsymbol{\theta}}} \mathbf{A}' \right)^{-1} \mathbf{A}\,\mathbf{ln}(\widehat{\boldsymbol{\theta}}), \end{aligned} \tag{5.12}$$

where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model and $\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\theta}}}$ is an estimate of its corresponding asymptotic covariance matrix under the mechanism $\mathcal{M}$, obtained according to the suggestion presented in Section 4.1. The case $u < S(R-1)$ is detailed in Appendix C.2.

Estimators of the asymptotic covariance matrix $\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\beta}}LL}$ of the ML estimate $\widehat{\boldsymbol{\beta}}$ may be obtained along the lines suggested earlier. The ML estimate $\widehat{\boldsymbol{\theta}}(M_{LL})$ of $\boldsymbol{\theta}$ under M_{LL} may be obtained from (5.8); an estimate of its asymptotic covariance matrix under the $\mathcal{M}$ mechanism obtained via the *delta* method is

$$\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\theta}}(M_{LL})} = \widehat{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'}} \widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\beta}}LL} \left(\widehat{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'}} \right)' = \widehat{\mathbf{V}}_{LL}\mathbf{X}\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\beta}}LL}\mathbf{X}'\widehat{\mathbf{V}}_{LL}, \tag{5.13}$$

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

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 statistic for the joint model $(M, \mathcal{M})$ conditional on the assumed MAR mechanism can be partitioned as

$$\begin{aligned} Q_L(M, \mathcal{M}|\text{MAR}) &= -2\ln \frac{L_1(\widehat{\boldsymbol{\theta}}(M)|\mathbf{N}; M)L_2(\{\widehat{\alpha}_{t(cs)}(\mathcal{M})\}|\mathbf{N}; \mathcal{M})}{L_1(\widehat{\boldsymbol{\theta}}|\mathbf{N})L_2(\{\widehat{\alpha}_{t(cs)}\}|\mathbf{N}; \text{MAR})} \\ &= Q_L(M) + Q_L(\mathcal{M}|\text{MAR}), \end{aligned} \tag{5.14}$$

where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ under the saturated model, and $\widehat{\boldsymbol{\theta}}(M)$, under the model M, $\{\widehat{\alpha}_{t(cs)}\}$ are the ML estimates of $\{\alpha_{t(cs)}\}$ under the MAR mechanism and $\{\widehat{\alpha}_{t(cs)}(\mathcal{M})\}$, 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 reduced model for $\boldsymbol{\theta}$. If the parameter of interest is $\boldsymbol{\theta}$, the likelihood ratio statistic for the goodness-of-fit test of model M is

$$\begin{aligned} Q_L(M|\mathcal{M}) &= -2\ln \frac{L_1(\widehat{\boldsymbol{\theta}}(M)|\mathbf{N})}{L_1(\widehat{\boldsymbol{\theta}}|\mathbf{N})} \\ &= -2\sum_{s=1}^{S} \mathbf{N}'_s\big[\mathbf{ln}\big(\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)\big) - \mathbf{ln}\big(\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s\big)\big], \end{aligned} \tag{5.15}$$

and is independent of the more restrictive mechanism $\mathcal{M}$ than the assumed MAR mechanism. In contrast with the likelihood ratio statistic, the Pearson, Neyman and Wald statistics no longer have the same advantageous property of being independent of the more restrictive mechanism $\mathcal{M}$. Computational formulae for these goodness-of-fit statistics as well as their asymptotic null distributions are given in Appendix B.2.

For tests of hypotheses of the form $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{C}_0$, where $\mathbf{C}$ is a $c \times p$ full rank matrix ($c \leq p$) and $\mathbf{C}_0$ is a $c \times 1$ vector with known elements (usually, $\mathbf{C}_0 = \mathbf{0}_c$), we may appeal to the Wald statistic

$$Q_W(H|M, \mathcal{M}) = \big(\mathbf{C}\widehat{\boldsymbol{\beta}}(M) - \mathbf{C}_0\big)'\big(\mathbf{C}\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\beta}}(M)}\mathbf{C}'\big)^{-1}\big(\mathbf{C}\widehat{\boldsymbol{\beta}}(M) - \mathbf{C}_0\big), \tag{5.16}$$

which follows an asymptotic $\chi^2_{(c)}$ distribution under the null hypothesis.

5.2 WLS and hybrid ML/WLS inferences on functional linear models under MAR, MCAR and MNAR mechanisms

With the purpose of fitting functional linear models of $\boldsymbol{\theta}$ under an MCAR mechanism, Koch et al. (1972) considered a two-stage procedure according to which WLS methodology is used to fit such models to WLS estimates $\widetilde{\boldsymbol{\theta}}$ obtained in a first stage.

In the light of the functional asymptotic regression for complete data suggested by Imrey et al. (1981, 1982), it is also possible to apply WLS methods to fit functional linear models to ML estimates $\widehat{\boldsymbol{\theta}}$, or to some other best asymptotically normal (BAN) estimates, obtained under any missingness mechanism in a first stage, as suggested by Paulino (1991) in a multinomial setup. Using this hybrid methodology, we may draw inferences about $\boldsymbol{\theta}$ more easily, mainly in the context of nonignorable models for the missingness mechanism under a product-multinomial framework.

We consider functional linear models of the form

$$M_F : \mathbf{F} \equiv \mathbf{F}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}, \tag{5.17}$$

where $\mathbf{F}(\boldsymbol{\theta}) = (F_i(\boldsymbol{\theta}), i = 1, \ldots, u)'$ is a $u \times 1$ vector defining the $u \leq S(R-1)$ functions of interest, and 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 a $u \times p$ model specification matrix with $r(\mathbf{X}) = p \leq u$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector with unknown parameters. The equivalent constraint formulation of (5.17) is $\mathbf{U}\mathbf{F}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}$, where $\mathbf{U}$ is a $(u-p) \times u$ full rank matrix such that $\mathbf{U}\mathbf{X} = \mathbf{0}_{(u-p),p}$.

Let $\widetilde{\boldsymbol{\theta}}$ denote any BAN estimator of $\boldsymbol{\theta}$ reflecting all the available data, for example, 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 obtained under any MNAR mechanism. Similarly, let $\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}}$ represent a consistent estimator of the covariance matrix of $\widetilde{\boldsymbol{\theta}}$ under the same missingness mechanism. Given that for sufficiently large sample sizes, $\widetilde{\boldsymbol{\theta}} \overset{a}{\sim} N_u(\boldsymbol{\theta}, \widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}})$, we have $\widetilde{\mathbf{F}} \equiv \mathbf{F}(\widetilde{\boldsymbol{\theta}}) \overset{a}{\sim} N_u(\mathbf{F}, \widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}})$, where we assume that $\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}} = \widetilde{\mathbf{G}}\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}}\widetilde{\mathbf{G}}'$, with $\widetilde{\mathbf{G}} \equiv \mathbf{G}(\widetilde{\boldsymbol{\theta}})$, is nonsingular. Therefore, the WLS estimator of $\boldsymbol{\beta}$ in (5.17) is

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}}^{-1}\widetilde{\mathbf{F}}, \tag{5.18}$$

and an estimate of its asymptotic covariance matrix is $\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\beta}}} = (\mathbf{X}'\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}}^{-1}\mathbf{X})^{-1}$. The WLS estimator of the functions $\mathbf{F}$ under M_F is $\mathbf{X}\widetilde{\boldsymbol{\beta}}$ and, recalling the *delta* method, an estimate of its asymptotic covariance matrix is $\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}(M_F)} = \mathbf{X}\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\beta}}}\mathbf{X}'$.

The goodness-of-fit of the model M_F conditionally on the missingness mechanism $\mathcal{M}$ (MCAR, MAR or MNAR) can be tested with the Wald statistic

$$Q_W(M_F|\mathcal{M}) = (\mathbf{U}\widetilde{\mathbf{F}})'\big(\mathbf{U}\widetilde{\mathbf{V}}_{\widetilde{\mathbf{F}}}\mathbf{U}'\big)^{-1}\mathbf{U}\widetilde{\mathbf{F}}, \tag{5.19}$$

which follows an asymptotic null distribution $\chi^2_{(u-p)}$. Reductions in the dimension of $\boldsymbol{\beta}$ may also be assessed via Wald tests analogous to (5.16).

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

functions and associated first derivatives are $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}\ln(\boldsymbol{\theta})$ and $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{D}_{\boldsymbol{\theta}}^{-1}$ or $\mathbf{F}(\boldsymbol{\theta}) = \exp[\mathbf{A}\ln(\boldsymbol{\theta})]$ and $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\exp[\mathbf{A}\ln(\boldsymbol{\theta})]}\mathbf{A}\mathbf{D}_{\boldsymbol{\theta}}^{-1}$. Note that these last two matrices $\mathbf{G}(\boldsymbol{\theta})$ may be obtained using the chain rule for differentiation.

When (5.17) corresponds to strictly linear or log-linear models, we may use the results of Section 5.1 to obtain the first stage ingredients.

6 Illustration

We now apply the methods presented in the last two sections to the examples described in Section 1. For brevity, we carry out both analyses under the MAR assumption, after assessing whether the MCAR missingness mechanism may be a reasonable simplification; one MNAR mechanism is considered in Example 1 for illustration purposes. In Section 8, we indicate how additional sensitivity analyses of MNAR models may be conducted and comment on some problems associated to them.

Example 1. The association between maternal smoking and child wheezing status given the home city may be assessed through the logarithms of 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 $(\pi_{11(s)}, \pi_{12(s)}, \ldots, \pi_{33(s)})' = (\theta_{r(s)}, r = 1, \ldots, 9)' = \boldsymbol{\theta}_s$. Taking the order of the categories into account, to assess whether the conditional independence is tenable, as in Lipsitz and Fitzmaurice (1996), we first consider a homogeneous linear-by-linear association model with unit-spaced response scores (Agresti, 2002), that is, we let $\omega_{ij(s)} = \beta$, and then test whether $\beta = 0$. For such purposes, we may consider a log-linear model $\mathbf{A}\ln(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}$ with

$$\begin{aligned} \mathbf{A} &= \mathbf{I}_2 \otimes \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}, \\ \mathbf{X}_L &= \mathbf{1}_8, \qquad \boldsymbol{\beta} = \beta. \end{aligned} \tag{6.1}$$

Alternatively, the log-linear model corresponding to a homogeneous association (or to a no three-factor interaction) model may be obtained by taking

$$\mathbf{X}_L = \mathbf{1}_2 \otimes \mathbf{I}_4 \quad \text{and} \quad \boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})'. \tag{6.2}$$

Conditionally on this model, independence may be assessed via a test of the hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}_4$, where $\mathbf{C} = \mathbf{I}_4$. 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 $\mathbf{U}_L\mathbf{A}\ln(\boldsymbol{\theta}) = \mathbf{0}_4$, where $\mathbf{U}_L = [(1, -1) \otimes \mathbf{I}_4]$.

We assume a MAR mechanism for the subsequent analyses, because the MCAR mechanism does not seem reasonable ($p < 0.0001$) as suggested by the goodness-of-fit statistics for the MCAR mechanism conditionally on the MAR assumption using either the ML approach of Section 4.1 ($Q_L = 45.54$, $Q_P = 46.16$, $Q_N = 48.15$) or the WLS methodology of Section 4.2 ($Q_N = 44.75$). Fitting the log-linear model (6.1) under the ML methodology of Section 5.1, there is no evidence against the homogeneous linear-by-linear association model ($Q_L = 5.25$, $Q_P = 5.93$, $Q_N = 4.89$, $Q_W = 5.52$, df $= 7$), resulting in p-values ranging from 0.5480 to 0.6729. As the Wald statistic, in this case, uses only the estimates under the saturated model for the marginal probabilities of categorization, the statistic has the same value whether we fit the model using the ML approach or the hybrid (ML/WLS) methodology of Section 5.2 jointly with ML estimates under the MAR mechanism of Section 4.1. The estimate (standard error) of β is 0.2003 (0.0680) for the ML methodology, and 0.2036 (0.0685) for the hybrid approach, respectively, leading to p-values equal to 0.0032 ($Q_W = 8.67$) and to 0.0030 ($Q_W = 8.83$) for the conditional independence hypothesis under model (6.1). The corresponding likelihood ratio statistic ($Q_L = 8.41$, $p = 0.0037$) may also be easily obtained by the difference between the goodness-of-fit statistics (Q_L) of the independence and the homogeneous linear-by-linear association models. The Pearson ($Q_P = 8.75$, $p = 0.0031$) and Neyman ($Q_N = 8.07$, $p = 0.0045$) statistics are computed with expressions analogous to those in Appendix B.2 by comparing the estimates of the expected frequencies of the response classes, $\{\widehat{n}_{stc} = \sum_{r \in \mathcal{C}_{stc}} \widehat{y}_{str}\}$, under both models. Since the homogeneous linear-by-linear association model is a special case of the homogeneous association model (6.2), the latter also exhibits a good fit ($p = 0.3838$ for Q_N, and $p > 0.5400$ for the other statistics). The conditional independence test under this unordered case generates $Q_L = 10.57$, $Q_P = 11.01$, $Q_N = 10.11$, $Q_W(\text{ML}) = 10.90$ and $Q_W(\text{ML/WLS}) = 11.75$ ($p = 0.0318$, 0.0264, 0.0386, 0.0277, 0.0193). These results are similar to the ordered and unordered score tests statistics 8.06 ($p = 0.0045$) and 9.98 ($p = 0.0408$) obtained by Lipsitz and Fitzmaurice (1996), which also follow asymptotic χ^2_1 and χ^2_4 null distributions respectively, albeit we expected that they would be closer to Q_P (a score-type statistic) than to Q_N. Indeed, this may be due to their use of the alternative estimation method for the variance of the score vector proposed by Berndt et al. (1974).

When we fit the saturated MNAR model described in the last paragraph of Section 3 with the hybrid (ML/WLS) approach, the previous conclusions are maintained; the corresponding p-values are 0.3828 and 0.4124, respectively, for the goodness-of-fit of models (6.1) and (6.2), and 0.0001 and 0.0011, for their corresponding tests of conditional independence. These p-values suggest that under this MNAR model, the association between maternal smoking and child wheezing status is stronger than that obtained via the MAR assumption; this conclusion is also corroborated by the larger estimate (and lower standard error) of β in (6.1), namely 0.2398 (0.0626).

Example 2. In their analysis, Woolson and Clarke (1984) first assume that there are no cohort effects, that is, that the marginal probability of obesity does not vary between 1977 and 1981 conditionally on the gender and the age group of the children at the year of the measurement. This can be expressed as a (strictly) linear model $\mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{A} = \mathbf{I}_{10} \otimes \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} = \mathbf{I}_{10} \otimes \begin{pmatrix} (0,1) \otimes \mathbf{1}_4' \\ \mathbf{1}_2' \otimes (0,1) \otimes \mathbf{1}_2' \\ \mathbf{1}_4' \otimes (0,1) \end{pmatrix}$$

and

$$\mathbf{X} = \mathbf{I}_2 \otimes \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}' . \quad (6.3)$$

Woolson and Clarke (1984) propose a quadratic relationship on age for each gender, to obtain a more parsimonious structure, by taking

$$\mathbf{X} = \mathbf{I}_2 \otimes (\mathbf{1}_8, \mathbf{age}, \mathbf{age}^2), \quad (6.4)$$

where $\mathbf{age} = (6, 8, 10, 8, 10, 12, 10, 12, 14, 12, 14, 16, 14, 16, 18)'$ are the midpoints of the age intervals and $\mathbf{age}^2$ are its squared values. Alternatively, we could consider a piecewise linear regression model

$$\mathbf{X} = (\mathbf{1}_{30}, \mathbf{1}_2 \otimes (-4, -2, 0, -2, \mathbf{0}_{11}')', (0,1)' \otimes \mathbf{1}_{15}), \quad (6.5)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, β_2 is the linear variation of the probability of obesity between ages 6 and 10 for boys and girls, β_1 is the probability of obesity for boys assumed constant between 10 and 18 years old, and β_3 is the difference between the probability of obesity for girls and boys assumed constant between ages 6 and 18.

Although we are mainly interested in modelling the marginal probabilities of obesity, we may consider a reduced structure for the association parameters among the responses of the same individuals in the longitudinal setting using the log-linear model $\mathbf{ln}(\boldsymbol{\theta}) = (\mathbf{I}_{10} \otimes \mathbf{1}_8)\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{X} = \big(\mathbf{I}_{10} \otimes [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3], \mathbf{1}_{10} \otimes \big[\boldsymbol{v}_2 * (\boldsymbol{v}_1 + \boldsymbol{v}_3), \boldsymbol{v}_1 * \boldsymbol{v}_3, \boldsymbol{v}_1 * \boldsymbol{v}_2 * \boldsymbol{v}_3\big]\big), \quad (6.6)$$

where $\boldsymbol{v}_1 = (0,1)' \otimes \mathbf{1}_4$, $\boldsymbol{v}_2 = \mathbf{1}_2 \otimes (0,1)' \otimes \mathbf{1}_2$, $\boldsymbol{v}_3 = \mathbf{1}_4 \otimes (0,1)'$, $*$ denotes the element-wise multiplication, $\boldsymbol{\nu}$ contains the parameters associated to the natural constraints, the first 30 parameters included in $\boldsymbol{\beta}$ are associated to the marginal probabilities of obesity, and the last 3 (namely, φ_1, φ_2 and φ_3), to the association among the repeated measurements, which are assumed homogeneous over gender

and age levels. Here, φ_1 corresponds to the association between successive times (1977–1979 and 1979–1981), φ_2, between the first and last measurement, and φ_3 allows these paired associations to vary according the level of the third measurement. A first-order homogeneous Markov chain is a special case of this model and may be evaluated by the hypothesis $\varphi_2 = \varphi_3 = 0$.

Many reanalyses of the Woolson and Clarke (1984) data set (or some subset of it) have appeared in the literature. Among them, we mention Lipsitz et al. (1994) and Azzalini (1994). Both considered the MAR assumption; the first compared the hybrid (ML/WLS) approach with applications of WLS methodology of Koch et al. (1972) and Woolson and Clarke (1984), and the second constructed separate models for the marginal outcome and the longitudinal association structure, assuming a Markovian dependence for the latter.

In the following analyses, we assume the MAR mechanism since $Q_L = 152.6$ ($p = 0.0237$) and $Q_P = 143.9$ ($p = 0.0680$), df $= 120$, indicate that the MCAR mechanism may not be reasonable; here we must point out that the asymptotic tests may be imprecise due to the sparse configuration of the data in Table 2 [there are 81 (31%) small frequencies (<5), including 17 zeros]. Indeed, this makes the Neyman statistic (B.2) and WLS methodology of Section 4.2 untrustworthy because the results vary substantially accordingly to the values chosen to replace the null frequencies. For instance, substituting zeros by any value smaller than 0.125 generates negative WLS estimates for some probabilities of the saturated model. Lipsitz et al. (1994) did not mention which strategy they used to substitute null frequencies to perform the two-stage WLS procedure and we were not able to reproduce exactly their corresponding results labelled "KO." The ML approach we consider here does not require the substitution of zero frequencies. However, as Newton–Raphson and Fisher's scoring algorithms do not impose constraints on the probabilities, some linear models may provide negative estimates in some iteration when there are null frequencies. Actually, this happened for models (6.3), (6.4) and (6.5). For these cases, replacing the null frequencies by 10^{-6} bypassed the problem. On the contrary, the hybrid (ML/WLS) approach did not generate negative estimates for the probabilities; the corresponding results of Lipsitz et al. (1994), which they labelled "ML," appear to have used the 10^{-6} replacement.

Using the ML approach, the log-linear model (6.6) seems to fit well ($Q_L = 37.5$, $Q_P = 24.6$, $Q_W = 21.8$, df $= 37$, $p > 0.4400$), and there is no evidence supporting further simplification to a Markovian dependence ($Q_W = 85.3$, df $= 2$). We proceed by fitting linear models (6.3), (6.4) and (6.5) by WLS methodology using the ML estimates from the log-linear model (6.6) for $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\mathbf{V}}_{\widetilde{\boldsymbol{\theta}}}$ of Section 5.2. The quadratic relationship (6.4) does not appear to fit well ($p = 0.0717$, df $= 24$). Lipsitz et al. (1994) fitted this linear model by the hybrid methodology without imposing any association structure and obtained $p = 0.053$, although they would have obtained $p = 0.0972$ without replacing the null frequencies. The less restrictive model (6.3) fits rather well ($p = 0.5773$, df $= 16$). Indeed, a careful examination

of the WLS estimates of this model suggested that the marginal probability of obesity increases between age intervals 5–7 and 9–11, but appears to stabilize from this point until ages 17–19, and the difference between the probabilities for boys and girls may be taken as approximately constant in time. Model (6.5) reflects such behaviour and has an acceptable fit ($p = 0.2429$, df $= 27$), with estimates $\widetilde{\boldsymbol{\beta}} = (0.2159, 0.0314, 0.0252)'$ and standard errors $(0.0075, 0.0031, 0.0104)'$. The gender effect for models (6.3) and (6.5), both yielding $p = 0.0157$ when Wald tests are considered, are substantially stronger than for model (6.4), $p = 0.1042$. Again, without the reduction of the association structure, Lipsitz et al. (1994) obtained $p = 0.098$ (this would have been $p = 0.1420$ if zeros had not been replaced). Replacing the null frequencies does not alter any of the results of the linear models fitted subsequently to the log-linear model.

7 Simulations

A series of simulation studies would be required to assess the properties of all statistics presented in the last sections in small and moderate sized samples. Parsimoniously, we focus only on the ML and hybrid (ML/WLS) estimators for the log-linear model (6.1). For the same reason, we also consider only the MAR assumption, because MNAR models require lengthy evaluations as those considered in Poleto et al. (2011a), for example.

We generated 10,000 Monte Carlo replicates of product-multinomial distributions with parameters obtained from the ML fit to the data in Example 1. In Table 3, we display the Monte Carlo estimates of the bias, standard deviation (SD) and mean squared error (MSE) of the estimators of β in (6.1) as well as the mean of the corresponding standard errors [Mean(SE)], ratio between Mean(SE) and SD, and 95% coverage probability (CovP).

Table 3 *Monte Carlo estimates of the bias, standard deviation (SD) and mean squared error (MSE) of the estimators of β in (6.1), the mean of the standard errors [Mean(SE)], ratio between Mean(SE) and SD, and 95% coverage probability (CovP)*

$\mathbf{N}_{++}$	Approach	Bias	SD	MSE	Mean(SE)	Mean(SE)/SD	95% CovP
(100, 100)	ML	0.0053	0.1823	0.0333	0.1720	0.944	0.954
	hybrid	0.0501	0.1622	0.0288	0.1857	1.145	0.975
(200, 200)	ML	0.0020	0.1185	0.0141	0.1173	0.989	0.956
	hybrid	0.0224	0.1105	0.0127	0.1205	1.090	0.970
(661, 477)	ML	0.0018	0.0693	0.0048	0.0687	0.990	0.948
	hybrid	0.0078	0.0676	0.0046	0.0692	1.023	0.953
(10,000, 10,000)	ML	0.0001	0.0162	0.0003	0.0161	0.993	0.948
	hybrid	0.0004	0.0162	0.0003	0.0161	0.995	0.949

The bias of the ML estimator is between 4 to 10 times smaller than that obtained via the hybrid approach, but the SD and MSE obtained via ML are unexpectedly larger than the corresponding ones produced via the hybrid approach; this is also true for the largest sample size, although the difference has a smaller magnitude. These differences decrease when the sample size increases. It is also curious that for the smaller sample sizes, the Mean(SE) is larger than the SD when the hybrid estimator is considered. This may have avoided the anticipated decrease of the 95% CovP in these cases. Although we do not expect that in general, the hybrid approach will always generate smaller MSEs than the ML procedure nor that the confidence intervals of the hybrid approach will usually behave conservatively, our results suggest that this methodology, indeed, may have an acceptable behaviour in some cases.

8 Concluding remarks

We present a unified matrix notation to describe methods for fitting functional linear models to product-multinomial data when there are missing responses and develop computational subroutines for such purposes. Strictly linear and log-linear models assuming MAR and MCAR mechanisms are considered under ML methodology. More general functional linear models are treated via a WLS approach under a more restrictive MCAR assumption. Greater flexibility is achieved via a hybrid strategy, where relatively simple (e.g., saturated) models for the measurement process are fitted in the first step via ML under MCAR, MAR or MNAR missingness mechanisms, and the estimated marginal probabilities of categorization and their covariance matrix are used in a second step to fit more general functional linear models via WLS. Here, the unique additional programming effort appears in the first step under MNAR mechanisms, where the user needs to employ built-in optimization functions to obtain ML estimates.

The inferential framework described by Paulino (1991) is a special case of ours when there are no explanatory variables (subpopulations), that is, $S = 1$, and the number of linear and log-linear functions to be modelled and fitted via ML is equal to the number of parameters of the multinomial distribution, that is, the number of rows of $\mathbf{A}$ in (5.1) and (5.9) are $u = R - 1$. Hence, our extension allows the handling of explanatory variables and parameters that are common to several subpopulations, as well as the fitting of larger classes of linear and log-linear models via the ML methodology.

We illustrated the methodology with two examples under a MAR assumption and one under a MNAR model. As there is no way to decide among saturated MAR and MNAR mechanisms merely on statistical grounds (Molenberghs et al., 2008), sensitivity analyses should be considered. An informal route is to fit and compare different meaningful MNAR models. A more formal strategy, adopted by Kenward et al. (2001), is to consider over-parameterized missingness models.

The so-called sensitivity parameters are then replaced by known values to enable the estimation of the remaining parameters. These are repeatedly estimated for a series of fixed values of the former. The union of the estimates results in the so-called ignorance region, and the union of the credible regions, in the so-called uncertainty region. Vansteelandt et al. (2006) indicate three methods for constructing the uncertainty regions and also provide appropriate definitions of consistency and coverage. Both kinds of sensitivity analyses can be conducted in combination with the hybrid (ML/WLS) approach using the same principle described for the MNAR mechanism. This was performed for Example 1 in Poleto (2006) and raised some difficulties with the conclusion that maternal smoking is associated to wheezing in children. When there is prior information or when classical sensitivity analysis is not feasible due to a large number of sensitivity parameters, a Bayesian framework may offer advantages, as Poleto et al. (2011b) indicate. However, these authors showed that both approaches have subjective components that can impact results in nontrivial ways, and therefore, a careful evaluation is indispensable.

Our simulations suggested that the hybrid approach may be a viable alternative to the ML methodology in certain cases when the latter cannot be easily employed. Although additional simulation studies are required to improve our understanding of the scenarios where the hybrid strategy may (or not) be adequate, general rules usually cannot be formulated. When MNAR mechanisms are considered, Poleto et al. (2011a) showed that ML estimators and likelihood ratio tests have undesirable asymptotic properties either if estimates of the conditional probabilities of missingness are on the boundary of the parameter space, or if the parameters of saturated models are nonidentifiable; even in standard cases the bias may be low only for large sample sizes (500 to 20,000), but it was always smaller than the naive analysis based on the units with no missing data, if the MNAR model is correctly specified. The same behaviour is likely shared by the hybrid approach due to the use of the ML methodology in the first step.

The saturated models considered for Examples 1 and 2 include, respectively, 18 and 80 marginal probabilities of categorization ($S \times R$). This illustrates that the number of parameters increases exponentially with the number of variables. The formulation of models becomes naturally more complex with the increasing number of parameters, but poses no additional computational difficulties; the algorithm takes, in general, less than a second to converge. Fitting models to sparse tables, however, might bring in some problems as we mentioned in the discussion of Example 2 in Section 6.

Appendix A: Fisher information matrices

A.1 Saturated models for the marginal probabilities of categorization

The expectation of $-\mathbf{H}(\overline{\boldsymbol{\theta}})$, using (4.5), (4.7) and (4.8), lead to the Fisher information matrices under the MAR and the MCAR mechanisms corresponding to $\overline{\boldsymbol{\theta}}$, denoted by $\boldsymbol{\mathcal{I}}(\overline{\boldsymbol{\theta}}, \{\boldsymbol{\alpha}_{st}^{\mathrm{MAR}}\})$ and $\boldsymbol{\mathcal{I}}(\overline{\boldsymbol{\theta}}, \{\boldsymbol{\alpha}_{st}^{\mathrm{MCAR}}\})$, which are block diagonal matrices

with blocks respectively, given by

$$\begin{aligned}\mathcal{I}_s(\overline{\boldsymbol{\theta}}_s, \{\boldsymbol{\alpha}_{st}^{\mathrm{MAR}}\}) &= n_{s++}\sum_{t=1}^{T_s}\overline{\mathbf{Z}}_{st}\left(\mathbf{D}_{\overline{\boldsymbol{\alpha}}_{st}^{\mathrm{MAR}}}\mathbf{D}_{\overline{\boldsymbol{\theta}}_{st}}^{-1} + \frac{\alpha_{t(sR_{ts})}}{1-\mathbf{1}'_{R_{st}-1}\overline{\boldsymbol{\theta}}_{st}}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right)\overline{\mathbf{Z}}'_{st}, \quad \text{(A.1)}\\ & s = 1, \ldots, S,\end{aligned}$$

where $\overline{\boldsymbol{\alpha}}_{st}^{\mathrm{MAR}} = (\mathbf{I}_{R_{st}-1}, \mathbf{0}_{R_{st}-1})\boldsymbol{\alpha}_{st}^{\mathrm{MAR}} = (\alpha_{t(cs)}, c = 1, \ldots, R_{st}-1)'$, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$ and

$$\begin{aligned}\mathcal{I}_s(\overline{\boldsymbol{\theta}}_s, \{\boldsymbol{\alpha}_{st}^{\mathrm{MCAR}}\}) &= n_{s++}\sum_{t=1}^{T_s}\alpha_{t(s)}\overline{\mathbf{Z}}_{st}\left(\mathbf{D}_{\overline{\boldsymbol{\theta}}_{st}}^{-1} + \frac{1}{1-\mathbf{1}'_{R_{st}-1}\overline{\boldsymbol{\theta}}_{st}}\mathbf{1}_{R_{st}-1}\mathbf{1}'_{R_{st}-1}\right)\overline{\mathbf{Z}}'_{st}, \quad \text{(A.2)}\\ & s = 1, \ldots, S.\end{aligned}$$

A.2 Log-linear models for the marginal probabilities of categorization

The Fisher information matrix under the $\mathcal{M}$ (MAR or MCAR) mechanism for the log-linear model discussed in Section 5.1 is expressed as

$$\begin{aligned}\mathcal{I}_{LL}(\boldsymbol{\beta}, \{\boldsymbol{\alpha}_{st}^{\mathcal{M}}\}) = \sum_{s=1}^{S}\mathbf{X}'_s\Bigg[n_{s++}\mathbf{I}_R - \sum_{t=2}^{T_s}(\mathbf{D}_{\mathbf{v}_{st}^{\mathcal{M}}} - \mathbf{D}_{\mathbf{w}_{st}^{\mathcal{M}}}\mathbf{Z}_{st}\mathbf{Z}'_{st})\Bigg]\\ \times\{\mathbf{D}_{\boldsymbol{\theta}_s(\boldsymbol{\beta})} - \boldsymbol{\theta}_s(\boldsymbol{\beta})[\boldsymbol{\theta}_s(\boldsymbol{\beta})]'\}\mathbf{X}_s,\end{aligned} \quad \text{(A.3)}$$

where

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

Appendix B: Additional goodness-of-fit statistics

B.1 Saturated models for the marginal probabilities of categorization

The Pearson and Neyman statistics for testing MCAR conditionally on the MAR mechanism in Section 4.1 are

$$\begin{aligned}Q_P(\mathrm{MCAR}|\mathrm{MAR}) &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}}\frac{(n_{stc} - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s)^2}{n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s}\\ &= \sum_{s=1}^{S}(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s)'(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s})(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s),\end{aligned} \quad \text{(B.1)}$$

$$
\begin{aligned}
Q_N(\text{MCAR}|\text{MAR}) &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{stc} - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s)^2}{n_{stc}} \\
&= \sum_{s=1}^{S}(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s)'(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}_{\mathbf{p}_s}^{-1})(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s),
\end{aligned}
\tag{B.2}
$$

where $\mathbf{N}_{s+} = (n_{st+} \otimes \mathbf{1}'_{R_{st}}, t = 1, \ldots, T_s)'$ is the vector with the same dimension as $\mathbf{N}_s$ that contains the total observed frequencies of units with each missingness pattern in the sth subpopulation sequentially repeated according to the number of classes in each pattern (note that $\mathbf{p}_s = \mathbf{D}_{\mathbf{N}_{s+}}^{-1}\mathbf{N}_s$).

B.2 Unsaturated models for the marginal probabilities of categorization

In Section 5.1, the Pearson and Neyman statistics for testing (M, MCAR) conditionally on the MAR mechanism, where M is M_L or M_{LL}, are

$$
\begin{aligned}
&Q_P(M, \text{MCAR}|\text{MAR}) \\
&\quad= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{stc} - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M))^2}{n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M)} \\
&\quad= \sum_{s=1}^{S}(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M))'(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)})(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)),
\end{aligned}
\tag{B.3}
$$

$$
\begin{aligned}
&Q_N(M, \text{MCAR}|\text{MAR}) \\
&\quad= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{stc} - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M))^2}{n_{stc}} \\
&\quad= \sum_{s=1}^{S}(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M))'(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}_{\mathbf{p}_s}^{-1})(\mathbf{p}_s - \mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)).
\end{aligned}
\tag{B.4}
$$

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

$$
\begin{aligned}
&Q_P(M|\text{MAR}) \\
&\quad= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{stc} - n_{s++}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M)\widehat{\alpha}_{t(cs)})^2}{n_{s++}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M)\widehat{\alpha}_{t(cs)}} \\
&\quad= \sum_{s=1}^{S}(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)])'(\mathbf{D}_{\mathbf{N}_s}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)})(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)]),
\end{aligned}
\tag{B.5}
$$

$$
\begin{aligned}
&Q_N(M|\text{MAR}) \\
&\quad = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{stc} - n_{s++}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M)\widehat{\alpha}_{t(cs)})^2}{n_{stc}} \qquad \text{(B.6)} \\
&\quad = \sum_{s=1}^{S}\big(\mathbf{1}_{R+l_s} - \mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s}\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)\big)'\mathbf{D}_{\mathbf{N}_s}\big(\mathbf{1}_{R+l_s} - \mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s}\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)\big), \\
&Q_P(M|\text{MCAR}) \\
&\quad = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M))^2}{n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M)} \qquad \text{(B.7)} \\
&\quad = \sum_{s=1}^{S}\big(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)]\big)'\big(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s(M)}\big)\big(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)]\big), \\
&Q_N(M|\text{MCAR}) \\
&\quad = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{c=1}^{R_{st}} \frac{(n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s - n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s(M))^2}{n_{st+}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s} \qquad \text{(B.8)} \\
&\quad = \sum_{s=1}^{S}\big(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)]\big)'\big(\mathbf{D}_{\mathbf{N}_{s+}}\mathbf{D}^{-1}_{\mathbf{Z}'_s\widehat{\boldsymbol{\theta}}_s}\big)\big(\mathbf{Z}'_s[\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}}_s(M)]\big),
\end{aligned}
$$

where $\widehat{\alpha}_{t(cs)} = n_{stc}/(n_{s++}\mathbf{z}'_{stc}\widehat{\boldsymbol{\theta}}_s)$.

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

$$Q_W(M_L|\mathcal{M}) = (\mathbf{UA}\widehat{\boldsymbol{\theta}})'\big(\mathbf{UA}\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\big)^{-1}\mathbf{UA}\widehat{\boldsymbol{\theta}}, \qquad \text{(B.9)}$$

$$Q_W(M_{LL}|\mathcal{M}) = \big(\mathbf{U}_L\mathbf{A}\ln(\widehat{\boldsymbol{\theta}})\big)'\big(\mathbf{UAD}^{-1}_{\widehat{\boldsymbol{\theta}}}\widehat{\mathbf{V}}^{\mathcal{M}}_{\widehat{\boldsymbol{\theta}}}\mathbf{D}^{-1}_{\widehat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\big)^{-1}\mathbf{U}_L\mathbf{A}\ln(\widehat{\boldsymbol{\theta}}). \qquad \text{(B.10)}$$

Asymptotically, under the model M and the MAR mechanism

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

and, additionally under the MCAR mechanism,

$$
\begin{aligned}
Q_P(M|\text{MCAR}) &\overset{a}{\approx} Q_N(M|\text{MCAR}) \\
&\overset{a}{\approx} Q_W(M|\text{MCAR}) \xrightarrow{a} \chi^2_{(u-p)}, \\
Q_L(M, \text{MCAR}|\text{MAR}) &\overset{a}{\approx} Q_P(M, \text{MCAR}|\text{MAR}) \\
&\overset{a}{\approx} Q_N(M, \text{MCAR}|\text{MAR}) \xrightarrow{a} \chi^2_{(u-p+g)}.
\end{aligned}
$$

Appendix C: Augmentation of models

C.1 Linear models

For the (strictly) linear model (5.3), when $u < S(R-1)$, we need to augment the model to

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

with an $[S(R-1)-u] \times SR$ matrix $\mathbf{A}_0$, basis of the orthocomplement of the vector space generated by $(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R)'$. This formulation encompasses the former, but has $S(R-1)-u$ additional nuisance parameters, $\boldsymbol{\beta}_0$. This requires us to substitute (5.4) and (5.5), respectively, by

$$\overline{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \left[\mathbf{I}_S \otimes (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})\right] \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \\ \mathbf{A}_0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{1}_S \\ \boldsymbol{\beta}_0 \end{pmatrix} \tag{C.2}$$

and

$$\mathbf{W} = \left[\mathbf{I}_S \otimes (\mathbf{I}_{R-1}, \mathbf{0}_{R-1})\right] \begin{pmatrix} \mathbf{A} \\ \mathbf{I}_S \otimes \mathbf{1}'_R \\ \mathbf{A}_0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X} & \mathbf{0}_{u,S(R-1)-u} \\ \mathbf{0}_{S,p} & \mathbf{0}_{S,S(R-1)-u} \\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u} \end{pmatrix}. \tag{C.3}$$

In (5.6), we also need to replace $\mathbf{A}$ by $(\mathbf{A}', \mathbf{A}'_0)'$ and $\mathbf{X}$ by

$$\begin{pmatrix} \mathbf{X} & \mathbf{0}_{u,S(R-1)-u} \\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u} \end{pmatrix}. \tag{C.4}$$

C.2 Log-linear models

For the (generalized) log-linear model (5.9), when $u < S(R-1)$, we must include a basis of the orthocomplement of the vector space generated by $(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R)'$, that is, an $[S(R-1)-u] \times SR$ matrix $\mathbf{A}_0$ such that the model

$$M_{LL} : \begin{pmatrix} \mathbf{A} \\ \mathbf{A}_0 \end{pmatrix} \ln(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{X}_L\boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \end{pmatrix} \tag{C.5}$$

may be re-expressed in the form (5.7) as

$$M_{LL} : \ln(\boldsymbol{\theta}) = (\mathbf{I}_S \otimes \mathbf{1}_R)\boldsymbol{\nu} + \left(\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{X}_L, \mathbf{A}'_0(\mathbf{A}_0\mathbf{A}'_0)^{-1}\right) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \end{pmatrix}. \tag{C.6}$$

In (5.12), we need to substitute $\mathbf{A}$ by $(\mathbf{A}', \mathbf{A}'_0)'$ and $\mathbf{X}_L$ by

$$\begin{pmatrix} \mathbf{X}_L & \mathbf{0}_{u,S(R-1)-u} \\ \mathbf{0}_{S(R-1)-u,p} & \mathbf{I}_{S(R-1)-u} \end{pmatrix}. \tag{C.7}$$

Acknowledgments

This research received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, and Fundação para a Ciência e Tecnologia (FCT) through the research centres CEMAT-IST and CEAUL-FCUL, Portugal.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley. MR1914507

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81**, 767–775.

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; Corrigenda 1232. MR0940999; MR0997605

Berndt, E. R., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3**, 653–666.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association* **63**, 542–551. MR0232485

Chen, T. T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30**, 629–642. MR0403086

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with comments). *Journal of the Royal Statistical Society, Ser. B* **39**, 1–38. MR0501537

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. MR0381144

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., Darroch, J. N., Freeman, D. H. Jr. and Tolley, H. D. (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. MR0651474

Imrey, P. B., Koch, G. G., Stokes, M. E., Darroch, J. N., Freeman, D. H. Jr. and Tolley, H. D. (1982). Categorical data analysis: Some reflections on the log linear model and logistic regression. Part II: Data analysis. *International Statistical Review* **50**, 35–63. MR0668609

Kenward, M. G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science* **13**, 236–247. MR1665713

Kenward, M. G., Goetghebeur, E. and Molenberghs, G. (2001). Sensitivity analysis for incomplete categorical data. *Statistical Modelling* **1**, 31–48.

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.

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 Methods and 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.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1994). Weighted least squares analysis of repeated categorical measurements with outcomes subject to nonresponse. *Biometrics* **50**, 11–24.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley. MR1925014

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, Ser. B* **70**, 371–388. MR2424758

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

Molenberghs, G., Goetghebeur, E., Lipsitz, S. R. and Kenward, M. G. (1999). Nonrandom missingness in categorical data: Strengths and limitations. *The American Statistician* **53**, 110–118.

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. MR1160735

Paulino, C. D. and Silva, G. L. (1999). On the maximum likelihood analysis of the general linear model in categorical data. *Computational Statistics & Data Analysis* **30**, 197–204. MR1702889

Paulino, C. D. and Soares, P. (2003). Analysis of rates in incomplete Poisson data. *The Statistician* **52**, 87–99. MR1973885

Poleto, F. Z. (2006). Analysis of categorical data with missingness. M.Sc. thesis, Universidade de São Paulo (in Portuguese).

Poleto, F. Z., Singer, J. M. and Paulino, C. D. (2011a). Missing data mechanisms and their implications on the analysis of categorical data. *Statistics and Computing* **21**, 31–43. MR2746601

Poleto, F. Z., Paulino, C. D., Molenberghs, G. and Singer, J. M. (2011b). Inferential implications of over-parameterization: A case study in incomplete categorical data. *International Statistical Review* **79**, 92–113.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592. MR0455196

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. MR0899519

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall. MR1692799

Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16**, 953–979. MR2281311

Williamson, G. D. and Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics* **49**, 194–203. MR1279436

Woolson, R. F. and Clarke, W. R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society, Ser. A* **147**, 87–99.

F. Z. Poleto
J. M. Singer
Instituto de Matemática e Estatística
Universidade de São Paulo
Caixa Postal 66281
São Paulo, SP, 05314-970
Brazil
E-mail: fpoleto@ime.usp.br
jmsinger@ime.usp.br

C. D. Paulino
Instituto Superior Técnico
Universidade Técnica de Lisboa
Av. Rovisco Pais
Lisboa, 1049-001
Portugal
E-mail: dpaulino@math.ist.utl.pt