Discrete choice models for nonignorable missing data

Esmeralda A. Ramalho* Richard J. Smith
Departamento de Economia Department of Economics
Universidade de Évora University of Warwick, UK
and CEMAPRE, Portugal

This draft: December 2002

Abstract

In econometrics, the data sets available for the practitioner have often some missing values. This paper suggests a likelihood-based unified approach to deal with several nonignorable missing data problems when the variable of interest is discrete and the probability of response conditional on the variable of interest is independent of the covariates. In particular, we address the case where only the value of the variable of interest is missing for some subjects and the case where both the variable of interest and the covariates are missing for some sampling units. Moreover, we also consider these two problems when a supplementary random sample consisting of observations of all covariates is available. Our approach consists of a reinterpretation of the missing data problem in discrete choice models. We propose a formulation for the nonresponse problems which is a modification of the standard regression model for choice-based sampling and, then, we extend Imbens’ (1992) efficient generalized method of moments estimators suggested for choice-based samples to handle them. A small Monte Carlo simulation study concerning most of the estimators suggested revealed very promising results.

JEL: C25, C51.

Keywords: nonresponse, discrete choice models, generalized method of moments.

*The author gratefully acknowledges partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. Helpful comments and suggestions were given by seminar participants at the Erasmus University Rotterdam, the 57th ESEM, Venice, and the Winter ESEM, Budapest. Address for correspondence: Esmeralda A. Ramalho, Departamento de Economia, Universidade de Évora, Largo dos Colegiãis 2, 7000-803 Évora, Portugal. E-mail: ela@uevora.pt.
1 Introduction

In econometrics, the data sets available for the practitioner have often some missing values. Frequently, survey sampling is conducted to gather complete information on all sampling units but, due to a variety of reasons, for a fraction of the subjects, either no data at all is available or information on one or more variables is missing. Indeed, often some sampling units refuse to participate in the study or to answer all the questionnaire, the interviewer is not able to contact all the sampling units or fails to ask all questions, and in data processing some questionnaires or part of questionnaires may be destroyed. Conversely, there are also cases where the presence of missing values is planned. For example, in variable probability sampling (VPS) an observation is randomly drawn from the population, the stratum to which it belongs is identified, and it is kept in the sample with a given probability of retention, defined by the agent who collects the sample. While in the latter situation, as the sampling scheme deliberately generates incomplete data, the mechanism which governs the missingness pattern is known, in the former, as the data is missing for reasons beyond the control of the researcher, in general, nothing is known about the missingness mechanism. This paper focus on problems where missingness is uncontrolled by the researcher.

An enormous statistical literature has been developed on this topic, from which we emphasise the books by Little and Rubin (1987) and Schafer (1997). Two forms of missing data are commonly distinguished: unit nonresponse, where for some sampling units no data at all is available, and item nonresponse, where only part of the information is missing. To deal with the former class of problems, most of the literature suggests the use of weighting adjustments, which involve the assignment of weights to respondents to compensate for the systematic differences relative to nonrespondents. For the latter form of nonresponse, the majority of the papers proposes imputation inference procedures in which the missing values are filled-in to produce complete data sets. However, many empirical studies do not follow any of these approaches. They simply discard all the sampling units with missing values and employ the usual inference procedures for random sampling (RS), which seriously bias the results when the characteristics of respondents and nonrespondents differ systematically.

In econometrics, the issue of nonresponse has been addressed mostly in the context of panel

\footnote{Moreover, the statistical literature often deals with two stage sampling designs where in a first stage the main sample is collected and in a second stage further variables, more expensive and/or difficult to collect, are measured only for a part of the survey participants.}
data studies, where often some sampling units drop out after participating in the initial waves of the survey, causing attrition [see, for example, Ridder (1990), Fitzgerald, Gottschalk and Moffitt (1998) and Hirano, Imbens, Ridder and Rubin (2001)]. However, Horowitz and Manski (1995, 1998, 2001) provide a general discussion on nonparametric identification of regressions with missing data on either/both the variable of interest or/and the covariates.

In this paper we propose a likelihood-based unified approach to deal with the presence of missing data in cross-section data where the variable of interest is discrete. We address cases where the variable of interest, and possibly the covariates, are missing for some sampling units, the response or nonresponse being dependent on the variable of interest. This setup is appropriated to handle situations where, due to the nature of some of the questions contained in the survey, a fraction of the sample either omits the answer to those questions or refuses to participate in the survey. Specifically, we address the case where only the value of the variable of interest is missing for some subjects, which we designate here as item nonresponse (INR), and the case where both the variable of interest and the covariates are missing for some sampling units, termed unit nonresponse (UNR). Moreover, we also consider these two problems when a supplementary random sample (SRS) consisting of observations of all covariates is available, which are denoted by, respectively, INRS and UNRS. This setting generalizes that for INR and UNR because although the variable of interest is also missing for all individuals included in the SRS, as this sample is random and is assumed to be independent of the main sample, similarly to INR and UNR, only the missing data mechanism of the main sample has to be dealt with. Thus, the analysis is mainly focussed on the general problems of INRS and UNRS, being sometimes specialized for INR and UNR.

In order to undertake the joint analysis of all cases, we consider the presence of three types of sampling units. Two of them belong to the main sample: those for which all the data is available and those for which the information is absent or is incomplete, designated, respectively, as respondents and nonrespondents. Thus, the term response is only reserved for complete responses, while the term nonresponse may describe either an incomplete or an inexistent response. The third class of sampling units are those included in the SRS. We assume that all incomplete data patterns of interest are underpinned by the same (unknown) missing data mechanism, defined in terms of the conditional probability of response given the variable of interest. Consequently, we assume that the individual characteristics included in the covariates do not contain any additional information on the unit response/nonresponse relative to that provided by the variable of interest.

The nonignorable nature of the nonresponse arises because the rate of response is different
across the \((C + 1)\) different values taken by the variable of interest. Thus, the observed data provide a distorted picture of the features of the population of interest. This creates a situation where, for likelihood-based inference, the model describing the available data is a complicated function of the structural model, defined by the conditional probability of the variable of interest given the covariates, assumed for the target population. In fact, besides the structural model, also the missing data mechanism, defined by \((C + 1)\) conditional probabilities of response given the variable of interest, has to be specified. Moreover, a further complication may arise. As the observation of the sampling units depends on the variable of interest, the covariates are not ancillary for the parameters of interest, except for cases where they are measured for all individuals (INRS and INR). This prevents efficient estimation to be conditional on the covariates, requiring the formalization of its marginal distribution for maximum likelihood (ML) estimation.

The likelihood-based estimators proposed in this paper, besides the assumption that the conditional probability of response given the variable of interest is independent of the covariates, require only the correct specification of the structural model. On the one hand, prior knowledge on the probabilities defining the missing data mechanism is not necessary. For each value of the variable of interest, these probabilities are written as a ratio of the proportion of units choosing that alternative in the sample and in the population, and these proportions are then taken as additional parameters to be estimated. On the other hand, the formulation of the distribution of the covariates is circumvented by maximizing the log-likelihood function over all discrete distributions whose support contains the covariate values observed in the data. Thus, we handle this problem semiparametrically, by replacing the unknown distribution by its empirical distribution function.

Our approach consists of a reinterpretation of the missing data problem in discrete choice models. Assuming that the initial unobservable sample is random, it arises straightforwardly that, due to the dependence of the probability of response on the variable of interest, in the observable incomplete sample the fraction of units choosing each alternative differs from the corresponding proportion in the population. Although for different reasons, the same sort of divergence is present in choice-based (CB) samples, a stratified sampling scheme where each of the discrete values taken by the variable of interest defines one stratum, for which the proportion in the sample is chosen by the sampling agent. Thus, exploiting the similarity of the two problems, we formulate all the aforementioned incomplete data patterns by modifying the common model for CB sampling and extend the efficient generalized method of moments (GMM) estimators suggested by Imbens (1992) for CB sampling to handle them.
This paper is organized as follows. Section 2 formalizes the regression models for the missing data problems of interest. GMM estimators for these models are developed and compared in section 3. Section 4 reports some Monte Carlo evidence on the performance in practice of some of the proposed estimators. Finally, section 5 concludes. Some technical proofs are relegated to the appendix.

2 Model specification

This section discusses the construction of regression models to handle the missing data problems of interest. Subsection 2.1 states the general notation employed throughout this paper. Subsection 2.2 defines and examines the assumed missing data mechanism. Subsection 2.3 modifies the CB sampling setup to describe the missing data problems. Subsection 2.4 presents the observed data likelihood functions which will be used in section 3 to derive likelihood-based GMM estimators.

2.1 Notation

Consider a sample of \( i = 1, \ldots, N \) individuals and let \( Y \) be the response variable of interest, taking values on a set of \( (C + 1) \) mutually exclusive alternatives, \( \mathcal{Y} = \{0, 1, \ldots, C\} \), with \( X \) a vector of \( k \) exogenous variables. Both \( Y \) and \( X \) are random variables defined on \( \mathcal{Y} \times \mathcal{X} \) with population joint density function

\[
f(y,x) = \Pr(y|x, \theta) f(x),
\]

where \( \Pr(y|x, \theta) \) is a function known up to the parameter vector \( \theta \) and the marginal density function \( f(x) \) for \( X \) is unknown. Our interest is consistent estimation of and inference on the parameter vector \( \theta \) in \( \Pr(y|x, \theta) \). With no loss of clarity in the exposition, the dependence of \( f(y,x) \) on \( \theta \) is omitted in the notation of (1). The same convention is adopted from now on in the notation of all joint density functions.

In the population, the probability of observing an individual for which \( Y = y \) is

\[
Q_y = \int_X \Pr(y|x, \theta) f(x) dx,
\]

where \( 0 < Q_y < 1 \) and \( \sum_{y \in \mathcal{Y}} Q_y = 1 \). Often, this probability may be known, the available estimates for \( Q_y \) being obtained from a large random sample, for example a census. When this aggregate information is available, similarly to what is usually done in the literature for CB sampling, we deal with it as if it was exact; see, for example, Manski and Lerman (1977), Imbens (1992) and Wooldridge (1999, 2001).
Assume that a random sample of size $N$ is to be collected, but only $n$ individuals provide all the information required. The former sample is designated as the initial or incomplete sample, while the latter is termed the complete sample. Consider that alternative $Y = y$ is chosen by $N_y$ individuals, from which only $n_y$ have complete questionnaires. Hence, $N = \sum_{y \in Y} N_y$ and $n = \sum_{y \in Y} n_y$. In addition, assume also that an independent SRS of size $m$ is drawn from the population of interest and define $N_m = N + m$ and $n_m = n + m$. As all incomplete data problems considered involve missing data on $(Y, X)$ or on $Y$, we always observe $n_y$, $n$ and $m$ but never $N_y$ because $Y$ is always missing for some individuals.

The size of the initial random sample, $N$, is always available for INRS (and INR), since the covariates are measured for all individuals. However, for UNRS (and UNR) the total number of individuals involved in the survey may or may not be known for the econometrician. Throughout this paper we formalize all the models assuming knowledge on the initial sample size $N$ for three reasons. Firstly, it is possible to follow the same approach for both INRS (and INR) and UNRS (and UNR). Secondly, the models for UNRS (and UNR) are straightforwardly adapted for when that information is not available. Finally, the inclusion of this information improves inference on the parameters of interest; see Li and Qin (1998) for a discussion of several examples of biased data where the incorporation of this information improves semiparametric likelihood-based inference.

2.2 The missing data mechanism

Assume that, conditional on $Y$, the response is independent of the individual characteristics contained in $X$, that is, the influence of $X$ over the response/nonresponse occurs only through $Y$. Define the binary indicator $R$ which takes the value 1 if $(Y, X)$ is fully observed or 0 if either $Y$ or $(Y, X)$ is missing. Thus, the conditional probability of observing a respondent unit given $Y$ is

$$P_y = \Pr (R = 1|Y = y, x) = \Pr (R = 1|Y = y).$$

(3)

In case a SRS of covariates is available, define the binary indicator $S$ which takes the value 1 when the sampling unit belongs to the supplementary data set and 0 otherwise. In this setup, although the variable of interest is not measured in the SRS, as this data set is random and independent from the main sample, we merely need to describe the missingness pattern of the main data set, which is still given by $P_y$ of (3), since $\Pr (R = 1|Y = y, x, S = 0) = \Pr (R = 1|Y = y)$.

Our integrated approach for all missing data problems considered is based on this formulation. In all cases, we assume that $0 < P_y < 1$. On the one hand, in the main incomplete sample, if, for
a given alternative \( Y = j, P_j = 0 \), there would not exist complete data in any of the sampling units choosing \( j \). On the other hand, \( P_j = 1 \) would indicate that none of the individuals choosing \( Y = j \) would have missing values. Combining (3) and (2), we may obtain the probability of observing a respondent unit,

\[
\Pr(R = 1) = \sum_{y \in \mathcal{Y}} P_y Q_y,
\]

which is, in general, unknown, because although \( Q_y \) may sometimes be known, usually the non-response mechanism, defined by \( P_y \), is unknown.

When the rate of response is the same for all alternatives, that is, \( P_y \) is identical for all \( y \), the data are said to be missing completely at random (MCAR), according to the definition of Little and Rubin (1987). This yields a random complete sample, with which, naturally, RS estimation methods may be used. In this case, the nonresponse is ignorable, since individuals with missing values are no different from those with complete information.

Note that with the missing data process in (3), if we had considered the circumstance where only information on \( X \) would be missing, the data would be said to be missing at random (MAR), because the probability of recording \( X \) would not depend on \( X \), after controlling for \( Y \), a variable which would be observed for all subjects. This case falls out of the scope of this paper as the missingness mechanism is ignorable for likelihood based inference [as long as (3) does not depend on the parameters of interest \( \theta \)]; see Rubin (1976) and Little and Rubin (1987). It is worth to remark, though, that most of the literature on nonresponse focus on data MAR [see, for example, Little and Rubin (1987) and Schafer (1997)] dealing mainly with procedures to impute the missing values. However, in econometrics, nonignorable nonresponse has been approached in the extensive literature on sample selection, pioneered by Heckman (1976), and also in some papers dealing with attrition in panel data [see, for example, Fitzgerald, Gottschalk and Moffitt (1998) and Hirano, Imbens, Ridder and Rubin (2001)].

2.3 Formulation of the missing data problems

Our reinterpretation of the missing data problems involves treating each group of respondents and nonrespondents in the population, associated with each discrete value of the variable of interest, as if it was a stratum. For both INRS and UNRS a further stratum, including a SRS of units for which only the covariates are measured, is added. Thus, we first consider \((C + 1)\) strata, each of which containing the respondent subjects for each value of \( Y \). The proportion of each of these strata in the sample and in the population is denoted by, respectively, \( H_y \) and \( Q_y \). Then,
we consider \((C + 1)\) additional strata containing the nonrespondent individuals for each response

\(Y\). Each of these strata has an unknown sampling proportion \(H^\text{nr}_y\) and the same population

proportion of \(Q_y\). Consequently, the initial random sample is interpreted as a combination of two

CB samples, one for the respondent and other for the nonrespondent units. Finally, the stratum

containing the SRS has a proportion of \(\Pr(S = 1) = H_S\) in the sample, while in the population,

as the supplementary sample is random, we observe units from this stratum with probability 1.

Probabilities \(H_y\) and \(H^\text{nr}_y\) are defined differently according to the use or not of a SRS. We

first examine the formulation which admits the use of a SRS because it encompasses the other as

a particular case. In this setup, \(H_y\) is the probability of observing a respondent unit reporting

\(Y = y\),

\[H_y = \Pr(Y = y, R = 1, S = 0), \quad (5)\]

while \(H^\text{nr}_y\) is the (unknown) proportion of nonrespondent individuals for which the variable of

interest is \(Y = y\):

\[H^\text{nr}_y = \Pr(Y = y, R = 0, S = 0). \quad (6)\]

Summing \(H_y\) and \(H^\text{nr}_y\) over \(Y\), we obtain the probability of observing, respectively, a respondent

unit, \(\Pr(R = 1, S = 0) = \sum_{y \in Y} H_y\), and a nonrespondent unit \(\Pr(R = 0, S = 0) = \sum_{y \in Y} H^\text{nr}_y\).

Summing these two last probabilities yields the proportion of the main sample in the global data

set which includes the main and the supplementary samples, \(\Pr(S = 0)\), that is

\[1 - H_S = \sum_{y \in Y} \sum_{r=0}^1 \Pr(Y = y, R, S = 0) = \sum_{y \in Y} H_y + \sum_{y \in Y} H^\text{nr}_y. \quad (7)\]

Additionally, due to the independence of the main and the supplementary samples, we may

write the marginal probability of \(Y = y\) in the population, \(Q_y = \Pr(Y = y)\), as

\[Q_y = \Pr(Y = y|S = 0) = \frac{\sum_{r=0}^1 \Pr(Y = y, R, S = 0)}{1 - H_S} = \frac{H_y + H^\text{nr}_y}{1 - H_S}. \quad (8)\]

This result will be used later on to avoid the estimation of the unknown probabilities \(H^\text{nr}_y\). Also,

as \(\Pr(Y = y, S = 0) = Q_y (1 - H_S)\), the conditional probability of response \(P_y\) of (3) is given by

\[P_y = \Pr(R = 1|Y = y, S = 0)\]
\[
\Pr(Y = y, R = 1, S = 0) \quad \Pr(Y = y, S = 0) \\
H_y = \frac{P_y}{Q_y (1 - H_S)}. \tag{9}
\]

From (9), it becomes obvious that, as \(0 < P_y < 1\), \(0 < H_y < Q_y (1 - H_S)\). Moreover, data MCAR are characterized by \(H_y\) fixed for all \(y\), because the rate \(P_y\) is fixed. Because in all cases \(H_y\) may be estimated from the incomplete sample as \(\frac{n_y}{N_m}\), equation (9) may be used to calculate the probability \(P_y\) when either \(Q_y\) is known or estimated by the methods proposed in section 3.

On the other hand, when the initial sample size \(N\) is not known under UNRS, the main adaptation of the setting just described involves the suppression of the \((C + 1)\) strata containing nonrespondents, since in the main sample we merely need to consider the respondent individuals. Hence, now \(H_y\) is defined as the sampling probability of observing \(Y = y\) in the main data set conditional on \(R = 1\):

\[
H_y = \Pr(Y = y, S = 0 | R = 1). \tag{10}
\]

Consequently, summing \(H_y\) over \(Y\) yields \(\sum_{y \in Y} H_y = \Pr(S = 0 | R = 1)\), which, due to the independence of the supplementary and the main sample, is simply \(\Pr(S = 0) = \sum_{y \in Y} H_y = 1 - H_S\).

Moreover, the probability \(Q_y\) no longer may be written in terms of \(H_y\) as in (8) and, instead of (9), the relation between \(P_y, H_y, H_S\) and \(Q_y\) is now given by

\[
H_y = \frac{\Pr(Y = y, R = 1, S = 0)}{\Pr(R = 1)} \\
= \frac{P_y Q_y (1 - H_S)}{\sum_{y \in Y} P_y Q_y}. \tag{11}
\]

From (11), it becomes clear that \(H_y\) no longer needs to be less than \(Q_y\) and \(H_y = Q_y (1 - H_S)\) for all \(y\) characterize both cases of absence of missing data and data MCAR. Also, we may conclude that, in this case, when \(H_y\) and \(Q_y\) are either known or estimated, we can not obtain the probabilities \(P_y\). However, for any two choices \(Y = y\) and \(Y = j\) we can estimate the ratio of the conditional response probabilities as \(\frac{P_j}{P_y} = \frac{H_j Q_y}{Q_j H_y}\), which takes the value 1 for all \(y\) when the data are MCAR.

When the SRS is unavailable or is not utilized, we merely deal with the main sample. Thus, all probabilities defined throughout this section are straightforwardly adapted by making \(H_S = 0\) and suppressing \(S = 0\) from the notation, as now all sampling units are associated to \(S = 0\). As these adjustments are very simple and they are also applicable to the likelihood functions to be defined in the next subsection, INR and UNR will only be addressed again in the estimation framework of section 3.
To conclude this subsection, it is worth to note that, alternatively, the nonresponse problems in analysis could also have been addressed by analogy with VPS, a sampling scheme referred to in section 1 which deliberately produces missing data. In this particular endogenous stratified sampling mechanism the subjects are kept in the sample according to a pre-defined probability of retention, chosen by the sampling agent. Hence, it seems appropriated to describe our missing data patterns by treating \( P_y \) as if it was the probability of retention, and taking them as additional parameters to estimate. However, we will not explore this avenue because the identification problems for UNRS (and UNR) arising when \( N \) is unknown, discussed bellow (11), would require a differentiated formulation relative to the remaining cases.

2.4 Observed data likelihood functions

This subsection derives the individual likelihood functions for the observed data under INRS and UNRS, as well as other sampling densities of interest which are important to characterize them. We first analyse INRS because the likelihood function of the observed data may be modified to obtain that for UNRS by eliminating the information on the covariates provided by nonrespondents. In fact, for both INRS and UNRS, the same data is observed for the respondent units and for the individuals of the SRS, respectively, \((Y, X, R = 1, S = 0) \) and \((X, S = 1) \). For the nonrespondents, we observe either \((X, R = 0, S = 0) \) for INRS or merely \((R = 0, S = 0) \) for UNRS.

Consider \( V = (Y, X, R) \). For INRS the likelihood function for a single unit of the observable data is

\[
\ell_{\text{INRS}}(v, s) = \left[ h(y, x, r = 1, s = 0)^r h(x, r = 0, s = 0)^{1-r} \right]^{1-s} h(x, s = 1)^s \\
= \left\{ \Pr(y, R = 1, S = 0) h(x|y)^r \left[ \sum_{y \in \mathcal{Y}} \Pr(y, R = 0, S = 0) h(x|y) \right]^{1-r} \right\}^{1-s} \left[ H_S f(x) \right]^s \\
= \left\{ \left[ \frac{H_y}{Q_y} \Pr(y|x, \theta) f(x) \right]^r \left[ \sum_{y \in \mathcal{Y}} \frac{Q_y (1 - H_S) - H_y}{Q_y} \Pr(y|x, \theta) f(x) \right]^{1-r} \right\}^{1-s} \left[ H_S f(x) \right]^s \\
= \left\{ \left[ \frac{H_y}{Q_y} \Pr(y|x, \theta) f(x) \right]^r \left\{ 1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta) f(x) \right\}^{1-r} \right\}^{1-s} \left[ H_S f(x) \right]^s \\
\tag{12}
\]

In the second line of (12) \( h(x|y, r, s = 0) = h(x|y) \) because, due to the independence of the main and the supplementary samples, \( h(x|y, r, s = 0) = h(x|y, r) \), which, due to the missing data
mechanism assumed in (3), is written as \( h(x|y, r) = h(x|y) \). Moreover, in the third line, the dependence on the unknown probabilities \( \Pr(y, R = 0, S = 0) = H^y_{y^r} \) was eliminated using the relation in (8).

In (12) the contribution of the units of the initial sample, associated with the indicator \((1 - S)\), is composed of two parts. The first term contains the information provided by the respondent units and may be interpreted as the complete data likelihood, while the second term accommodates the information on \( X \) provided by the nonrespondent units. The component indexed to \( S \) incorporates information on \( X \) provided by the individuals of the SRS. Notice that the data on \( X \) reported by nonrespondent and units of the SRS enter the likelihood function in a different way. It is clear that only the behavior of the former subjects, which are included in the main sample, is determined by the missing data mechanism assumed in section 2.2.

For UNRS, relative to INRS, as the nonrespondent units do not provide any information, we merely known the total amount of nonrespondents (because \( N \) is assumed known). Relative to (12), only the term associated with \((1 - R)\) needs to be modified,

\[
l_{UNRS}(v, s) = \left[ h(y, x, r = 1, s = 0) \right]^r \Pr(r = 0, s = 0)^{1-r} \left[ 1 - H_S - \sum_{y \in Y} \frac{H_y}{Q_y} \int_X \Pr(y|x, \theta) f(x) \, dx \right]^{1-s} h(x, s = 1)^s
\]

no longer being a function of \( X \). In fact, it merely incorporates information on the total sample size \( N \), which is employed in the estimation of \( H_y \) and \( H_S \).

So far, we have specified the joint density functions of the observed data. However, if some of the variables are ancillary for the parameters of interest \( \theta \), inference must be conditional on them. The analysis of the marginal sampling density function of the covariates under both INRS and UNRS reveals that the two missing data problems have a different nature. For INRS, this sampling density,

\[
h_{INRS}(x) = f(x) \sum_{s=0}^{1} \sum_{r=0}^{1} \left[ \sum_{y \in Y} \frac{H_y}{Q_y} \Pr(y|x, \theta) \right]^r \left[ 1 - H_S - \sum_{y \in Y} \frac{H_y}{Q_y} \Pr(y|x, \theta) \right]^{1-r} H^s_S
\]

\( = f(x) \frac{\sum_{y \in Y} \frac{H_y}{Q_y} \Pr(y|x, \theta) + 1 - H_S - \sum_{y \in Y} \frac{H_y}{Q_y} \Pr(y|x, \theta) + H_S}{H^s_S} \]

\( = f(x), \)

(14)
coincides with the population version \( f(x) \), not being a function of \( \theta \). Thus, inference must be conditional on \( X \), being based on

\[
I_{\text{INRS}}(y, r, s | x) = \left[ \frac{H_y}{Q_y} \Pr(y|x, \theta) \right]^r \left[ 1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta) \right]^{1-r} H_S^s. \tag{15}
\]

Conversely, for UNRS, the sampling density function of \( X \),

\[
h_{\text{UNRS}}(x) = \sum_{s=0}^{1} \sum_{r=0}^{1} \left\{ \left[ \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta) f(x) \right]^r \left[ 1 - H_S - \sum_{y \in \mathcal{Y}} H_y \right]^{1-r} \right\} [H_S f(x)]^s
\]

\[
= f(x) \left[ H_S + \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta) \right] + 1 - H_S - \sum_{y \in \mathcal{Y}} H_y, \tag{16}
\]
depends on \( \theta \). Thus, as \( X \) is not ancillary for \( \theta \), in general, the use of a likelihood function which is conditional on \( X \) yields inefficient estimators.\(^2\) So, the GMM estimators suggested later on are based on (13).

On the other hand, the joint probability of the indicators \((R, S)\) is identical for both INRS and UNRS. The calculations for INRS yield

\[
\Pr(R = r, S = s) = \left( \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} \frac{H_y}{Q_y} \Pr(y|x, \theta) f(x) \, dx \right)^r \left\{ \int_{\mathcal{X}} \left[ 1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta) \right] f(x) \, dx \right\}^{1-r} H_S^s \tag{17}
\]

from which it is clear that both \( R \) and \( S \) are ancillary for \( \theta \), since (17) is not a function of this parameter vector. Hence inference must be conditional on these two indicators. However, similarly to Imbens (1992) and Imbens and Lancaster (1996) in the endogenous stratified sampling framework, the GMM estimators we propose next, being based on likelihood functions which are not conditional on \( R \) and \( S \), given in (13) and (15), conform with the principle of conditionality. Basically, this is implemented by introducing the ancillary statistics \( \hat{H}_y = \frac{n_y}{N_m} \) and \( \hat{H}_S = \frac{m}{N_m} \) in the estimation procedure.

Similarly to section 2.3, the setting for UNRS may be adapted for when the total sample size \( N \) is unknown. Consider now that all the analysis concerning the main sample is conditional on

\(^2\) For a discussion on the issue of the ancillarity of the covariates in problems where data are MAR, see Lawless, Kalbfleisch and Wild (1999).
The likelihood of the observed data is

\[ l_{\text{UNRS}}(y, x, r = 1, s) = h(y, x, s) = 0 | r = 1)^{r(1-s)} h(x, s = 1) \]

in which, relative to (13), merely the indicator \( R \) and the terms associated with the indicator \((1 - R)\) were suppressed, because \( R = 1 \) for all individuals of the main sample. This likelihood function (and the simpler version for UNR where \( S = 0 \)), coincides with that for CB sampling with (without) a SRS; see, for example, Cosslett (1981a). Thus, inference procedures proposed for CB samples may be used. Note also that both (16) and (17) are simplified by eliminating \( 1 - H_S - \sum_{y \in Y} H_y \), as now \( 1 - H_S = \sum_{y \in Y} H_y \); see comments below (10).

Finally, it is important to notice that, in all missing data patterns, the component associated to \( R(1 - S) \), identified as the complete data likelihood, differs from the population joint density function of \((Y, X)\) in (1), the likelihood function used under RS. Hence, unless data are MCAR, in which case \( \frac{H_y}{Q_y} \) is a constant, being irrelevant for likelihood based inference, the procedures for RS may not be used with the complete sample.

### 3 Generalized method of moments estimation

This section adapts the method suggested by Imbens (1992) for efficient GMM estimation under CB sampling to handle data arising from the patterns of nonresponse considered. Basically, a set of moment indicators is derived for INRS and UNRS, which may be employed when either the marginal probabilities \( Q_y \) are unknown or known, the vector of parameters of interest being, respectively, \( \varphi = (H, \theta, Q) \) or \( \varphi_Q = (H, \theta) \), with \( Q = (Q_0, \ldots, Q_C) \) and \( H = (H_0, \ldots, H_C, H_S) \) [or \( H = (H_0, \ldots, H_C) \) under INR and UNR]. The parameter vector \( H \) is estimated together with the remaining parameters of interest in order to condition the analysis on the ancillary statistics \( \hat{H}_y = \frac{n_y}{N_m} \) and \( \hat{H}_S = \frac{m}{N_m} \) (or \( \hat{H}_y = \frac{n_y}{n_m} \) and \( \hat{H}_S = \frac{m}{n_m} \) for INR and UNR). We first deal with UNRS, case where the extension of Imbens (1992) method is more direct as, similarly to CB sampling, the analysis is not undertaken conditional on the covariates. Then we utilize some of the procedures involved in this method to obtain estimators for INRS, which are based on a likelihood function conditional on \( X \).

Our reinterpretation of the incomplete data problems in discrete choice models based on the CB sampling setting suggests that some of the estimators originally proposed for this setup may be used to handle some of these missing values patterns. Namely, all of these estimators may be
used to deal with UNR when $N$ is unknown. In fact, later on we demonstrate that our estimators, when simplified to deal with this case, coincide with those proposed by Imbens (1992). Similarly, Cosslett’s (1981a) ML estimators for CB samples combined with a SRS of covariates, may be employed to describe UNRS ignoring the information on the size of the initial sample. However, in the same sense that Imbens (1992) simplified Cosslett’s (1981a,b) estimators for CB samples, we derive GMM estimators for UNRS which are much simpler than those of Cosslett (1981a). Furthermore, our estimators embed Lancaster and Imbens’ (1996) efficient GMM estimators for case-control binary models with contaminated controls, where there are two strata, one consisting of a random sample where only the covariates are observable, the other including units choosing 1.3

On the other hand, a likelihood function suggested by Hausman and Wise (1981) for a VPS scheme where the covariates are observed for all individuals and the variable of interest is measured according to the probability of retention associated with each stratum, could be used to describe INR. In effect, one merely has to define one stratum for each value of $Y$ and take the $(C + 1)$ probabilities of retention (which are known under VPS but are unknown in our missing data problems) as additional parameters to estimate. However, as, for reasons discussed previously (see section 2.3), we did not consider the VPS framework, we propose GMM estimators for INR which are derived by the same method of all the other nonresponse problems.

The remainder of this section is organized as follows. Subsection 3.1 derives the moment indicators for INRS and UNRS. These moment indicators are used in subsection 3.2 to obtain ten alternative GMM estimators which are appropriate to handle all the missing data patterns considered in this paper. Subsection 3.3 presents a brief analysis and comparison of these estimators. Finally, subsection 3.4 discusses particular estimation issues which arise in multiplicative intercept models (MIM).

### 3.1 Derivation of the moment indicators

This subsection first derives the moment indicators suitable for the missing data pattern induced by UNRS. Then, using a similar method, the moment indicators appropriate for INRS are obtained.

Under UNRS, in order to avoid the specification of the marginal distribution of $X$, let $f(x)$ be discrete with $L$ points of support $x^l$, $l = 1, 2, ..., L$, and associated probability mass parameters

3Note that in this case, $Y = \{0, 1\}$, $P_{y=0} = 0$ and $P_{y=1} = 1$. Hence, the assumption of $0 < P_y < 1$ is relaxed to $0 \leq P_y \leq 1$. 

14
\[ \text{Pr}\left(X = x^l\right) = \pi_l, \pi_l > 0, \ l = 1, 2, ..., L, \ L > J \text{ with } J \text{ being the number of strata considered.} \]

The resultant log-likelihood function based on (13) is

\[ L(H, \theta, \pi) = \sum_{i=1}^{N_m} (1 - s_i) r_i \left[ \ln H_{y_i} + \ln \text{Pr} \left( y_i | x^{l_i}, \theta \right) - \ln \sum_{l=1}^{L} \pi_l \text{Pr} \left( y_i | x^l, \theta \right) + \ln \pi_{l_i} \right] + \]

\[ \sum_{i=1}^{N_m} (1 - s_i) (1 - r_i) \ln \left( 1 - HS_i - \sum_{y \in \mathcal{Y}} H_y \right) + \]

\[ \sum_{i=1}^{N_m} s_i \left( \ln HS_i + \ln \pi_{l_i} \right). \tag{19} \]

The maximization of (19) is undertaken with respect to the vector of parameters \((H, \theta, \pi)\), subject to the restriction \(\sum_{l=1}^{L} \pi_l = 1\), the resultant first order conditions being

\[ s \left( \hat{H}, \hat{\theta}, \hat{\pi} \right)_{H_t} = \sum_{i=1}^{N_m} (1 - s_i) \left[ \frac{r_i I(y_i = l)}{H_t} - \frac{(1 - r_i)}{1 - HS - \sum_{y \in \mathcal{Y}} H_y} \right] = 0 \tag{20} \]

\[ s \left( \hat{H}, \hat{\theta}, \hat{\pi} \right)_{H_S} = \sum_{i=1}^{N_m} \left[ \frac{s_i I(s_i = 1)}{H_S} - \frac{(1 - s_i)(1 - r_i) I(s_i = 1)}{1 - HS - \sum_{y \in \mathcal{Y}} H_y} \right] = 0 \tag{21} \]

\[ s \left( \hat{H}, \hat{\theta}, \hat{\pi} \right)_{\pi_z} = \sum_{i=1}^{N_m} (1 - s_i) r_i \left[ \frac{I(l_i = z)}{\hat{\pi}_z} - \frac{\text{Pr} \left( y_i | x^{z_i}, \hat{\theta} \right)}{\sum_{z=1}^{Z} \hat{\pi}_z \text{Pr} \left( y_i | x^l, \hat{\theta} \right)} \right] + \frac{s_i I(l_i = z)}{\hat{\pi}_z} \tag{22} \]

\[ s \left( \hat{H}, \hat{\theta}, \hat{\pi} \right)_{\mu} = \sum_{l=1}^{L} \hat{\pi}_l - 1 = 0, \tag{24} \]

where \( t = 0, 1, ..., C, \mu \) is the Lagrange multiplier associated with the restriction \(\sum_{l=1}^{L} \pi_l = 1\) and \(z = 1, ..., L\).

Performing operations similar to those in Imbens (1992), which use the fact that the ML estimator for \(Q_y\) can be written from (2) as

\[ \hat{Q}_y = \sum_{l=1}^{L} \hat{\pi}_l \text{Pr} \left( y | x^l, \hat{\theta} \right), \tag{25} \]

the dependence on the discrete distribution assumed for \(X\) is removed from the system (20)-(24). We first multiply all the terms in (23) by \(\hat{\pi}_z\) and sum over \(z\), to obtain

\[ \hat{\mu} = \sum_{i=1}^{N_m} (1 - s_i) r_i \left[ \frac{\sum_{z=1}^{Z} \hat{\pi}_z I(l_i = z)}{\hat{\pi}_z} - \frac{\sum_{z=1}^{Z} \hat{\pi}_z \text{Pr} \left( y_i | x^z, \hat{\theta} \right)}{\sum_{l=1}^{L} \hat{\pi}_l \text{Pr} \left( y_i | x^l, \hat{\theta} \right)} \right] \]

\[ + \sum_{i=1}^{N_m} s_i \frac{\sum_{z=1}^{Z} \hat{\pi}_z I(l_i = z)}{\hat{\pi}_z} \]

\[ = m \]

\[ = N_m H_S. \]
which, replaced in (23), yields

\[ \hat{\pi}_z = \frac{1}{N_m} \sum_{i=1}^{N_m} [(1 - s_i) r_i + s_i] I_{(i,z)} \left[ \hat{H}_S + \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{(1 - s_i) r_i}{Q_{y_i}} \Pr \left( y_i | x^z, \hat{\theta} \right) \right]^{-1} \]

\[ \approx \frac{1}{N_m} \sum_{i=1}^{N_m} [(1 - s_i) r_i + s_i] I_{(i,z)} \left[ \hat{H}_S + \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^z, \hat{\theta} \right) \right]^{-1}. \quad (26) \]

This estimator for \( \hat{\pi}_z \) reflects the distortion of \( h_{UNRS}(x) \) in (16) relative to \( f(x) \), induced by the nonresponse pattern. If the main sample had no missing values, we would have \( \hat{\pi}_z = \frac{1}{N_m} \sum_{i=1}^{N_m} I_{(i,z)} \), the usual nonparametric ML estimator for a mass point at each of the \( L \) points of support [see, for example, Coslett (1997)], since we would have \( R = 1 \) for all units of the main sample and \( \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^z, \hat{\theta} \right) = \frac{H_y}{Q_y} \sum_{y \in Y} \Pr \left( y | x^z, \hat{\theta} \right) = \left( 1 - \hat{H}_S \right) \sum_{y \in Y} \Pr \left( y | x^z, \hat{\theta} \right) = \left( 1 - \hat{H}_S \right)^4. \]

Then, \( \hat{\pi}_z \) is substituted in both the last term of (22),

\[ \sum_{i=1}^{N_m} (1 - s_i) r_i \frac{\sum_{l=1}^{L} \hat{\pi}_l \nabla \Pr \left( y_i | x^l, \hat{\theta} \right)}{\sum_{l=1}^{L} \hat{\pi}_l \Pr \left( y_i | x^l, \hat{\theta} \right)} = \sum_{i=1}^{N_m} (1 - s_i) r_i \frac{1}{N_m} \sum_{j=1}^{N_m} [(1 - s_j) r_j + s_j] I_{(j,l)} \left[ \hat{H}_S + \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^l, \hat{\theta} \right) \right]^{-1} \nabla \Pr \left( y_i | x^l, \hat{\theta} \right) \]

\[ = \sum_{j=1}^{N_m} [(1 - s_j) r_j + s_j] \left[ \hat{H}_S + \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^l, \hat{\theta} \right) \right]^{-1} \frac{1}{N_m} \sum_{i=1}^{N_m} (1 - s_i) r_i \nabla \Pr \left( y_i | x^l, \hat{\theta} \right) \]

\[ = \sum_{i=1}^{N_m} [(1 - s_i) r_i + s_i] \left[ \hat{H}_S + \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^l, \hat{\theta} \right) \right]^{-1} \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \nabla \Pr \left( y_i | x^l, \hat{\theta} \right). \quad (27) \]

and in \( \hat{Q}_y \) of (25),

\[ \hat{Q}_y = \frac{1}{N_m} \sum_{i=1}^{N_m} [(1 - s_i) r_i + s_i] I_{(i,l)} \left[ \hat{H}_S + \sum_{y \in Y} \frac{\hat{H}_y}{Q_y} \Pr \left( y | x^l, \hat{\theta} \right) \right]^{-1} \Pr \left( y_i | x^l, \hat{\theta} \right) \]

\[ = \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{[(1 - s_i) r_i + s_i] \Pr \left( y_i | x^l, \hat{\theta} \right)}{\hat{H}_S + \sum_{y \in Y} \frac{H_y}{Q_y} \Pr \left( y | x^l, \hat{\theta} \right)}. \quad (28) \]

\footnote{This derivation uses the fact that, with no missing data, the ratio \( \frac{\hat{H}_y}{Q_y} \) is a constant across all \( Y \) and uses equation (9) with \( P_y = 1 \).}
Equation (28) is used to obtain estimating functions for \( Q \), which are necessary because the term (27), included in the estimating function for \( \theta \) is a function of these probabilities.

The resulting system of estimating functions is then composed of

\[
g(v, s, \varphi)_H = (1 - s) rI_{(y=t)} - H_t \tag{29}
\]

\[
g(v, s, \varphi)_S = s - H_S \tag{30}
\]

\[
g(v, s, \varphi)_\theta = (1 - s) r \nabla \theta \ln \Pr(y|x, \theta) - \frac{[(1 - s) r + s] \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta)}{H_S + \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta)} \tag{31}
\]

\[
g(v, s, \varphi)_{Q_t} = Q_t - \frac{[(1 - s) r + s] \Pr(t|x, \theta)}{H_S + \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y|x, \theta)}, \tag{32}
\]

which are used as moment indicators in the GMM framework.

Under INRS, due to the ancillarity of \( X \) for \( \theta \), inference must be conducted conditional on \( X \). Even so estimation may proceed in a very similar way to that under UNRS; recall that in section 2 we showed that the formulation of both nonresponse patterns is very similar and in both cases the indicators \( R \) and \( S \) are ancillary for \( \theta \).

In order to condition the analysis on \( \hat{H}_y = \frac{n_y}{N_m} \) and \( \hat{H}_S = \frac{m}{N_m} \), we consider again the estimation of both \( H_y \) and \( H_S \) using, respectively, (29) and (30). On the other hand, the estimating functions for \( \theta \) are derived from the log-likelihood function based on (15),

\[
L(H, \theta, Q) = \sum_{i=1}^{N_m} (1 - s_i) r_i [\ln H_{y_i} + \ln \Pr(y_i|x, \theta) - \ln Q_{y_i}] + \\
\sum_{i=1}^{N_m} (1 - s_i) (1 - r_i) \ln \left[1 - H_{S_i} - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y_i|x, \theta)\right] + \sum_{i=1}^{N_m} s_i \ln H_{S_i}. \tag{33}
\]

As the estimating equation for \( \theta \),

\[
s(\hat{H}, \hat{\theta}, \hat{Q}) = \sum_{i=1}^{N_m} (1 - s_i) \left[r_i \nabla \theta \ln \Pr(y_i|x, \hat{\theta}) - (1 - r_i) \frac{\sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y_i|x, \hat{\theta})}{1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \Pr(y_i|x, \hat{\theta})}\right] = 0, \tag{34}
\]

depends on \( Q \), we use a similar approach to that employed under UNRS to obtain estimating functions for these probabilities. Again, we consider the definition of the ML estimator for \( Q_y \) given in (25) and assume that \( f(x) \) be discrete with \( L \) points of support \( x^l, l = 1, 2, ..., L \), and associated probability mass parameters \( \Pr(X = x^l) = \pi_l, \pi_l > 0, L > J, l = 1, 2, ..., L \). To remove the dependence of (25) on \( \pi_l \) we need to obtain a nonparametric ML estimator for each of the \( L \) mass points of \( f(x) \). This estimator can not be derived from the first order condition for \( \pi_z \) as before, because (33) is conditional on \( X \). However, as now \( h_{INRS}(x) \) in (14) coincides with \( f(x) \)
because $X$ is observed for all individuals, $\hat{\pi}_z = \frac{1}{N_m} \sum_{i=1}^{N_m} I(l_i = z)$; see also comments below (26).

The resulting ML estimator for $Q_y$ is

$$
\hat{Q}_y = \frac{1}{N_m} \sum_{i=1}^{N} \Pr \left( y_i | x_i, \theta \right). \tag{35}
$$

Thus, we propose the estimation of $\varphi$ and $\varphi_Q$ by GMM using the moment indicators:

$$
g(v, s, \varphi)_{H_t} = (1 - s) r I(y = t) - H_t \tag{36}
$$

$$
g(v, s, \varphi)_{H_S} = s - H_S \tag{37}
$$

$$
g(v, s, \varphi)_y = (1 - s) \left[ r \nabla_\theta \ln \Pr(y | x, \theta) - (1 - r) \frac{\sum_{y \in \mathcal{Y}} \frac{H_y}{\partial y} \nabla_\theta \Pr(y | x, \theta)}{1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{\partial y} \Pr(y | x, \theta)} \right] \tag{38}
$$

$$
g(v, s, \varphi)_{Q_t} = Q_t - \Pr(t | x, \theta). \tag{39}
$$

### 3.2 Alternative estimators

In the GMM framework the objective function to be minimized is

$$
\Upsilon_{N_m}(\varphi) = g_{N_m}(v, s, \varphi)' W_{N_m} g_{N_m}(v, s, \varphi), \tag{40}
$$

where $g_{N_m}(v, s, \varphi) = \frac{1}{N_m} \sum_{i=1}^{N_m} g(v_i, s_i, \varphi)$ is the sample counterpart of the moment conditions $E[g(v, s, \varphi)] = 0$, with the moment indicators $g(v, s, \varphi)$ given in (29)-(32) and (36)-(39) and $E[.]$ denoting expectation taken over $l_{UNRS}(v, s)$ of (13) and $l_{INRS}(y, r, s|x)$ of (15) for, respectively, UNRS and INRS, and $W_{N_m}$ is a positive semi-definite weighting matrix.

When the marginal choice probability in the population $Q_y$ is unknown, the estimation of $\varphi$ is a just-identified problem. Under the usual regularity conditions required for GMM, see Newey and McFadden (1994, Theorems 2.6, 3.4), the resulting GMM estimator, $\hat{\varphi}$, is consistent for the true value $\varphi^0$ and satisfies

$$
\sqrt{N_m} \left( \hat{\varphi} - \varphi^0 \right) \xrightarrow{d} N \left( 0, G^{-1} \Omega G^{-1} \right), \tag{41}
$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\Omega = E \left[ g(v, s, \varphi) g(v, s, \varphi)' \right]$ and $G = E \left[ \nabla_\varphi g(v, s, \varphi)' \right]$. Under UNRS, when $X$ is discrete, $\hat{\varphi}$ is a ML estimator, being, thus, efficient. Otherwise, asymptotic efficiency, in the semiparametric sense, can be proved by an analogous demonstration to that of Imbens (1992, Theorem 3.3); see the appendix.

On the other hand, if $Q_y$ is known, this probability is replaced in (29)-(32) or (36)-(39) and an overidentified GMM estimator for $\varphi_Q$ is obtained. Assuming that the regularity conditions in
Newey and McFadden (1994, Theorems 2.6, 3.4) hold, the optimal estimator, \( \hat{\varphi}_Q \), obtained from using the weighting matrix \( W_{N_m} = \Omega_{N_m}^{-1} \) in the objective function (40), where \( \Omega_{N_m} \) is a consistent estimator of \( \Omega \), is consistent for \( \varphi_Q^0 \) and satisfies

\[
\sqrt{N_m} \left( \hat{\varphi}_Q - \varphi_Q^0 \right) \xrightarrow{d} N \left[ 0, \left( G'\Omega^{-1}G \right)^{-1} \right],
\] (42)

with \( \Omega \) and \( G \) defined below (41) with the obvious adaptations resulting from the replacement of \( \varphi \) by \( \varphi_Q \). Under UNRS, similarly to \( \hat{\varphi} \), asymptotic efficiency can be proved as in Imbens (1992, Theorem 3.3); see the appendix.

Alternative estimators for \( \varphi \) and for \( \varphi_Q \) may be obtained under UNRS for cases where \( N \) is unknown. Considering the general setup for UNRS, first, in all derivations and results \( N_m \) is replaced by \( n_m \). Consequently, \( H_y \) is estimated by \( \frac{n_y}{n_m} \) and the estimating function for \( H_S \) of (30) is suppressed, because now \( H_S \) may be estimated by \( 1 - \sum_{y \in Y} \hat{H}_y \); see comments below (10). Second, the indicator \( R \) is replaced by 1 in (29), (31) and (32). This allows all observations to enter the calculation of the second terms of both (31) and (32), because now \( X \) is available for all the units in the sample, since we do not consider the strata containing nonrespondents. Here the expectations are taken over \( l_{UNRS} (y, x, r = 1, s) \) of (18).

Finally, estimators for both \( \varphi \) and \( \varphi_Q \) under UNR (\( N \) known or unknown) and UNR, may be straightforwardly obtained from the respective versions which consider a SRS. In all derivations and results \( N_m \) and \( n_m \) are replaced by, respectively, \( N \) and \( n \). In the moment indicators we make \( H_S = 0 \) and \( S = 0 \) and suppress the estimating functions for \( H_S \) in (30) and (37). Naturally, expectations are now taken with respect to the density functions referred to before, but with \( H_S = 0 \) and \( S = 0 \). The estimators derived for UNR with \( N \) unknown coincide with those suggested by Imbens (1992) for CB sampling.

### 3.3 Joint analysis of the different estimators

Throughout this section we employed a similar methodology to derive GMM estimators for \( \varphi \) and \( \varphi_Q \) under several missing data patterns. The estimators obtained use different information, ranging from the case where only the respondents are observed, UNR, to that where we have data on respondents, nonrespondents and a SRS, INRS, including the intermediate cases where, besides the data on respondents, we use information on \( X \) provided by nonrespondents (INR) or by units of the SRS (UNRS). The analysis of the respective systems of moment indicators in (29)-(32) and (36)-(39) (and their simplified forms for UNR and INR), allows the examination
of both the common characteristics of the different estimators and the mechanisms by which the information on $X$ is incorporated in the estimation procedure.

The estimating functions for $H_y$ and $H_S$, $g(v, s, \varphi)_{H_t}$ and $g(v, s, \varphi)_{H_S}$, are identical in all cases and are used to condition the analysis on the ancillary statistics $\hat{H}_y$ and $\hat{H}_S$. The moment indicators for $\theta$, $g(v, s, \varphi)_{\theta}$, have two components. The first term is, in all cases, the score function of the RS ML (RSML) estimator for $\theta$ and, being a function of $(Y, X)$, it is only calculated for the respondent units. The remaining term merely depends on $X$, being calculated for respondents and units of the SRS in UNRS, for nonrespondents in INRS and INR and for respondents in UNR. Finally, the estimating functions for $Q_y$, $g(v, s, \varphi)_{Q_t}$, use information from the same units as the second term of $g(v, s, \varphi)_{\theta}$ under UNRS and UNR, while for INRS and INR data from all units observed (respectively, those of both the main and the SRS and those of the main sample) are used. It becomes, thus, apparent that, relative to UNR, the estimators for INR and UNRS include further information on $X$ through the second terms of $g(v, s, \varphi)_{\theta}$ and $g(v, s, \varphi)_{Q_t}$. In the case where more data is available, INRS, relative to INR, the supplementary data set is only used in the estimation of $Q_y$.

When data are MCAR, $P_y$ is constant for all $y$, as well as the ratio $\frac{H_y}{Q_y} = c$; see the comments below (9). This yields $\sum_{y \in Y} \frac{H_y}{Q_y} \nabla_{\theta} \Pr(y|x, \theta) = c \sum_{y \in Y} \nabla_{\theta} \Pr(y|x, \theta) = 0$ and the consequent suppression of the second term of $g(v, s, \varphi)_{\theta}$, which then becomes identical to the score function of the RSML estimator. Additionally, in $g(v, s, \varphi)_{Q_t}$, $\sum_{y \in Y} \frac{H_y}{Q_y} \Pr(y|x, \theta) = c$. As $g(v, s, \varphi)_{H_t}$, $g(v, s, \varphi)_{\theta}$ and $g(v, s, \varphi)_{Q_t}$ become only function of, respectively, $H$, $\theta$, and $\varphi$, we are in presence of a two steps GMM estimator: $\hat{H}$ and $\hat{\theta}$ obtained in the first step are then used to estimate $\hat{Q}$ in the second step. In these conditions, the calculations required to obtain the asymptotic variance of $\theta$ are identical to those employed with the RSML estimator, with the difference that the sample size in the last estimator is $n$. Thus, with the exception of UNR with $N$ unknown, in which case our GMM estimators use a sample of size $n$, in all the other cases the asymptotic variance of $\theta$ is improved relative to RSML estimators, since the sample size is larger.5

In general, it is clear that, unless data are MCAR, RSML estimators applied the complete sample are biased. The problem in a practical situation is that we do not known if the missingness mechanism which governs our data is ignorable or not, unless there exists information on the probabilities $Q_y$, in which case we may compare them with the sampling proportions $H_y$ and roughly conclude about the nature of the missing data. Thus, the proposal of specification tests

---

5Note that, these improvements are merely due to the increase of the sample size, because the information on $X$ provided by nonrespondents or units of the SRS is not used.
for the null hypotheses of data MCAR, that is $H_0 : H_y - cQ_y = 0$ for all $y$ or $H_0 : H_y - Q_y = 0$ for all $y$, for cases where $N$ is, respectively, known or unknown, appears an important major contribution in this area to be developed in future research.

### 3.4 The particular case of multiplicative intercept models

So far we proposed estimators to deal with different patterns of nonresponse governed by a non-ignorable missing data mechanism. We concluded that, unless data are MCAR, the conventional RS estimators applied with the complete data set are inconsistent. However, Carroll, Ruppert and Stefanski (1995) p. 184 and Allison (2001), p. 7, refer that, as long as the probability of response conditional on the variable of interest is independent of the covariates, precisely the missingness mechanism we assume, the estimators of the slope parameters of logit models are still consistent. Given that in this paper we propose discrete choice models appropriate for dealing with incomplete samples which are a modification of the usual formulation for CB sampling, we interpret that conjecture as a result of the particular properties of MIM, which include the logit model as a particular case, so widely discussed in the area of CB sampling. On the one hand, both the intercept terms and the marginal choice probabilities $Q_y$ are not separately identified when these last probabilities are unknown. On the other hand, except for the shift in intercept terms, all the parameters are consistently estimated by the RSML estimator, for which the individual likelihood function is $Pr(y|x, \theta)$; see, for example, Hsieh, Manski and McFadden (1985) and Weinberg and Wacholder (1993).

Basically, we found that the model for UNR kept the two characteristics referred to above, since only a slight modification over the CB sampling formulation was required. However, neither of those two properties can be extended for none of the other cases, unless the incomplete units are discarded, yielding a UNR case.

Define the MIM as in Hsieh, Manski and McFadden (1985),

$$Pr(y|x, \nu_y, \theta_{1y}) = \frac{\nu_y V_y(\theta_{1y})}{\sum_{y \in Y} \nu_y V_y(\theta_{1y})},$$

where $\theta_{0y}$ and $\theta_{1y}$ are, respectively, the constant term and the vector of slope parameters associated with alternative $Y$, $\nu_y = f(\theta_{0y})$, $\nu_0 = 1$, $V_0(\theta_{1y}) = 1$, $V_y(\theta_{1y}) > 0$, $\nabla_{\theta_{0y}} V_y(\theta_{0y}) = \nu_y(\theta_{0y})$ and $\nabla_{\theta_{1y}} V_y(\theta_{1y}) = x'_y V_y(\theta_{1y})$ for all $y$.

As in Imbens’ (1992) method for CB sampling, under UNR the identification problems for intercept terms of MIM become apparent because the moment

---

6 The multinomial logit model arises when $\nu_y = e^{\theta_{0y}}$, $V_y(\theta_{1y}) = e^{x'\theta_{1y}}$, and $\theta_{0y=0} = \theta_{1y=0} = 0$. 21
indicators for the intercept parameters are perfectly correlated with the ones that allow the estimation of $Q_y$. In fact, writing the moment indicators (31) for a given $\theta_0$, $t = 0, 1, \ldots, C$,

$$g(v, s, \varphi)_{\theta_0} = r \left[ I_{(y=t)} - \frac{H_{Q_t} \nu_t V_t(\theta_1)}{\sum_{y \in Y} H_{Q_y} \nu_y V_y(\theta_1)} \right],$$  

(44)
it becomes obvious that the result coincides with $g(v, s, \varphi)_{H_t}$ in (32) when multiplied by $\frac{H_{Q_t}}{Q_t}$ and summed to $g(v, s, \varphi)_{H_t}$ in (29). Thus, the identification of $\theta_0$ is only possible when $Q_t$ is known, in which case $Q_t$ is replaced in the moment indicators $g(v, s, \varphi)_{\theta_0}$ and $g(v, s, \varphi)_{\theta_1}$, and $g(v, s, \varphi)_{Q_t}$ is suppressed.

The particular property of MIM which causes these identification problems allows the employment of RS procedures to estimate the slope parameters $\theta_1$ under UNR. This becomes apparent by noticing that the moment indicators for a given $\theta_1$, for UNR, given by (44) pre-multiplied by $x'_t$, apart from the distortion in the constant terms, which now are given by $\frac{H_{Q_t}}{Q_t} \nu_t$, coincide with the moment indicators employed with RS,

$$g(v, s, \varphi)_{\theta_1}^{rs} = x'_t \left[ I_{(y=t)} - \text{Pr}(t|x, \nu_y, \theta_1) \right].$$  

(45)

Thus, we may simply use (45) to consistently estimate $\theta_1$. This property does not hold in cases where the data on $X$ provided by nonrespondents and/or units of the SRS are used (UNRS, INRS and INR) as none of the moment indicators for $\theta$ under those missing data patterns may be written in the RS form expressed in (45). Thus, although the RSML estimator may be used in all cases with the complete data set for consistent estimation of the slope parameters, if one wishes to include that information on $X$ in the estimation procedure, the GMM estimators for UNRS, INRS or INR proposed in previous section must be used.

4 A Monte Carlo simulation study

In this section we undertake a small Monte Carlo simulation study to investigate the performance in practice of some of the estimators developed previously in binary probit models. Subsection 4.1 describes the design of the experiments and subsection 4.2 examines the results.

4.1 Experimental design

In all experiments we handle binary data with $Y = \{0, 1\}$. The variable of interest, $Y$, conditional on the scalar $X = x$, is assumed to be generated as a probit model characterized by $\text{Pr}(1|x, \theta) =$
Table 1: Experimental designs describing the missing data patterns

<table>
<thead>
<tr>
<th>Experiment designation</th>
<th>P_a</th>
<th>P_0</th>
<th>P^*</th>
<th>H_S</th>
<th>H_1</th>
<th>H_0</th>
<th>n_1</th>
<th>n_0</th>
<th>n</th>
<th>N−n</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.750</td>
<td>0.250</td>
<td>0.225</td>
<td>75</td>
<td>75</td>
<td>300</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0.760</td>
<td>0.920</td>
<td>0.826</td>
<td>0.167</td>
<td>0.475</td>
<td>0.192</td>
<td>171</td>
<td>69</td>
<td>240</td>
<td>60</td>
</tr>
<tr>
<td>c</td>
<td>0.227</td>
<td>0.920</td>
<td>0.247</td>
<td>0.375</td>
<td>0.106</td>
<td>0.144</td>
<td>51</td>
<td>69</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>d</td>
<td>0.373</td>
<td>0.480</td>
<td>0.778</td>
<td>0.375</td>
<td>0.175</td>
<td>0.075</td>
<td>84</td>
<td>36</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>e</td>
<td>0.920</td>
<td>0.920</td>
<td>1.000</td>
<td>0.074</td>
<td>0.639</td>
<td>0.213</td>
<td>207</td>
<td>69</td>
<td>276</td>
<td>24</td>
</tr>
</tbody>
</table>

\(\Phi(x\theta)\). In order to produce a probability \(Q = 0.75\) of observing \(Y = 1\) in the population, we set \(\theta = 0.251\) and \(X\) was generated as a mixture of normal distributions with mean 3 and variance 4, where the variate is \(N(2, 1.2915)\) with probability 0.7 and \(N(5.333, 1.2915)\) with probability 0.3.

In all the examples the total sample size is \(N = 300\). We analyse five experimental designs, characterized by different combinations of conditional response probabilities for \(Y = 1\) and \(Y = 0\), respectively denoted \(P_a\) and \(P_0\), which produce different ratios \(P^* = \frac{P_a}{P_0}\). The size of the SRS is \(m = (N − n)\), so that the improvements due to combining information on \(X\) from this independent random sample (under UNRS) or from the same number of nonrespondents from the initial sample (under INR) may be compared. Different combinations of \(P_a\) and \(P_0\) produced different proportions for individuals responding \(Y = 1\) \((Y = 0)\) in the main sample and for the SRS, denoted, respectively, \(H_1\) \((H_0)\) and \(H_S\).\(^7\) Table 1 summarizes the main characteristics of the five experimental designs. For comparison purposes, in the first experimental design we consider, design \(a\), there are no missing values. Then, from experiment \(b\) to \(c\), the number of incomplete responses \((N − n)\) is increased, as well as the differential between \(P_a\) and \(P_0\). In design \(d\) we considered a relatively large ratio \(P^*\) (a little smaller than that of \(b\)) associated with a small complete sample size \((n = 120\) as in design \(c\)), in order to distinguish the effects of varying \(P^*\) and \(n\). Finally, experiment \(e\) assumes that the data are MCAR, that is \(P^* = 1\). All computations were done using the package S-Plus. Each experiment was based on 1000 replications.

Eight estimators were compared. Four assume the nonexistence of information on \(Q\), including the common estimator for RS which merely uses the complete data set, designated RSMLE, and the estimators proposed in this paper for UNR, UNRS, and INR, denoted, respectively, UNRE, UNRSE, and INRE. The remaining four estimators incorporate information on \(Q\), being denoted as QRSML E, QUNRE, QUNRSE, and QINRE. The moment indicators for UNRS and INRS are written from, respectively, \((29)-(32)\) and \((36)-(39)\) and are described in Table 2. The moment indicators for UNR and INR used the simplifications described in section 3.2 and the moment indicator for \(\theta\) for RSMLE is obtained from \(g(v, s_\varphi)_{\theta}\) in UNR with the second term suppressed.

\(^7\)Note that, from equation \((9)\), \(H_1 = P_aQ(1 − H_S)\) and \(H_0 = P_0(1 − Q)(1 − H_S)\), with \(H_S = \frac{mn}{N}\).
Table 2: Binary models: individual moment indicators

<table>
<thead>
<tr>
<th>M.I.</th>
<th>UNRSE</th>
<th>INRSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g(v, r, \varphi)_{H_1}$</td>
<td>$(1-s)r y - H_1$</td>
<td>$(1-s)r y - H_1$</td>
</tr>
<tr>
<td>$g(v, r, \varphi)_{H_0}$</td>
<td>$(1-s)r (1-y) - H_0$</td>
<td>$(1-s)r (1-y) - H_0$</td>
</tr>
<tr>
<td>$g(v, r, \varphi)_{H_S}$</td>
<td>$s - H_S$</td>
<td>$s - H_S$</td>
</tr>
<tr>
<td>$g(v, r, \varphi)_{Q}$</td>
<td>$Q - \frac{(1-s)s + rP}{H_S + B} R$</td>
<td>$Q - \frac{(1-s)s + rP}{H_S + B} R$</td>
</tr>
</tbody>
</table>

$P = \Pr (1|x, \theta), p = \nabla_{x\theta} \Pr (1|x, \theta), R = \frac{H_0}{Q} - \frac{H_0}{1-Q},$ and $B = \frac{H_0}{1-Q} + RP$

4.2 Results

The summary statistics for probit models are presented in Table 3, which provides the mean and the median bias in percentage terms and the standard deviation across the replications for the estimates of $\theta$. Figures 1 and 2 show the estimated sampling distributions of the estimates of $\theta$ for designs b, c and d. The first figure considers the four estimators in which information on $Q$ is not used (RSMLE, UNRE, UNRSE and INRE) and QUNRE which, among the estimators which use the known value of $Q$, gave the worst performance. The second figure illustrates the behaviour of RSMLE compared with that of the four estimators where $Q$ is known, QRSMLE, QUNRE, QUNRSE and QINRE.

![Figure 1 about here](image1)

![Figure 2 about here](image2)

As expected, RSMLE perform well in designs a), where there are no missing values, and e), where data is MCAR. In these experiments the incorporation of aggregate information on $Q$ reduces the standard deviations across replications in more than 50%. However, in designs b, c and d, where the probability of response differs for $Y = 1$ and $Y = 0$, RSMLE suffer from large mean and median biases. These biases are less significant in designs b and d, which are characterized by ratios $P^*$ close to 1, being, thus, closer to a MCAR pattern, but, even in these cases, the biases achieve unacceptable figures of more than 10%. In these three experiments, the combination of information on $Q$ produces substantial improvements. Except for design c, where $P^*$ is very small, QRSMLE present small mean and median distortions, which sometimes are smaller than those presented by our modified GMM estimators, and a variability similar to that of the modified estimators for $Q$ known; see also Figure 2. These results are specially relevant. Imbens and Lancaster (1994) proposed the combination of macro and micro information with
<table>
<thead>
<tr>
<th>Experiment</th>
<th>Estimator</th>
<th>Bias Mean</th>
<th>Bias Median</th>
<th>St. D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>RSMLE</td>
<td>.008</td>
<td>.005</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>.015</td>
<td>.012</td>
<td>.011</td>
</tr>
<tr>
<td>b</td>
<td>RSMLE</td>
<td>.015</td>
<td>-.105</td>
<td>.110</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>.012</td>
<td>.015</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.010</td>
<td>.011</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.022</td>
<td>.017</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>UNRE</td>
<td>.024</td>
<td>.013</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>QUNRE</td>
<td>.013</td>
<td>.011</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>RSMLE</td>
<td>-.105</td>
<td>-.110</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>.015</td>
<td>.012</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.010</td>
<td>.011</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.022</td>
<td>.017</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>UNRE</td>
<td>.024</td>
<td>.013</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>QUNRE</td>
<td>.013</td>
<td>.011</td>
<td>.012</td>
</tr>
<tr>
<td>c</td>
<td>RSMLE</td>
<td>-.841</td>
<td>-.842</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>-.130</td>
<td>-.127</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.014</td>
<td>.012</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.016</td>
<td>.008</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>UNRE</td>
<td>.032</td>
<td>.036</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>QUNRE</td>
<td>.014</td>
<td>.011</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>QUNRSE</td>
<td>.016</td>
<td>.015</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>QINRE</td>
<td>.016</td>
<td>.014</td>
<td>.12</td>
</tr>
<tr>
<td>d</td>
<td>RSMLE</td>
<td>-.138</td>
<td>-.145</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>.018</td>
<td>.014</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.015</td>
<td>.021</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.010</td>
<td>.011</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>UNRE</td>
<td>.034</td>
<td>.027</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>QUNRE</td>
<td>.016</td>
<td>.013</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>QUNRSE</td>
<td>.015</td>
<td>.013</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>QINRE</td>
<td>.015</td>
<td>.014</td>
<td>.12</td>
</tr>
<tr>
<td>e</td>
<td>RSMLE</td>
<td>.009</td>
<td>.009</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>QRSMLE</td>
<td>.016</td>
<td>.014</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.015</td>
<td>.013</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>UNRE</td>
<td>.021</td>
<td>.015</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>QUNRE</td>
<td>.015</td>
<td>.013</td>
<td>.12</td>
</tr>
</tbody>
</table>
the aim of obtaining more efficient estimators for the parameters of interest. However, in these examples, it becomes clear that this is not the only advantage of this procedure, since it produces more robust estimators to the sampling issue of nonignorable nonresponse.\(^8\)

All estimators proposed to deal with nonignorable missing data perform relatively well. None of the results appears to be strongly affected by the experimental design, apart from some adverse effects on the variability of \(\hat{\theta}\) when the complete sample size \(n\) is reduced, which become more serious when \(P^*\) is close to 1, in design \(d\). In effect, the mean and median biases of the proposed estimators are small; see also Figures 1 and 2, in which the estimated densities of \(\hat{\theta}\) for these estimators are always centrally located around the true value of \(\theta = 0.251\). These statistics show that the inclusion of information on \(X\) from nonrespondents and units of the SRS is only relevant when \(Q\) is unknown. In fact, while UNRSE and INRE exhibit, in general, better results than UNRE, specially when \(P^*\) is reduced, the mean and the median biases are very similar for QUNRE, QUNRSE, and QINRE. On the other hand, in these examples, the inclusion of information on \(Q\) only appears to be relevant for bias correction under UNR, the case where less data is available.

Examining the column containing the standard deviations across replications, the improvements due to the knowledge of \(Q\) are obvious. In fact, QUNRE, QUNRSE and QINRE, relative to the respective versions where \(Q\) is estimated, reduce the dispersion to, at least, 31\% (compare also the line representing QUNRE in Figure 1, the estimator using information on \(Q\) with the worst behaviour in terms of dispersion, with those for UNRE, UNRSE and INRE). Moreover, in experiments \(b\), \(c\) and \(d\), the standard deviation across replications are reduced in both UNRSE and INRE relative to UNRE, and QUNRSE and QINRE relative to QUNRE, as a result of including information on \(X\) from the incomplete questionnaires. These improvements are more considerable in design \(c\), where the differential between \(P_{y=1}\) and \(P_{y=0}\) is large and the complete sample size is small, a situation in which the information on \(X\) incorporated in UNRS and INR has an increasing weight relative to that on \((Y, X)\) provided by units with complete responses. It is also clear that when \(Q\) is unknown the reductions of the variability are more significant in INRE than in UNRSE. Thus, as in our experiments the size of the SRS in UNRS equals the number of nonrespondents in INR, we may conclude that the values of \(X\) provided by the nonrespondents are more informative than those from the SRS. Finally, it is also worth to note that RSMLE underestimate the variability of the data, which is a common behaviour when a sampling problem,

\(^8\)Similar conclusions, concerning problems of misclassification in the variable of interest under CB sampling and measurement error in the covariates, were achieved in simulation studies conducted by Ramalho (2001, 2002).
Table 4: Probit model - summary statistics for $P^*$ estimates from 1000 replications

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Estimator</th>
<th>Bias Mean</th>
<th>Bias Median</th>
<th>St. D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>UNRE</td>
<td>.006</td>
<td>-.020</td>
<td>.204</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>-.001</td>
<td>-.030</td>
<td>.192</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.020</td>
<td>-.021</td>
<td>.203</td>
</tr>
<tr>
<td>c</td>
<td>UNRE</td>
<td>.006</td>
<td>-.047</td>
<td>.078</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.036</td>
<td>-.019</td>
<td>.107</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.005</td>
<td>-.023</td>
<td>.044</td>
</tr>
<tr>
<td>d</td>
<td>UNRE</td>
<td>.017</td>
<td>-.034</td>
<td>.253</td>
</tr>
<tr>
<td></td>
<td>UNRSE</td>
<td>.060</td>
<td>-.021</td>
<td>.284</td>
</tr>
<tr>
<td></td>
<td>INRE</td>
<td>.031</td>
<td>-.018</td>
<td>.216</td>
</tr>
<tr>
<td>e</td>
<td>UNRE</td>
<td>.007</td>
<td>-.023</td>
<td>.239</td>
</tr>
</tbody>
</table>

not only nonresponse but also several forms of measurement error, is ignored; see, for example, Hausman et al. (1998) and Chesher (1998), who deal with two different forms of measurement error.

Additionally, for each experiment with missing data (designs $b$ to $e$) we estimated the ratio $P^*$ for cases where $Q$ is unknown using expression (9). The mean and the median bias in percentage terms and the standard deviation across the replications for these estimates are presented in Table 4. The conclusions are similar to those for $\theta$. The mean and median biases are small and exhibit the worst results for UNR with the two smaller values of $P^*$ (experiments $c$ and $d$) and for UNRS, when $m$ is small (experiment $b$). Also, the variability of these estimates seems to be dependent on $P^*$: the standard deviations are smaller in design $c$, with the smallest $P^*$, and then increase dramatically in the other cases, specially in design $d$, where a relatively large value of $P^*$ is associated to a small complete sample size $n$.

In general, these Monte Carlo experiments show the importance of using all the available information in the estimation procedure. Undoubtedly, the aggregate information on $Q$ is the major source of improvements, followed by data on $X$ from incomplete responses, and, finally, data on $X$ from a SRS. The use of one of these two last forms of information when $Q$ is available appears not to offer any advantage. However, the incorporation of the known value of $Q$ appears to be beneficial in all cases.

5 Conclusion

This paper proposed an unified methodology to deal with several nonignorable missing data problems when the variable of interest is discrete. We specified the regression models of interest
by modifying the setup usually employed with CB sampling. Then, Imbens’ (1992) efficient GMM estimators proposed for CB samples were extended to deal with all the missing data problems of interest.

Our framework, besides the correct specification of the structural model, merely requires that the probability of response conditional on the variable of interest is independent of the covariates. The former assumption is present in all likelihood-based analyses. As for the latter, it is meet in many practical situations. In cases of INR (and INRS) it is reasonable to assume that the covariates influence similarly the variable of interest and the willingness to report the value of that variable. Under UNR (and UNRS), that assumption is likely to be meet in cases where the refusal to participate in the survey is specially motivated by the unwillingness to reveal the variable of interest.

The advantages of developing an integrated approach for several nonresponse patterns are obvious. Not only the same methodology could be employed for both model specification and the derivation of the estimators for all cases, but also allowed the investigation of the relations between the different nonresponse problems. In fact, we had a nonresponse pattern, INRS, which encompassed all the others. Discarding the information on the covariates provided by nonrespondents and individuals of the SRS, respectively, UNRS and INR, were obtained. Then, the suppression of similar information from these two cases yielded UNR. Thus, straightforwardly we could analyse how that data was incorporated in the estimation procedure. Moreover, it also became apparent that when the structural model was a MIM and the aim was the consistent estimation of the slope parameters of interest, the RS estimation methods could be employed in all cases with the complete data set.

A small Monte Carlo simulation study concerning most of the estimators suggested revealed very promising results. In effect, considering an initial random sample of size 300 and various combinations of rates of response across two alternatives, we concluded that the estimators suggested here have a negligible bias, which was especially reduced in cases where data on the covariates, provided by either the nonrespondents or the units of a SRS, were incorporated in the estimation procedure. In contradistinction, RSML estimators were considerably biased in all cases where the response rates across the alternatives were different, even in experiments where this differential was not very substantial. The combination of aggregate information on the marginal choice probabilities with the micro data arose as an important issue. Not only our modified estimators were substantially improved, but also the estimators for RS employed with the complete data set became much more robust to nonignorable nonresponse.
Appendix: efficiency of the generalized method of moments estimators for UNRS and UNR

Following Imbens (1992), the efficiency of our estimators, for both the cases where the exact value $Q_y$ is known or no information on this quantity is available, can be proved by showing that the Cramér-Rao lower bounds associated with a sequence of parametric models which satisfy the same regularity conditions as our model, converges to the asymptotic covariance matrix of our semiparametric estimators.

To construct the sequence of parametric models recall that $X$ has density $f(x)$ in $\mathcal{X}$. For any $\varepsilon > 0$, partition $\mathcal{X}$ into $L_\varepsilon$ subsets $\mathcal{X}_l$ where, for $l \neq m$, $\mathcal{X}_l \cap \mathcal{X}_m = \emptyset$ and, if $x, z \in \mathcal{X}_l$, then $|x - z| < \varepsilon$. Define $\phi_{lx} = 1$ if $x \in \mathcal{X}_l$ and 0 otherwise, and $f_\varepsilon(x) = f(x) \left[ \sum_{l=1}^{L_\varepsilon} \phi_{lx} f(x) dx \right]^{-1}$, such that $f(x, \varpi) = f_\varepsilon(x) \sum_{l=1}^{L_\varepsilon} \phi_{lx} \varpi_l$, where $\varpi_l = \Pr(x \in \mathcal{X}_l) = \int_{\mathcal{X}_l} f(x) dx$ and $f_\varepsilon(x)$ is a known function.

Under UNRS, the parametric model indexed by $\varepsilon$, which result from substituting $f(x, \varpi)$ in (13), is

$$I_{UNRS, \varepsilon}(v, s) = \left( \frac{H_y \Pr(y|x, \theta) f_\varepsilon(x) \sum_{l=1}^{L_\varepsilon} \phi_{lx} \varpi_l \int_{\mathcal{X}_l} \Pr(y|x, \theta) f_\varepsilon(x) \phi_{lx} dx}{\sum_{l=1}^{L_\varepsilon} \varpi_l \int_{\mathcal{X}_l} \Pr(y|x, \theta) f_\varepsilon(x) \phi_{lx} dx} \right)^{1-r} \left( 1 - H_S - \sum_{y \in \mathcal{Y}} H_y \right)^{1-s}$$

$$\left( H_S f_\varepsilon(x) \sum_{l=1}^{L_\varepsilon} \phi_{lx} \varpi_l \right)^s,$$

which, as $f_\varepsilon(x)$ is a known function, depend on the unknown vector of parameters $(H, \theta, \phi_{lx})$.

Constructing the log-likelihood function, taking the first order derivatives and noting that the ML estimator for $Q_y$ is written from (2) as

$$\hat{Q}_y = \sum_{l=1}^{L_\varepsilon} \varpi_l \int_{\mathcal{X}_l} \Pr(y|x, \hat{\theta}) f_\varepsilon(x) \phi_{lx} dx,$$

the dependence on $\varpi_l$ can be removed by the same procedure employed to remove dependence on $\hat{\pi}_l$ in the system (20)-(24). The resultant moment indicators are

$$g_\varepsilon(v, s, \varphi)_{H_i} = (1 - s) r I_{(y=t)} - H_i$$

$$g_\varepsilon(v, s, \varphi)_{H_S} = s - H_S$$

$$g_\varepsilon(v, s, \varphi)_{\varpi_l} = (1 - s) r \nabla_{\varphi} \ln \Pr(y|x, \theta) - [(1 - s) r + s] \nabla_{\varphi} \ln \left[ H_S + \sum_{y \in \mathcal{Y}} H_y \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} f_\varepsilon(x) \Pr(y|x, \theta) dx \right]$$
\[ g_{e}(v, s, \varphi)_{Q_{e}} = Q_{y} - \frac{\left(1 - s\right) r + s}{H_{S} + \sum_{y \in Y} \frac{\partial}{\partial \theta} \sum_{l=1}^{L} \phi_{l}(x) f_{l}(x) \Pr(y|x, \theta) dx} \text{.} \]

To compare the asymptotic covariance matrix of this parametric estimator with that of our semiparametric estimator, define \( E_{e}[\Pr(y|x, \theta)] = \sum_{l=1}^{L} \phi_{l}(x) f_{l}(x) \Pr(y|x, \theta) dx \) and \( E_{e}[\nabla_{\theta} \Pr(y|x, \theta)] \) similarly. Hence, it is clear that these systems correspond to, respectively, (29)-(32) with \( \Pr(y|x, \theta) \) and \( \nabla_{\theta} \Pr(y|x, \theta) \) replaced by their expectations.

Assuming that \( \Pr(y|x, \theta) \), \( \nabla_{\theta} \Pr(y|x, \theta) \) and \( \nabla_{\theta\theta} \Pr(y|x, \theta) \) are continuously differentiable with respect to \( x \), there is uniform convergence of \( E_{e}[\Pr(y|x, \theta)] \), \( E_{e}[\nabla_{\theta} \Pr(y|x, \theta)] \) and \( E_{e}[\nabla_{\theta\theta} \Pr(y|x, \theta)] \) to \( \Pr(y|x, \theta) \), \( \nabla_{\theta} \Pr(y|x, \theta) \), and \( \nabla_{\theta\theta} \Pr(y|x, \theta) \), respectively. Thus, in the case when there is no information on \( Q_{y} \), the limits of \( \Omega_{e} = E_{e}[g_{e}(v, s, \varphi) g_{e}(v, s, \varphi)'] \) and \( G_{e} = E_{e}[\nabla_{\varphi} g_{e}(v, s, \varphi)'] \) equal those of \( \Omega \) and \( G \) and the covariance matrix, \( G_{e}^{-1} \Omega_{e} G_{e}^{-1} \), the Cramér-Rao bound, converges to \( G^{-1} \Omega G^{-1} \), which implies that our semiparametric estimators are efficient. Analogously, in presence of exact information on \( Q_{y} \), merely by re-defining \( \Omega_{e} \) and \( G_{e} \), the same conclusion is reached, since the covariance matrix of the optimal parametric-based GMM estimator \( (G_{e} \Omega_{e}^{-1} G_{e}')^{-1} \) converges to \( (\Omega^{-1} G)^{-1} \).

References


251-299.


data when the total sample size is known”, *Journal of the Royal Statistical Society*, Series B, 60, pp. 243-254.


Figure 1: Probit model with missing data - estimated sampling distributions for the parameter estimates

Notes: RSMLE (solid line), UNRE (dotted line), UNRSE (dot-dashed line), INRE (dashed line) and QUNRE (three-dot-dashed line).
Figure 2: Probit model with missing data - estimated sampling distributions for the parameter estimates

Notes: RSMLE (solid line), QRSMLE (dotted line), QUNRE (dot-dashed line), QUNRSE (dashed line) and QINRE (three-dot-dashed line).