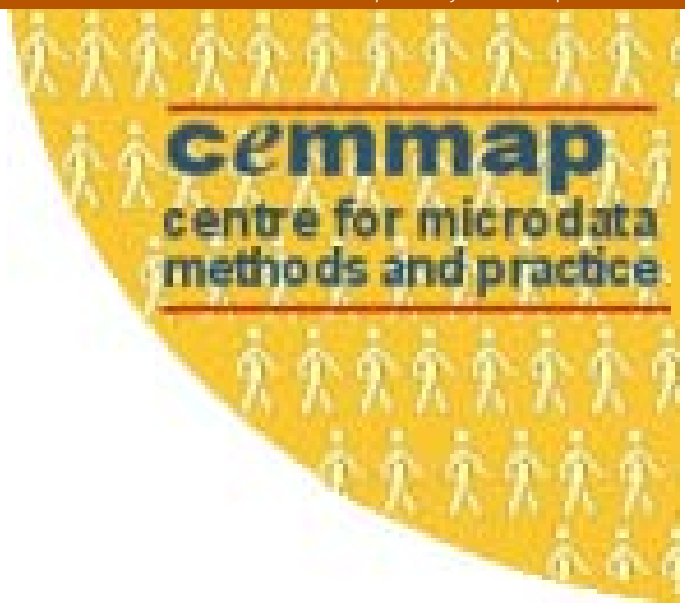


DISCRETE CHOICE NONRESPONSE

Esmerelda A. Ramalho
Richard J. Smith

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP07/03

Discrete Choice Nonresponse*

Esmeralda A. Ramalho
CEMAPRE
and
Departamento de Economia
Universidade de Évora
Portugal

Richard J. Smith†
CEMMAP
and
Department of Economics
University of Warwick
U.K.

First Draft: January 2002
This Revision: July 2003

Abstract

Missing values are endemic in the data sets available to econometricians. This paper suggests a unified likelihood-based approach to deal with several nonignorable missing data problems for discrete choice models. Our concern is when either the dependent variable is unobserved or situations when both dependent variable and covariates are missing for some sampling units. These cases are also considered when a supplementary random sample of observations on all covariates is available. A unified treatment of these various sampling structures is presented using a formulation of the nonresponse problems as a modification of choice-based sampling. Extensions appropriate for nonresponse are detailed of Imbens' (1992) efficient generalized method of moments (GMM) estimator for choice-based samples. Simulation evidence reveals very promising results for the various GMM estimators proposed in this paper.

JEL Classification: C25, C51.

Keywords: Generalized Method of Moments Estimation, Missing Completely at Random, Nonignorable Nonresponse, Semiparametric Efficiency.

---

*The first author gratefully acknowledges partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. Earlier versions of this paper were presented at the 2002 Econometric Society European Meeting, Venice, the 2002 European Winter Meetings of the Econometric Society, Budapest, and seminars at Birkbeck College, Erasmus University, Rotterdam, and the University of Warwick. We are grateful to participants at those seminars and meetings for their helpful comments.

†Corresponding Author. Mailing Address: Department of Economics, University of Warwick, Coventry CV4 7AL, U.K. E-mail Address: r.j.smith@warwick.ac.uk

1 Introduction

Survey sampling is principally conducted to gather complete information on all sampling units. Due to a variety of reasons, nonresponse is an unfortunate but endemic feature of sample surveys. For a fraction of the subjects either no data at all are available or information on one or more variables is missing. Indeed, some sampling units may simply refuse to participate at all in the study or answer the questionaire incompletely. The interviewer may not be able to contact all the sampling units or fails to ask all questions. Some questionnaires or parts thereof may be destroyed in data processing. Conversely, there are also cases where the presence of missing values is a deliberate part of the sampling process. In variable probability sampling, for example, an observation is randomly drawn from the population and the stratum to which it belongs is identified, the observation being retained in the sample with a probability defined by the agent who collects the sample.[1] Because the latter sampling scheme deliberately generates incomplete data, the mechanism which governs the missingness pattern is known. In the former situation, which is the subject of this paper, in contradistinction, nothing is generally known about the missingness mechanism as data is missing for reasons beyond the control of the researcher.

In econometrics, nonresponse has been addressed primarily in the context of panel data studies, where often some sampling units will drop out after participating in the initial waves of the survey; see, for example, Ridder (1990), Fitzgerald, Gottschalk and Moffitt (1998) and Hirano, Imbens, Ridder and Rubin (2001). In contrast, Horowitz and Manski (1995, 1998, 2001) provide a general discussion of nonparametric identification for regression with missing data on either (both) the variable of interest or (and) the covariates. An enormous statistical literature has also been developed to address the

[1]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 obtained but only for a subset of the survey participants.

issue of nonresponse; see *inter alia* 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. For the former class, most of the literature suggests the use of weighting adjustments, which involve the assignment of weights to respondents to compensate for their systematic differences relative to nonrespondents. For the latter form of nonresponse, most papers propose imputation inference procedures in which the missing values are filled in to produce complete data sets.

Many empirical studies, however, do not adopt either of the above approaches or that taken in this paper, simply discarding all sampling units with missing values and employing the usual inference procedures associated with random sampling (RS). This practice may seriously bias results when the characteristics of respondents and nonrespondents differ systematically, that is, when the missingness mechanism is endogenous. The non-ignorable nature of nonresponse arises because the rate of response may differ across the possible values taken by the dependent variable which thereby causes the observed data to provide a distorted picture of the features of the population of interest. Therefore, for likelihood-based inference, an appropriate model for a description of the available data becomes a complicated function of the structural model defined by the assumed population conditional distribution of the dependent variable given the covariates and the missing data mechanism. As the observation of the sampling units may depend on the dependent variable, an additional complication may arise because the covariates are no longer ancillary for the parameters of interest, rendering conditional maximum likelihood (ML) estimation given the covariates inefficient.

This paper proposes a unified likelihood-based approach for parametric discrete choice models with missing data in a cross-section context. We address cases where the discrete reponse variable, and possibly the covariates, are missing for some sampling units. This set-up is adapted 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 at all. Specifically, we address cases where

observations on the response variable only are missing for some subjects, designated as *item nonresponse* (INR), and where both response variable and covariates are missing for some sampling units, termed *unit nonresponse* (UNR). Furthermore, we reconsider these two situations, denoted respectively as INRS and UNRS, when a supplementary random sample (SRS) is available, consisting of observations on all covariates and assumed independent of the main sample. Such additional information might naturally arise from census data; see Cosslett (1981a). Analysis focusses on the more general INRS and UNRS, which are then specialized for INR and UNR.

To provide a unified framework, we allow three types of sampling unit to be present. Two types belong to the main sample: those for which all the data is available and those for which information is absent or is incomplete, designated respectively respondents and nonrespondents. We therefore reserve the terms *response* for a complete response, while *nonresponse* describes either an incomplete or the absence of response. The third class of sampling unit is those included in the SRS. All incomplete data patterns are underpinned by the same (unknown) missing data mechanism which is assumed to be completely determined by the response variable.[2] That is, individual characteristics included in the covariates do not contain any additional information on unit response/nonresponse over and above that provided by the response variable. The probabilities defining the missing data mechanism do not require prior knowledge, being treated as additional parameters to be estimated. The distribution of the covariates is handled semiparametrically. Central to our analysis is a similarity of nonresponse to choice-based (CB) samples. Consequently, all of the aforementioned incomplete data patterns may be formulated as modifications of CB sampling. Therefore, Imbens' (1992) efficient generalized method of moments (GMM) approach may be adapted and extended to our context.

This paper is organized as follows. Section 2 formalizes the model specification for the

[2]This assumption may be straightforwardly relaxed to permit a degree of dependence on the covariates also if this dependence is expressed in terms of a finite partition of the covariate sample space, for example, a partition defined by discrete-valued covariates. However, to achieve an economy of notation, we confine attention in the main part of the text to a missingness mechanism determined purely in terms of the discrete choice dependent variable with appropriate modifications given in a series of footnotes which deal with missingness defined additionally in terms of covariates.

missing data problems of interest. Section 3 details the observed data likelihoods. GMM estimators are developed and compared in section 4. Specification tests are described in section 5. Section 6 reports some simulation evidence on the performance of some of the proposed estimators. Finally, section 7 concludes. Some technical details are relegated to Appendices.

2 Model Specification

2.1 Some Notation

The response variable is denoted by Y and takes values on a set $\mathcal{Y}$ of $(C+1)$ mutually exclusive alternatives, $\mathcal{Y} = \{0, 1, ..., C\}$. Let $X \in \mathcal{X}$ be a p-vector of weakly exogenous covariates. The random variables Y and X are assumed to be defined on $\mathcal{Y} \times \mathcal{X}$ with population joint density function

$$f(y, x, \theta) = \mathcal{P}\{y|x, \theta\} f_X(x), \tag{2.1}$$

where the discrete probability $\mathcal{P}\{.|., \theta\}$ is known up to the parameter vector θ of dimension p and the marginal density function $f_X(.)$ for X is unknown. The problem addressed in this paper is consistent estimation of and efficient inference on the parameter vector θ. Where there is no loss of clarity, we suppress the dependence on θ of (2.1) and other joint density functions.

Let θ_0 denote the true value of θ. The population probability of observing $Y = y$ is

$$\begin{aligned} Q_y &= \mathcal{P}\{Y = y\} \\ &= \int_{\mathcal{X}} \mathcal{P}\{y|x, \theta_0\} f_X(x) dx, \end{aligned} \tag{2.2}$$

where $0 < Q_y < 1$, $y \in \mathcal{Y}$, and $\sum_{y \in \mathcal{Y}} Q_y = 1$. The probabilities Q_y, $y \in \mathcal{Y}$, may in fact be known, for example, from a large random sample like a census. In such circumstances, this information is treated as if it were exact similarly to the approach in the choice-based (CB) sampling literature; see, for example, Manski and Lerman (1977), Imbens (1992) and Wooldridge (1999, 2001).

2.2 Survey Sampling Structure

The survey objective is to collect a random sample (RS) of size N of complete observations on Y and X. Suppose, however, that only n sampling units provide all the information requested. These respective samples are designated the *initial* (or *incomplete*) and *complete* samples whereas those sampling units in the initial sample which provide observations on either both Y and X or X only comprise the *main* sample.

Assumption 2.1 *(Initial Sample (IS).) The IS is a random sample of size N.*

We additionally assume that an independent supplementary random sample (SRS) of observations of size m on X is drawn from the population of interest.

Assumption 2.2 *(Supplementary Random Sample (SRS).) The SRS of observations of size m on X is independent of the main sample.*

Let the binary indicator S take value 1 when the sampling unit belongs to the supplementary data set and 0 otherwise. Also define $N_m = N + m$ and $n_m = n + m$.

Alternative $Y = y$ is chosen by N_y individuals, of whom only n_y provide complete questionnaires. Hence, $N = \sum_{y \in \mathcal{Y}} N_y$ and $n = \sum_{y \in \mathcal{Y}} n_y$.

As all incomplete data problems considered here involve missing data on (Y, X) or on Y only, we always observe n_y, n and m but never N_y, $y \in \mathcal{Y}$. The size of the initial random sample, N, is always available for item nonresponse (INR) (and INR with SRS (INRS)), since the covariates are measured for all units. For unit nonresponse (UNR) (and UNR with SRS (UNRS)) N may or may not be known to the econometrician. However, our exposition assumes knowledge of N for three reasons. Firstly, the same approach may be followed for both INRS (INR) and UNRS (UNR). Secondly, the analysis is straightforwardly adapted for UNRS (UNR) when N is unknown. Finally, inclusion of information on N improves inference for the parameters of interest.[3]

[3]See Li and Qin (1998) for a discussion of several examples of biased data where information on N improves semiparametric likelihood-based inference.

2.3 Missing Data Mechanism

A critical assumption is that, conditional on Y, unit response/nonresponse is independent of the covariates X, that is, the influence of X on response/nonresponse is only transmitted through the response variable Y.

Define the binary indicator

$$R = \begin{cases} 1 & \text{if } (Y, X) \text{ is fully observed} \\ 0 & \text{if either } Y \text{ or } (Y, X) \text{ is missing} \end{cases}.$$

Assumption 2.3 *(Conditional Probability of Response.) The conditional probability P_y of observing a respondent unit given $Y = y$ and X is independent of X; that is*

$$\begin{aligned} P_y &= \mathcal{P}\{R = 1|Y = y, X = x\} \\ &= \mathcal{P}\{R = 1|Y = y\}, \end{aligned} \tag{2.3}$$

where $0 < P_y < 1,\ y \in \mathcal{Y}$.

In all cases, we assume that $0 < P_y < 1$. If $P_y = 0$, alternative $Y = y$ would not be observed in the complete sample. If, on the other hand, $P_y = 1$, then there would be no missing values among units with $Y = y$.[4]

When a SRS is available, by Assumptions 2.2 and 2.3, $\mathcal{P}\{R = 1|Y = y, X = x, S = 0\} = \mathcal{P}\{R = 1|Y = y\}$. Hence, although Y is not observed in the SRS, the missingness pattern, namely P_y of (2.3), in the main sample is all that is required.

Combining (2.2) and (2.3), the probability of observing a respondent unit is

$$\mathcal{P}\{R = 1\} = \sum_{y \in \mathcal{Y}} P_y Q_y, \tag{2.4}$$

[4] As noted in the Introduction, Assumption 2.3 may be weakened to allow response/nonresponse to depend also on a finite partition of the sample space $\mathcal{X}$ of the covariates. Let $\mathcal{X}_{\mathsf{j}}$, $j \in \mathcal{J}$, $\mathcal{J} = \{1, ..., M\}$, be a partition of $\mathcal{X}$ such that $\mathcal{X}_{\mathsf{j}} \cap \mathcal{X}_{\mathsf{k}} = \emptyset$, $j \neq k$, and $\mathcal{X} = \cup_{\mathsf{j} \in \mathcal{J}} \mathcal{X}_{\mathsf{j}}$. Define the random variable $J = j$ if $X \in \mathcal{X}_{\mathsf{j}}$. Then Assumption 2.3 and (2.3) are modified to

$$\begin{aligned} P_{\mathsf{y}}^{\mathsf{j}} &= \mathcal{P}\{R = 1|Y = y, X = x\} \\ &= \mathcal{P}\{R = 1|Y = y, J = j\}, \end{aligned}$$

if $x \in \mathcal{X}_{\mathsf{j}}$, $j \in \mathcal{J}$.

which, because P_y is unknown, in general, will also be unknown even though Q_y may be known.[5]

If the rate of response were the same for all alternatives, that is, $P_y = P$, $y \in \mathcal{Y}$, the data are said to be *missing completely at random* (MCAR) [Little and Rubin (1987)] as in this case the complete sample is also random. Naturally, RS estimation methods may be used since nonresponse is ignorable, units with missing values being no different from those with complete information.

If only information on X was missing, according to the mechanism (2.3), data would be *missing at random* (MAR), because the probability of recording X would be independent of X after controlling for Y, which in this case would be observed for all subjects. This problem falls outside of the scope of this paper as the missingness mechanism is ignorable for likelihood-based inference if P_y of (2.3) does not depend on θ; see Rubin (1976) and Little and Rubin (1987). Most of the statistical literature on nonresponse focusses on data MAR, dealing mainly with procedures for imputing missing values; see, for example, Little and Rubin (1987) and Schafer (1997). In econometrics, the issue of nonignorable nonresponse has been considered 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)].

2.4 Missing Data Formulation by Stratification

An important point of departure for this paper is the adaptation of the approach taken in the CB sampling literature to the missing data problems considered here. In order to do so, we reinterpret respondents and nonrespondents as strata for each discrete value of Y. For both INRS and UNRS a further stratum including the SRS of units is added.

[5] If response/nonresponse depend on the finite partition $\mathcal{X} = \cup_{\mathsf{j}=1}^{\mathsf{J}} \mathcal{X}_{\mathsf{j}}$ then

$$\mathcal{P}\{R = 1\} = \sum_{\mathsf{y} \in \mathcal{Y}} \sum_{\mathsf{j} \in \mathcal{J}} P_{\mathsf{y}}^{\mathsf{j}} Q_{\mathsf{y}}^{\mathsf{j}},$$

where $Q_{\mathsf{y}}^{\mathsf{j}} = \mathcal{P}\{Y \in \mathcal{Y}, J = j\} = \int_{\mathcal{X}_{\mathsf{j}}} \mathcal{P}\{y|x,\theta\} f_{\mathsf{X}}(x)dx$, cf. (2.2).

First of all, then, there are $C+1$ strata containing the respondent subjects for each value of Y. The proportions of each of these strata in the sample and in the population are denoted by H_y and Q_y respectively; see (2.2). Secondly, $C+1$ additional strata contain the nonrespondent individuals for each response Y. Each of these strata has an unknown sampling proportion H_y^{nr} but the same population proportion Q_y. Therefore, the initial random sample is interpreted as a combination of two CB samples consisting of the respondent and the nonrespondent sampling units. Finally, the stratum containing the SRS has a proportion of $\mathcal{P}\{S=1\} = H_S$ 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_y^{nr} are defined differently according to the presence or otherwise of a SRS and whether N is known or unknown. We firstly examine H_y and H_y^{nr} in the presence of a SRS. The absence of a SRS is dealt with as a special case.

2.4.1 Known N

The probability of observing a respondent unit and $Y=y$ is

$$H_y = \mathcal{P}\{Y=y, R=1, S=0\}, \tag{2.5}$$

while the corresponding probability for nonrespondent units is

$$H_y^{nr} = \mathcal{P}\{Y=y, R=0, S=0\}. \tag{2.6}$$

Aggregating over $\mathcal{Y}$ yields the probability of observing, respectively, respondent, $\mathcal{P}\{R=1, S=0\} = \sum_{y\in\mathcal{Y}} H_y$, and nonrespondent units, $\mathcal{P}\{R=0, S=0\} = \sum_{y\in\mathcal{Y}} H_y^{nr}$. A further summation reveals the proportion of the main sample in the full data set (the main and the supplementary samples)

$$\begin{aligned}
\mathcal{P}\{S &= 0\} = 1 - H_S \\
&= \sum_{r=0}^{1} \sum_{y\in\mathcal{Y}} \mathcal{P}\{Y=y, R=r, S=0\} \\
&= \sum_{y\in\mathcal{Y}} H_y + \sum_{y\in\mathcal{Y}} H_y^{nr}.
\end{aligned} \tag{2.7}$$

From Assumption 2.2, from independence, the marginal probability of observing $Y = y$ in the population may be rewritten as

$$\begin{aligned} Q_y &= \mathcal{P}\{Y = y | S = 0\} \\ &= \frac{\sum_{r=0}^{1} \mathcal{P}\{Y = y, R = r, S = 0\}}{1 - H_S} \\ &= \frac{H_y + H_y^{nr}}{1 - H_S}. \end{aligned} \tag{2.8}$$

This result will prove useful later as it permits the estimation of the unknown sample probabilities H_y^{nr}, $y \in \mathcal{Y}$, to be avoided.

Also, by Assumption 2.2, cf. (2.8), $\mathcal{P}\{Y = y, S = 0\} = Q_y (1 - H_S)$. Hence, by Assumptions 2.2 and 2.3,[6]

$$\begin{aligned} P_y &= \mathcal{P}\{R = 1 | Y = y, S = 0\} \\ &= \frac{\mathcal{P}\{Y = y, R = 1, S = 0\}}{\mathcal{P}\{Y = y, S = 0\}} \\ &= \frac{H_y}{Q_y (1 - H_S)}. \end{aligned} \tag{2.9}$$

From (2.9), as $0 < P_y < 1$ by Assumption 2.3, $0 < H_y < Q_y (1 - H_S)$. Moreover, data MCAR are characterized by H_y/Q_y constant for all y because P_y is invariant across $y \in \mathcal{Y}$. In all cases H_y may be estimated from the incomplete sample as n_y/N_m. Hence, equation (2.9) may be used to estimate P_y when Q_y is either known or estimated by the methods set out in section 4.

2.4.2 Unknown N

To adapt the above analysis for when N is unknown, the $(C + 1)$ strata containing nonrespondents are suppressed, since we now only consider respondent individuals in the main sample. Now H_y is defined as the sampling probability of observing $Y = y$ in the main sample conditional on $R = 1$:

$$H_y = \mathcal{P}\{Y = y, S = 0 | R = 1\}. \tag{2.10}$$

[6] If response/nonresponse depends on the finite partition $\mathcal{X} = \cup_{\mathsf{j}=1}^{\mathsf{J}} \mathcal{X}_{\mathsf{j}}$ then define $H_{\mathsf{y}}^{\mathsf{j}} = \mathcal{P}\{Y = y, J = j, R = 1, S = 0\}$ with a similar definition for $H_{\mathsf{y}}^{\mathsf{nrj}}$; cf. (2.5) and (2.6). Then $Q_{\mathsf{y}}^{\mathsf{j}} = (H_{\mathsf{y}}^{\mathsf{j}} + H_{\mathsf{y}}^{\mathsf{nrj}})/(1 - H_{\mathsf{S}})$ and $P_{\mathsf{y}}^{\mathsf{j}} = H_{\mathsf{y}}^{\mathsf{j}}/Q_{\mathsf{y}}^{\mathsf{j}}(1 - H_{\mathsf{S}})$. Cf. (2.8) and (2.9).

Consequently, $\mathcal{P}\{S=0|R=1\} = \sum_{y\in\mathcal{Y}} H_y$. From Assumption 2.2, as $\mathcal{P}\{S=0\} = \mathcal{P}\{S=0|R=1\}$,

$$1 - H_S = \sum_{y\in\mathcal{Y}} H_y.$$

However, the population probability Q_y may no longer be written in terms of H_y as in (2.8). Instead of (2.9), the relation between P_y, H_y, H_S and Q_y is now given by

$$\begin{aligned} H_y &= \frac{\mathcal{P}\{Y=y, R=1, S=0\}}{\mathcal{P}\{R=1\}} \\ &= \frac{P_y Q_y (1-H_S)}{\sum_{y\in\mathcal{Y}} P_y Q_y}; \end{aligned} \tag{2.11}$$

see (2.4) and (2.9).[7] From (2.11), H_y is no longer necessarily less than Q_y. Furthermore, in contrast with known N, $H_y = Q_y(1-H_S)$ for all y characterizes both data MCAR *and* the absence of missing data. For unknown N, from (2.11), even if H_y and Q_y are known, the probabilities P_y are not identified. However, for any two choices $Y = y_1$ and $Y = y_2$, their ratios may be estimated from $P_{y_1}/P_{y_2} = (H_{y_1}/H_{y_2})/(Q_{y_1}/Q_{y_2})$, which is of course 1 for all y for data MCAR.[8]

2.4.3 Unavailable SRS

When a SRS is unavailable or is not utilized, we merely deal with the main sample. All probabilities defined above are straightforwardly adapted for this situation by setting $H_S = 0$. As all sampling units are now associated with $S = 0$, $S = 0$ should also be suppressed. These alterations are also applicable to the likelihood functions defined in the next section.

[7] If response/nonresponse depends on the finite partition $\mathcal{X} = \cup_{\mathsf{j}=1}^{\mathsf{J}} \mathcal{X}_{\mathsf{j}}$ then if N is unknown $H_{\mathsf{y}}^{\mathsf{j}} = \mathcal{P}\{Y=y, J=j, S=0|R=1\}$; cf. (2.10). Then $1 - H_{\mathsf{S}} = \sum_{\mathsf{y}\in\mathcal{Y}} \sum_{\mathsf{j}\in\mathcal{J}} H_{\mathsf{y}}^{\mathsf{j}}$ and $H_{\mathsf{y}}^{\mathsf{j}} = P_{\mathsf{y}}^{\mathsf{j}} Q_{\mathsf{y}}^{\mathsf{j}}(1 - H_{\mathsf{S}})/\sum_{\mathsf{y}\in\mathcal{Y}} \sum_{\mathsf{j}\in\mathcal{J}} P_{\mathsf{y}}^{\mathsf{j}} Q_{\mathsf{y}}^{\mathsf{j}}$; cf. (2.11).

[8] An alternative formulation for nonresponse is possible by analogy with variable probability sampling (VPS) briefly outlined in the Introduction which deliberately produces missing data. This particular endogenous stratified sampling mechanism retains subjects in the sample with a pre-defined probability chosen by the sampling agent. Our missing data patterns might be obtained by regarding P_{y}, $y \in \mathcal{Y}$, as the probabilities of retention and treating them as additional parameters to be estimated. This avenue is not explored here because of the identification problems for UNRS (and UNR) arising when N is unknown and discussed below. Equation (2.11) would also require a different formulation relative to the other cases in a VPS-type framework whereas our approach produces a unified framework for estimation in all patterns of nonresponse discussed in this paper.

3 Observed Data Likelihoods

This section considers the individual likelihood functions for the observed data under both INRS and UNRS, as well as other sampling densities of interest which also provide important characterizations of INRS and UNRS. INRS is analysed first because the INRS observed data likelihood may be modified to obtain that for UNRS by eliminating the covariate information provided by nonrespondents. In fact, the same data on respondent and SRS units is observed for both INRS and UNRS, $(Y, X, R = 1, S = 0)$ and $(X, S = 1)$ respectively. For nonrespondents, we observe either $(X, R = 0, S = 0)$ for INRS or merely $(R = 0, S = 0)$ for UNRS. The generic notation $h(.)$ is used for sample density functions.

We only need consider INR and UNR again in section 4.[9]

3.1 INRS

The joint sample density function for Y, X, R and S is

$$
\begin{aligned}
h_{INRS}(y,x,r,s) &= \left[h(y,x,r=1,s=0)^r h(x,r=0,s=0)^{1-r}\right]^{1-s} h(x,s=1)^s \qquad (3.1)\\
&= \left\{[\mathcal{P}\{y,R=1,S=0\}h(x|y)]^r \left[\sum_{y\in\mathcal{Y}} \mathcal{P}\{y,R=0,S=0\}h(x|y)\right]^{1-r}\right\}^{1-s}\\
&\quad \times [H_S f_X(x)]^s\\
&= \left\{\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r \left[\sum_{y\in\mathcal{Y}} \frac{Q_y(1-H_S)-H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^{1-r}\right\}^{1-s}\\
&\quad \times [H_S f_X(x)]^s\\
&= \left\{\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r \left[\left(1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right)f_X(x)\right]^{1-r}\right\}^{1-s}\\
&\quad \times [H_S f_X(x)]^s.
\end{aligned}
$$

The second equality in (3.1) arises since $h(x|y,r,S=0) = h(x|y)$ because, from Assumption 2.2, $h(x|y,r,S=0) = h(x|y,r)$, and $h(x|y,r) = h(x|y)$ by Assumption 2.3. The

[9] In the following, if response/nonresponse depends on the finite partition $\mathcal{X} = \cup_{\mathsf{j}=1}^{\mathsf{J}}\mathcal{X}_{\mathsf{j}}$, H_{y} and Q_{y} should be replaced by $H_{\mathsf{y}}^{\mathsf{j}}$ and $Q_{\mathsf{y}}^{\mathsf{j}}$ respectively, see fns. 5 and 6, and integration over $\mathcal{X}$ ($\int_{\mathcal{X}}$) by summation over $j \in \mathcal{J}$ and integration over $\mathcal{X}_{\mathsf{j}}$ ($\sum_{\mathsf{j}\in\mathcal{J}}\int_{\mathcal{X}_{\mathsf{j}}}$).

third equality eliminates the dependence on the unknown probabilities $H_y^{nr} = \mathcal{P}\{y, R = 0, S = 0\}$ using (2.8).

The contribution of the units of the initial sample, associated with the indicator $1 - S$, to the sample density (3.1) is composed of two parts. The first term contains the information provided by respondent units and may be interpreted as the complete data likelihood, while the second term accommodates the information on the covariates provided by nonrespondent units. The third component of (3.1) is information on X provided by individuals in the SRS. Note that the data on X reported by nonrespondent and SRS units enter the density function in quite different ways. Only the behaviour of nonrespondents, which are included in the main sample, is affected by the missing data mechanism Assumption 2.3.

3.2 UNRS

Relative to INRS, nonrespondent units do not provide any information. Therefore, the joint sample density function for Y, X, R and S is

$$\begin{aligned} h_{UNRS}(y,x,r,s) &= \left[h(y,x,R=1,S=0)^r \mathcal{P}\{R=0,S=0\}^{1-r}\right]^{1-s} h(x,S=1)^s \qquad (3.2)\\ &= \left\{\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r \left[1 - H_S - \sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y}\int_{\mathcal{X}} \mathcal{P}\{y|x,\theta\}f_X(x)dx\right]^{1-r}\right\}^{1-s} \\ &\quad \times [H_S f_X(x)]^s \\ &= \left\{\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r \left[1 - H_S - \sum_{y\in\mathcal{Y}} H_y\right]^{1-r}\right\}^{1-s} [H_S f_X(x)]^s . \end{aligned}$$

Relative to (3.1), only the term associated with nonrespondents is modified, no longer being a function of X. As will become apparent, this term merely incorporates information on the total sample size N which is employed in the estimation of H_y and H_S.

3.3 Ancillarity

The conditionality principle states that inference should be conducted conditionally on statistics ancillary for the parameters of interest θ; see Cox and Hinkley (1974). The

analysis of the marginal sampling density function of the covariates from the joint sample density functions (3.1) and (3.2) reveals that INRS and UNRS are quite different in nature.

3.3.1 INRS

The sampling density function of X is

$$\begin{aligned} h_{INRS}(x) &= f_X(x)\sum_{s=0}^{1}\sum_{r=0}^{1}\left\{\left[\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^r\left[1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^{1-r}\right\}^{1-s}[H_S]^s \\ &= f_X(x)\left[\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}+1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}+H_S\right] \\ &= f_X(x), \end{aligned} \tag{3.3}$$

the population density function $f_X(.)$ which is not a function of θ. Thus, inference should be conducted using the conditional density given X[10]

$$h_{INRS}(y,r,s|x)=\left\{\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^r\left[1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^{1-r}\right\}^{1-s}[H_S]^s. \tag{3.4}$$

3.3.2 UNRS

The sampling density of X is

$$\begin{aligned} h_{UNRS}(x) &= \sum_{s=0}^{1}\sum_{r=0}^{1}\left\{\left[\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r\left[1-H_S-\sum_{y\in\mathcal{Y}}H_y\right]^{1-r}\right\}^{1-s}[H_S f_X(x)]^s \\ &= f_X(x)\left[H_S+\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]+1-H_S-\sum_{y\in\mathcal{Y}}H_y, \end{aligned} \tag{3.5}$$

which depends on θ. Hence, X is not ancillary for θ and conditional maximum likelihood given X will be inefficient.[11] Efficient estimation should therefore be based on (3.2).

[10] Note that $f_{\mathsf{X}}(x)$ may be factored out of $h_{\mathsf{INRS}}(y,x,r,s)$ in (3.1) leaving $h_{\mathsf{INRS}}(y,r,s|x)$ of (3.4).

[11] For a discussion on the issue of covariate ancillarity for problems where data are MAR, see Lawless, Kalbfleisch and Wild (1999).

3.3.3 Joint Density of R and S

In contradistinction to those for X, the joint density of the indicators R and S is identical under both INRS and UNRS. Calculations for INRS yield

$$\begin{aligned}\mathcal{P}\{R &= r, S = s\} = \left(\left[\sum_{y\in\mathcal{Y}}\int_{\mathcal{X}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)dx\right]^r\right. \\ &\left.\left\{\int_{\mathcal{X}}\left[1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]f_X(x)dx\right\}^{1-r}\right)^{1-s}\left[H_S\int_{\mathcal{X}}f_X(x)dx\right]^s \\ &= \left[\left(\sum_{y\in\mathcal{Y}}H_y\right)^r\left(1-H_S-\sum_{y\in\mathcal{Y}}H_y\right)^{1-r}\right]^{1-s}[H_S]^s. \qquad (3.6)\end{aligned}$$

From (3.6), both R and S are ancillary for θ and, thus, inference should be conducted conditional on R and S. Although, similarly to Imbens (1992) and Imbens and Lancaster (1996) for endogenous stratified sampling, estimation is based on the likelihood functions (3.2) and (3.4), which are not conditional on R and S, the method does conform with the conditionality principle as, for example, the estimator for H_s is the marginal ML estimator $\hat{H}_S = m/N_m$ obtained from (3.6).

3.4 Unknown N

As noted in section 2.4, UNRS must be adapted if the initial sample size N is unknown. The main sample now consists only of those units for which $R = 1$. As now $1 - H_S = \sum_{y\in\mathcal{Y}} H_y$ the density function of the observed data is therefore

$$\begin{aligned}h_{UNRS}(y, x, r &= 1, s) = h(y, x, r = 1, s = 0)^{r(1-s)}h(x, s = 1)^s \\ &= \left[\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^{r(1-s)}[H_S f_X(x)]^s. \qquad (3.7)\end{aligned}$$

Relative to (3.2), the terms associated with the indicator $1 - R$ have been suppressed. This density function (and the simpler version for UNR obtained when $S = 0$) coincides with that for CB sampling with (without) a SRS; see, for example, Cosslett (1981a). Thus, inference procedures appropriate for CB samples may be used if N is unknown. In a similar fashion both (3.5) and (3.6) are simplified by the elimination of the term

$1 - H_S - \sum_{y \in \mathcal{Y}} H_y$; see the comments below (2.10).

For all of the above missing data patterns, an important aspect of the analysis is that the component associated with the joint indicator $R(1-S)$, which corresponds to the complete data density, differs from the population joint density function of Y and X in (2.1) which would be appropriate under RS. Hence, unless the data are MCAR, in which case H_y/Q_y is invariant to $y \in \mathcal{Y}$ and is thus irrelevant for likelihood-based inference, RS procedures should not be used with the complete sample.

4 Generalized Method of Moments

This section adapts efficient GMM estimation under CB sampling [Imbens (1992)] for the missing data patterns discussed above. To implement GMM for INRS and UNRS, a set of moment indicators is derived, which may be employed when either the marginal population stratum probabilities Q_y, $y \in \mathcal{Y}$, are unknown or known. The parameters of interest θ are estimated jointly with the population and sample stratum occupancy probabilities, Q_y and H_y, $y \in \mathcal{Y}$, respectively, and H_S. Let the parameter vector φ denote Q_y, H_y, $y \in \mathcal{Y}$, H_S and θ and φ_0 the true value of φ.

Our reinterpretation of incomplete data problems for discrete choice models using a CB sampling setting suggests that some of the estimators originally proposed for that set-up may be relevant here also. In particular, as noted in section 3.4, all CB sampling estimators may be used to deal with UNR when the initial sample size N is unknown. As we later demonstrate, 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 if information on N is ignored. However, in the same sense that Imbens (1992) simplified Cosslett's (1981a,b) estimators for CB samples, the GMM estimators for UNRS derived here are substantially simpler than those corresponding to Cosslett (1981a). Further-

more, 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 $Y = 1$.[12,13]

The remainder of this section is organized as follows. Section 4.1 derives the moment indicators for INRS and UNRS. These moment indicators are used in section 4.2 to obtain alternative GMM estimators, appropriate for handling all the missing data patterns considered in this paper. Section 4.3 presents a brief analysis and comparison of these estimators. Finally, section 4.4 discusses particular estimation issues which arise with multiplicative intercept models (MIM).

4.1 Moment Indicators

To avoid the need to specify the marginal distribution of X, we initially assume that X is discrete with L points of support x^l with associated probability mass $\mathcal{P}\{X = x^l\} = \pi_l$, $0 < \pi_l < 1$, $l = 1, 2..., L$; we impose $L > J$ where J is the number of strata considered. This subsidiary assumption is innocuous because the nuisance parameters $\pi = (\pi_1, ..., \pi_L)$ may be concentrated out as demonstrated in Appendix A.

4.1.1 INRS

Under INRS, X is ancillary for θ and, thus, by the conditionality principle, efficient inference is conducted conditional on X using the conditional likelihood (3.4). We prefer, however, to base analysis on the joint likelihood obtained from (3.1) which allows an identical approach to be adopted for estimation under both INRS and UNRS and semiparametric efficiency to be analysed in a similar fashion for both nonresponse schemes.

[12] In this case, $\mathcal{Y} = \{0, 1\}$, $P_0 = 0$ and $P_1 = 1$. Hence, Assumption 2.3 where $0 < P_y < 1$ is relaxed to $0 \leq P_y \leq 1$.

[13] Alternatively, INR could also be described by the 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. A stratum is defined for each value of Y and the $(C + 1)$ probabilities of retention (which are known under VPS but are unknown here) are treated as additional parameters to be estimated. However, for reasons discussed in fn. 8, we do not consider the VPS framework further here.

Appendix A.1 verifies directly the equivalence of the unconditional and conditional likelihood approaches. Recall from section 4.3.3 that the indicators R and S are also ancillary for θ.

The unconditional log-likelihood function based on (3.1) is

$$\begin{aligned}\log L_{INRS}(\varphi,\pi) &= \sum_{i=1}^{N_m}\Bigg\{(1-s_i)\,r_i\log\left[\frac{H_{y_i}}{Q_{y_i}}\mathcal{P}\{y_i|x^{l_i},\theta\}\pi_{l_i}\right]+ \\ &\quad (1-s_i)\,(1-r_i)\log\left[\left(1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x^{l_i},\theta\}\right)\pi_{l_i}\right] \\ &\quad +s_i\,(\log H_S+\log\pi_{l_i})\Bigg\},\end{aligned} \tag{4.1}$$

where $Q_y=\sum_{l=1}^{L}\pi_l\mathcal{P}\{y|x^l,\theta\}$, $y\in\mathcal{Y}$. Maximization of (4.1) is undertaken subject to the restriction $\sum_{l=1}^{L}\pi_l=1$.

Given the ancillarity of S, for inference conditional on S, the marginal ML estimator for H_S is the ancillary statistic $\hat{H}_S=m/N_m$; see (3.6). Hence, from Appendix A.1, (A.6), (A.8) and (A.9), the resultant system of GMM moment indicators is

$$H_t \;:\; (1-s)\,rI(y=t)-\frac{H_t}{Q_t}\frac{(1-s)(1-r)\mathcal{P}\{t|x,\theta\}}{1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}},\ t\in\mathcal{Y}, \tag{4.2}$$

$$H_S \;:\; s-H_S, \tag{4.3}$$

$$\theta \;:\; (1-s)\Bigg\{r\frac{\partial\log\mathcal{P}\{y|x,\theta\}}{\partial\theta}- \tag{4.4}$$

$$(1-r)\left[1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^{-1}\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\frac{\partial\mathcal{P}\{y|x,\theta\}}{\partial\theta}\Bigg\},$$

$$Q_y \;:\; Q_y-\mathcal{P}\{y|x,\theta\},\ y\in\mathcal{Y}, \tag{4.5}$$

where $I(.)$ denotes an indicator function.

The presence of the multiplicative factor associated with H_t in the second term of (4.2) indicates that additional information is conveyed by the covariate information from INRS nonrespondents for the stratum probabilities, H_t, over and above that of the sample proportions, n_t/N_m, $t\in\mathcal{Y}$.

4.1.2 UNRS

The log-likelihood function based on (3.2) is

$$\begin{aligned}\log L_{UNRS}(\varphi,\pi) &= \sum_{i=1}^{N_m}\Bigg\{(1-s_i)\,r_i\log\left[\frac{H_{y_i}}{Q_{y_i}}\mathcal{P}\{y_i|x^{l_i},\theta\}\pi_{l_i}\right]+ \\ &\quad (1-s_i)\,(1-r_i)\log\left(1-H_S-\sum_{y\in\mathcal{Y}}H_y\right) \\ &\quad +s_i\,(\log H_S+\log\pi_{l_i})\Bigg\},\end{aligned} \tag{4.6}$$

where $Q_y=\sum_{l=1}^{L}\pi_l\mathcal{P}\{y|x^l,\theta\}$, $y\in\mathcal{Y}$. Maximization of (4.6) is undertaken subject to the restriction $\sum_{l=1}^{L}\pi_l=1$.

Appendix A.2 presents the first order derivatives (A.10)-(A.14) arising from (4.6) which after some manipulation result in the system of GMM moment indicators given by

$$H_t \;:\; (1-s)\,rI(y=t)-H_t,\; t\in\mathcal{Y}, \tag{4.7}$$

$$H_S \;:\; s-H_S, \tag{4.8}$$

$$\begin{aligned}\theta \;:\; &(1-s)\,r\frac{\partial\log\mathcal{P}\{y|x,\theta\}}{\partial\theta}- \\ &[(1-s)\,r+s]\left[H_S+\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^{-1}\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\frac{\partial\mathcal{P}\{y|x,\theta\}}{\partial\theta},\end{aligned} \tag{4.9}$$

$$Q_y \;:\; Q_y-[(1-s)\,r+s]\left[H_S+\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^{-1}\mathcal{P}\{y|x,\theta\},\; y\in\mathcal{Y}. \tag{4.10}$$

Equations (4.7) and (4.8) reflect the ancillarity of R and S for θ resulting in the ancillary GMM estimators $\hat{H}_y=n_y/N_m$ and $\hat{H}_S=m/N_m$; cf. (4.2).

4.2 GMM Estimation

Let $g(\varphi)$ denote the vector of moment indicators obtained after stacking either (4.2)-(4.5) or (4.7)-(4.10). A subscript i denotes evaluation at observation (y_i,x_i,s_i,r_i), $i=1,...,N_m$.

The GMM objective function is defined by

$$\hat{J}(\varphi)=\hat{g}(\varphi)'\hat{W}\hat{g}(\varphi), \tag{4.11}$$

where $\hat{W}$ is a positive semi-definite weighting matrix. The vector $\hat{g}(\varphi) = \sum_{i=1}^{N_m} g_i(\varphi)/N_m$ is the sample counterpart of the moment conditions $E\left[g(\varphi_0)\right] = 0$, where $E\left[.\right]$ denotes expectation taken over $h_{UNRS}(y, x, r, s)$ of (3.2) or $h_{INRS}(y, r, s|x)$ of (3.4) for, respectively, UNRS and INRS. Let $\hat{\varphi}$ denote the minimiser of (4.11).

4.2.1 Unknown Q_y

When the population marginal choice probability Q_y is unknown, the parameter vector φ is just-identified. We adopt the following standard regularity conditions which are sufficient for the consistency and asymptotic normality of $\hat{\varphi}$. See Imbens (1992) and Newey and McFadden (1994, Theorems 2.6 and 3.4).

Assumption 4.1 *(a)* $\theta_0 \in int(\Theta)$, Θ *a compact subset of* $\mathcal{R}^p$*; (b)* $H_y > 0$, $y \in \mathcal{Y}$, *and* $H_S > 0$.

Assumption 4.2 *(a)* $\mathcal{P}\{y|x,\theta\}$ *is twice continuously differentiable in* $\theta \in \Theta$*; (b)* $\mathcal{P}\{y|x,\theta\}$ *and* $\partial\mathcal{P}\{y|x,\theta\}/\partial\theta$ *are continuous at each* $\theta \in \Theta$*; (c)* $\mathcal{P}\{y|x,\theta\} > 0$, $y \in \mathcal{Y}$, *for all* $x \in \mathcal{X}$ *and* θ *in an open neighbourhood of* θ_0*; (d)* $f_X(x) > 0$ *for all* $x \in \mathcal{X}$*; (e)* $1 - H_S > \sum_{y\in\mathcal{Y}}(H_y/Q_y)\mathcal{P}\{y|x,\theta\}$ *for all* $x \in \mathcal{X}$ *and* θ *in an open neighbourhood of* θ_0.

Assumptions 4.2 (c) and (d) ensure that $Q_y > 0$, $y \in \mathcal{Y}$. Assumption 4.2 (e) requires a positive sample (and population) probability of observing $R = 0$ and $S = 0$, an assumption which is not required for UNR with N unknown.

Let $G = E\left[\partial g(\varphi_0)/\partial\varphi'\right]$ and $\Omega = E\left[g(\varphi_0)g(\varphi_0)'\right]$.

Assumption 4.3 *(a)* $\hat{W} \xrightarrow{p} W$, W *positive definite; (b)* φ_0 *is the unique solution to* $E[g(\varphi_0)] = 0$*; (c)* $E[\sup_\varphi \|g(\varphi)\|^2] < \infty$ *and* $E[\sup_{\varphi\in\mathcal{N}} \|\partial g(\varphi)/\partial\varphi'\|] < \infty$ *where* $\mathcal{N}$ *is a neighbourhood of* φ_0*; (d)* Ω *is nonsingular; (e)* G *is full column rank.*

These conditions lead to the following result.

Theorem 1 *(Consistency and Asymptotic Normality of $\hat{\varphi}$.) If Assumptions 2.1-2.3 and 4.1-4.3 are satisfied then*

$$\begin{aligned} &\hat{\varphi} \xrightarrow{p} \varphi_0, \\ &N_m^{1/2}\left(\hat{\varphi} - \varphi_0\right) \xrightarrow{d} N(0, G^{-1}\Omega G'^{-1}), \end{aligned} \tag{4.12}$$

where $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution respectively.

When X is discrete, $\hat{\varphi}$ is the ML estimator for φ and is, thus, asymptotically first order efficient. Asymptotic efficiency, in the semiparametric sense, is proved analogously to Imbens (1992, Theorem 3.3). Appendix B provides such a proof for UNRS and UNR; a similar proof may be obtained for INRS and INR but at the expense of more algebraic complexity.

Theorem 2 *(Efficiency of $\hat{\varphi}$.) If Assumptions 2.1-2.3 and 4.1-4.3 are satisfied then $\hat{\varphi}$ achieves the semiparametric efficiency bound.*

4.2.2 Known Q_y

When Q_y is known, the system (4.7)-(4.10) or (4.2)-(4.5) is over-identified. Let φ now denote the unrestricted parameters with the definitions for G and Ω above Assumption 4.3 suitably adapted. Also let $\tilde{\varphi}$ be a preliminary consistent estimator of φ_0, obtained for example, by setting the metric $\hat{W}$ as the identity matrix in (4.11). The optimal GMM estimator is obtained using the weighting matrix $\hat{W} = \tilde{\Omega}^{-1}$ in (4.11), where $\tilde{\Omega} = \sum_{i=1}^{N_m} g_i(\tilde{\varphi}) g_i(\tilde{\varphi})'/N_m$. Similarly to Theorem 4.1 and, in particular, (4.12),

$$\hat{\varphi} \xrightarrow{p} \varphi^0, N_m^{1/2}(\hat{\varphi} - \varphi^0) \xrightarrow{d} N(0, (G'\Omega^{-1}G)^{-1}). \tag{4.13}$$

Asymptotic efficiency of $\hat{\varphi}$ may be proved similarly to Appendix B.

4.2.3 Unknown N

Firstly, in all the above derivations and results for UNRS, N_m is replaced by n_m. Consequently, H_y is now estimated by n_y/n_m and the estimating function for H_S of (4.8) is

suppressed, because now H_S may be estimated by $1 - \sum_{y \in \mathcal{Y}} \hat{H}_y$; see the comments below (2.10). Secondly, the indicator R is set to 1 in (4.7), (4.9) and (4.10). All observations now enter the calculation of the second terms in both (4.9) and (4.10) because X is observed for all units in the sample, since strata containing nonrespondents are no longer considered. Expectations are now taken over $h_{UNRS}(y, x, R = 1, s)$ of (3.7).

4.2.4 UNR and INR

Estimators for φ under UNR (N known or unknown) and INR may be straightforwardly obtained from their respective SRS versions. In all derivations and results N_m and n_m are replaced by, respectively, N and n. In the moment indicators we set $H_S = 0$ and $S = 0$ and suppress the estimating functions for H_S in (4.8) and (4.3). Naturally, expectations are now taken with respect to the density functions referred to above, but where now $H_S = 0$ and $S = 0$. It is interesting to note that the estimators for UNR with N unknown coincide with those suggested by Imbens (1992) for CB sampling.

4.3 Estimator Comparison

The various GMM estimators obtained above use different information, ranging from the case where only the respondents are observed, UNR, to that where data on respondents, nonrespondents and a SRS, INRS, are available. In the latter and intermediate cases, besides the data on respondents, we also use information on X provided by nonrespondents (INR) or by units of the SRS (UNRS). An analysis of the respective systems of moment indicators in (4.2)-(4.5) and (4.7)-(4.10) (and their simplified INR and UNR forms) allows both the common characteristics of the different estimators and the mechanisms by which the information on X is incorporated in the estimation procedure to be examined.

The moment indicator for H_S is identical in all cases where a SRS is present, reflecting the ancillarity of S for θ. The moment indicator for θ has two components. The first term in all cases is the score function of the RS ML (RSML) estimator for θ and, being a function of both Y and X, is only calculated for respondent units. The other term only involves X, being calculated for respondents and units of the SRS in UNRS, for

nonrespondents in INRS and INR and for respondents in UNR. The estimating functions for Q_y use information from the same units as the second term of the moment indicator for θ under UNRS and UNR, while for INRS and INR data from all units observed (respectively, those of both main sample and SRS and those of the main sample) are used. Thus, relative to UNR, GMM for INR and UNRS includes additional information on X through the second terms of the moment indicators for θ and Q_y. When most data is available, INRS, the SRS only contributes to the estimation of Q_y relative to INR.

4.4 Multiplicative Intercept Models

The GMM estimators proposed above deal with different patterns of nonresponse governed by a nonignorable missing data mechanism. In general, unless data are MCAR, conventional RS estimators applied with the complete data set are inconsistent. However, Carroll, Ruppert and Stefanski (1995, p.184) and Allison (2001, p.7), aver that, as long as the probability of response conditional on the dependent variable is independent of the covariates, precisely the missingness mechanism assumed in Assumption 2.3, then estimators for the slope parameters of logit models remain consistent in apparent contradiction to the results presented here. As is shown below, their conjecture results from the particular properties of multiplicative intercept models (MIM), which include the logit model as a particular case and are also widely discussed in the area of CB sampling. The literature on CB sampling demonstrates that, on the one hand, both intercept terms and marginal choice probabilities Q_y are not separately identified in MIM when these probabilities are unknown. On the other hand, except for the shift in intercept terms, all parameters in MIM are consistently estimated by the RSML estimator; see, for example, Hsieh, Manski and McFadden (1985) and Weinberg and Wacholder (1993).

UNR should preserve these two characteristics, since only a slight modification of the CB sampling formulation is required; see section 4.2.4. However, neither of these properties can be extended to the other cases unless incomplete units are discarded which would again reduce to UNR.

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

$$\mathcal{P}\{y|x, v_y, \theta_y^1\} = \frac{\nu_y V_y(\theta_y^1)}{\sum_{y\in\mathcal{Y}} \nu_y V_y(\theta_y^1)}. \tag{4.14}$$

where $\nu_y = v_y(\theta_y^0)$. The coefficients θ_y^0 and θ_y^1 are, respectively, the constant term and vector of slope parameters associated with alternative $Y = y$. We set $\nu_0(\theta_0^0) = 1$, $V_0(\theta_0^1) = 1$, $V_y(\theta_y^1) > 0$, $\partial\nu_y(\theta_y^0)/\partial\theta_y^0 = \nu_y(\theta_y^0)$ and $\partial V_y(\theta_y^1)/\partial\theta_y^1 = x_y V_y(\theta_y^1)$ for all y.[14]

Under UNR the identification problem for intercept terms of MIM becomes apparent because the moment indicators for the intercept parameters are perfectly correlated with those relevant for the estimation of Q_y; cf. Imbens (1992) for CB samples. The moment indicator (4.9) for θ_t^0 is

$$\theta_t^0 : r\left\{ I(y=t) - \left[\sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y}\nu_y V_y(\theta_y^1)\right]^{-1} \frac{H_t}{Q_t}\nu_t V_t(\theta_t^1)\right\}, \tag{4.15}$$

$t = 0, ..., C$. Clearly the moment indicator (4.15) coincides with H_t/Q_t times that for Q_t (4.10) plus that for H_t (4.7). Thus, identification of θ_y^0 is only possible when Q_y is known, in which case the known Q_y is substituted in the moment indicators for θ_y^0 and θ_y^1 and that for Q_y is suppressed.

The particular property of MIM which causes these identification problems allows the use of RS procedures to estimate the slope parameters θ_y^1 under UNR. This is apparent because the moment indicators for θ_t^1, given by (4.15) pre-multiplied by x_t, apart from the distortion in the intercept parameters, which are now $(H_y/Q_y)\nu_y$, $y \in \mathcal{Y}$, coincide with the RS moment indicators

$$x_t(I(y=t) - \mathcal{P}\{t|x, \nu_t, \theta_t^1\}). \tag{4.16}$$

Thus, (4.16) may be used to consistently estimate θ_t^1 under UNR.

This property, however, does not hold when 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 θ may be written in the RS form (4.16). Thus, although the RSML estimator

[14] The multinomial logit model arises when $\nu_y(\theta_y^0) = \exp(\theta_y^0)$, $V_y(\theta_y^1) = \exp(x'\theta_y^1)$ and $\theta_0^0 = 0$, $\theta_0^1 = 0$.

based on the complete sample is consistent for the slope parameters in all cases, if one wishes to include the additional information on X from the nonrespondents and/or the SRS in the estimation procedure, then the GMM estimators for UNRS, INRS or INR proposed in the previous sections must be used.

5 Specification Tests

5.1 MCAR

In general, unless data are MCAR, RSML applied to the complete sample will yield inconsistent estimators. In an application whether or not the missingness mechanism is ignorable would be unknown. If information on the population probabilities Q_y was available, a comparison of Q_y with the sampling proportion H_y might be used to draw rough conclusions about the nature of the missing data. More formally, specification tests for the null hypotheses of data MCAR may be constructed as is now described.

If the data are MCAR, both P_y and the ratio H_y/Q_y are constant for all $y \in \mathcal{Y}$; that is, $P_y = P$, $y \in \mathcal{Y}$. See the comments below (2.9) and (2.11). The MCAR null hypothesis H_0 for when the initial sample size N is known is

$$H_0 : \frac{H_y}{Q_y} = P(1 - H_S),\ y \in \mathcal{Y}, \tag{5.1}$$

and, when N is unknown,

$$H_0 : \frac{H_y}{Q_y} = 1 - H_S,\ y \in \mathcal{Y}, \tag{5.2}$$

which also corresponds to the absence of missing data.

GMM estimation under either version of H_0 using the moment indicator systems (4.2)-(4.5) or (4.7)-(4.10) is straightforward. The moment indicator for θ is identical to the score function of the RSML estimator since $\sum_{y\in\mathcal{Y}}(H_y/Q_y)\partial\mathcal{P}\{y|x,\theta\}/\partial\theta = (H_y/Q_y)$ $\sum_{y\in\mathcal{Y}}\partial\mathcal{P}\{y|x,\theta\}/\partial\theta = 0$; see (4.4) and (4.9). Additionally, the moment indicators for H_t under INRS and Q_y under UNRS utilise $\sum_{y\in\mathcal{Y}}(H_y/Q_y)\mathcal{P}\{y|x,\theta\} = H_y/Q_y$; see (4.2) and (4.10). From (3.4), the INRS conditional sample density becomes

$$h_{INRS}^{MCAR}(y,r,s|x) = \mathcal{P}\{y|x,\theta\}^{r(1-s)}\left[P^r(1-P)^{1-r}\right]^{1-s}\left[H_S^s(1-H_S)^{1-s}\right].$$

Therefore, the INRS MCAR estimators are $\tilde{P} = n/N$, $\tilde{H}_S = m/N_m$ and $\tilde{H}_y = \tilde{Q}_y\tilde{P}(1 - \tilde{H}_S)$, where, from (4.5), $\tilde{Q}_y = \sum_{i=1}^{N_m} \mathcal{P}\{y|x_i, \tilde{\theta}\}/N_m$, $y \in \mathcal{Y}$, and $\tilde{\theta}$ is the RSML estimator. For UNRS, from (3.2),

$$h_{UNRS}^{MCAR}(y, x, r, s) = \mathcal{P}\{y|x, \theta\}^{r(1-s)} \left[P^r(1-P)^{1-r}\right]^{1-s} \left[H_S^s(1-H_S)^{1-s}\right] f_X(x)^{r(1-s)+s}.$$

Therefore, the UNRS MCAR estimators are as above except, from (4.10), $\tilde{Q}_y = \sum_{i=1}^{n_m} \mathcal{P}\{y|x_i, \tilde{\theta}\}/n_m$, $y \in \mathcal{Y}$. If N is unknown, from (3.7), the UNRS sample density now becomes

$$h_{UNRS}(y, x, r = 1, s) = \mathcal{P}\{y|x, \theta\}^{r(1-s)} \left[H_S^s(1-H_S)^{r(1-s)}\right] f_X(x)^{r(1-s)+s}.$$

The UNRS estimators remain the same except that $\tilde{H}_S = m/n_m$. Let $\tilde{\varphi}$ denote the H_0 MCAR estimator for φ.

A test for data MCAR may be based on the difference of estimated GMM criteria (4.11) under null and alternative hypotheses; that is, the statistic

$$N_m \left[\hat{g}(\tilde{\varphi})'\hat{\Omega}^{-1}\hat{g}(\tilde{\varphi}) - \hat{g}(\hat{\varphi})'\hat{\Omega}^{-1}\hat{g}(\hat{\varphi})\right], \tag{5.3}$$

where $\hat{\Omega} = \sum_{i=1}^{N_m} \hat{g}(\hat{\varphi})\hat{g}(\hat{\varphi})'/N_m$ with obvious adjustments if there is no SRS and N_m replaced by n_m if N is unknown. Under the MCAR null hypothesis H_0, (5.1) or (5.2), the statistic (5.3) will converge in distribution to a chi-square random variable with respectively C and $C + 1$ degrees of freedom. See Newey and West (1987) for other asymptotically equivalent test statistics.

5.2 Missing Data Mechanism

Assumption 2.3 is crucial to the foregoing analysis. It requires that nonresponse is determined by covariates solely through the dependent variable although as explained above it may be relaxed to allow dependence on a finite partition of the covariate sample space.

Consider a general definition of the missingness mechanism

$$P_y(x) = \mathcal{P}\{R = 1|Y = y, X = x\};$$

cf. (2.3). The major alterations which are required concern the resultant specifications of the sample densities for INRS (INR) and UNRS (UNR); cf. sections 3.1 and 3.2. For INRS, $h(x|y)$ in (3.1) is replaced appropriately by $h(x|y, r = 1) = h(y, r = 1, x)/h(y, r = 1)$ or

$$\begin{aligned} h(x|y, r &= 1) = \frac{P_y(x)\mathcal{P}\{y|x,\theta\}f_X(x)}{\int_{\mathcal{X}} P_y(x)\mathcal{P}\{y|x,\theta\}f_X(x)dx} \\ &= \frac{P_y(x)}{P_y}\frac{\mathcal{P}\{y|x,\theta\}f_X(x)}{Q_y}. \end{aligned}$$

and

$$h(x|y, r = 0) = \frac{(1 - P_y(x))}{(1 - P_y)}\frac{\mathcal{P}\{y|x,\theta\}f_X(x)}{Q_y}.$$

Therefore,

$$\begin{aligned} h_{INRS}(y, x, r, s) &= \left\{\left[\frac{P_y(x)}{P_y}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}\right]^r \right. \\ &\quad \times \left.\left[\sum_{y\in\mathcal{Y}}\frac{(1 - P_y(x))}{(1 - P_y)}\left(1 - H_S - \frac{H_y}{Q_y}\right)\mathcal{P}\{y|x,\theta\}\right]^{1-r}\right\}^{1-s} \\ &\quad \times [H_S]^s f_X(x). \end{aligned} \tag{5.4}$$

Correspondingly, for UNRS with N known, cf. (3.2),

$$\begin{aligned} h_{UNRS}(y, x, r, s) &= \left\{\left[\frac{P_y(x)}{P_y}\frac{H_y}{Q_y}\mathcal{P}\{y|x,\theta\}f_X(x)\right]^r \right. \\ &\quad \times \left.\left[\sum_{y\in\mathcal{Y}}\int_{\mathcal{X}}\frac{(1 - P_y(x))}{(1 - P_y)}\left(1 - H_S - \frac{H_y}{Q_y}\right)\mathcal{P}\{y|x,\theta\}f_X(x)dx\right]^{1-r}\right\}^{1-s} \\ &\quad \times [H_S f_X(x)]^s. \end{aligned} \tag{5.5}$$

The proposed specification test is based on the Lagrange multiplier principle; see *inter alia* Newey and West (1987). Firstly, the response probabilities are parameterised as $P_y(x) = P_y(z_y'\eta_y)$ where $z_y = z_y(x)$ is a suitably chosen vector of independent functions of the covariates x and $P_y(0) = P_y$, $y \in \mathcal{Y}$. Secondly, log-likelihoods are constructed based on the sample densities (5.4) and (5.5); cf. (4.1) and (4.6) respectively. Thirdly, the moment indicators corresponding to η_y are obtained by differentiating the resultant

log-likelihoods and evaluation at $\eta_y = 0$, $y \in \mathcal{Y}$. For INRS:

$$\eta_t : P_y'(0)(1 - H_S)\frac{Q_t}{H_t} z_t \left((1-s)\, r I(y = t) - \frac{H_t}{Q_t} \frac{(1-s)(1-r)\mathcal{P}\{t|x,\theta\}}{1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \mathcal{P}\{y|x,\theta\}} \right), \; t \in \mathcal{Y},$$

where $P_y'(.)$ denotes the derivative of $P_y(.)$ with respect to its argument. Similarly for UNRS:

$$\eta_t : P_y'(0)(1 - H_S)\frac{Q_t}{H_t} z_t \left((1-s)\, r I(y = t) - H_t \right), \; t \in \mathcal{Y}.$$

Because the multiplicative factors $P_y'(0)(1 - H_S)(Q_t/H_t)$ are observation invariant they may be omitted so that the relevant moment indicator becomes z_t multiplied by the moment indicator for H_t, (4.2) or (4.7). For UNRS with N unknown an appropriate moment indicator is defined by analogy as that for UNRS with N known.[15]

Let $q(\varphi)$ define the vector of moment indicators obtained from $g(\varphi)$ defined in section 4.2 augmented by those given above for η_y, $y \in \mathcal{Y}$. Let $Q(\varphi) = \partial q(\varphi)/\partial \varphi'$. Correspondingly, define $\hat{q}(\varphi) = \sum_{i=1}^{N_m} q_i(\varphi)/N_m$, $\hat{Q}(\varphi) = \sum_{i=1}^{N_m} Q_i(\varphi)/N_m$ and $\hat{\Sigma}(\varphi) = \sum_{i=1}^{N_m} q_i(\varphi) q_i(\varphi)'/N_m$. Therefore, a GMM Lagrange multiplier specification test for Assumption 2.3 is given by

$$\mathcal{LM} = N_m \hat{q}(\hat{\varphi})' \hat{\Sigma}(\hat{\varphi})^{-1} \hat{Q}(\hat{\varphi}) \left(\hat{Q}(\hat{\varphi})' \hat{\Sigma}(\hat{\varphi})^{-1} \hat{Q}(\hat{\varphi}) \right)^{-1} \hat{Q}(\hat{\varphi}) \hat{\Sigma}(\hat{\varphi})^{-1} \hat{q}(\hat{\varphi});$$

cf. Newey and West (1987). For INR and UNR, N_m is replaced by N. Hence, $\hat{q}(\varphi) = \sum_{i=1}^{N} q_i(\varphi)/N$, $\hat{Q}(\varphi) = \sum_{i=1}^{N} Q_i(\varphi)/N$ and $\hat{\Sigma}(\varphi) = \sum_{i=1}^{N} q_i(\varphi) q_i(\varphi)'/N$. For UNR(S) and N unknown, the form of statistic is as for UNR(S) except N (N_m) is replaced by n (n_m). If Assumption 2.3 is satisfied $\mathcal{LM}$ has a limiting chi-square distribution with degrees of freedom given by $\sum_{y \in \mathcal{Y}} \dim(\eta_y)$.

[15] With N unknown, cf. (3.7),

$$h_{\mathsf{UNRS}}(y, x, r = 1, s) = \left[\frac{P_{\mathsf{y}}(x)}{P_{\mathsf{y}}} \frac{H_{\mathsf{y}}}{Q_{\mathsf{y}}} \mathcal{P}\{y|x,\theta\} f_{\mathsf{X}}(x) \right]^{\mathsf{r}(1-\mathsf{s})} \left[H_{\mathsf{S}} f_{\mathsf{X}}(x) \right]^{\mathsf{S}}.$$

However, the response probabilities P_{y} and $P_{\mathsf{y}}(x)$ are unidentified when N is unknown; see section 2.4.2.

6 Simulation Evidence

This section presents a simulation study based on a Probit model in order to investigate the performance in practice of some of the estimators developed in previous sections. Section 6.1 describes the experimental design and section 6.2 discusses the results.

6.1 Experimental Design

All experiments consider binary data; thus $\mathcal{Y} = \{0, 1\}$. The variable of interest, Y, conditional on the scalar covariate $X = x$, is generated by the Probit model characterized by $\mathcal{P}\{1|x, \theta\} = \Phi(x\theta)$ where $\Phi(.)$ denotes the standard normal distribution function. The scalar covariate X has mean 3 and variance 4 and is generated as a mixture of normally distributed variates, $N(2, 1.2915)$ with probability 0.7 and $N(5.333, 1.2915)$ with probability 0.3. To obtain a population probability $Q_1 = 0.75$ of observing $Y = 1$, we set the true value $\theta_0 = 0.251$. The initial sample size is $N = 300$ throughout.

Five experimental designs characterized by different ratios $P^* = P_1/P_0$ are analysed. The size of the SRS is $m = N - n$, so that the improvements due to combining information on X from this independent sample (under UNRS) or from the same number of nonrespondents from the initial sample (under INR) may be compared. The different combinations of P_1 and P_0 produce different proportions of individuals responding $Y = 1$ ($Y = 0$) in the main sample and the SRS, H_1 (H_0) and H_S respectively.[16] Table 1 summarizes the main characteristics of the five experimental designs. For comparison purposes, the first experimental design, Experiment a, contains no missing values. The number of incomplete responses $N-n$ is increased from Experiment b to Experiment c, as well as the differential between P_1 and P_0. Experiment d considers a relatively large ratio P^*, which is a little smaller than that in Experiment b, associated with a small complete sample size ($n = 120$ as in Experiment 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

[16] From (2.9), $H_1 = P_1 Q_1 (1 - H_S)$ and $H_0 = P_0 (1 - Q_1)(1 - H_S)$.

computations were done using S-Plus. Each experiment uses 1000 replications.

Table 1 about here

Eight estimators were compared. Four assume the absence of information on Q_1, including the RSML estimator (RSMLE), which uses the complete data set, and the estimators proposed in this paper for UNR (UNRE), UNRS (UNRSE) and INR (INRE). The remaining four estimators incorporate information on Q_1 and are denoted by QRSMLE, QUNRE, QUNRSE and QINRE. Table 2 lists the moment indicators for INRS and UNRS which are obtained from, respectively, (4.2)-(4.5) and (4.7)-(4.10). The moment indicators for INR and UNR use the simplifications described in section 4.2 and the moment indicator for θ for RSMLE is that in UNR with the second term suppressed. For INRS (INR) rather than use the efficient moment indicator (4.2), the simpler indicators $(1-s)\,rI(y=t)-H_t,\ t\in\{0,1\}$, as in UNRS (UNR), are implemented.

Table 2 about here

6.2 Results

Summary statistics 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 various estimators of θ. Figures 1 and 2 show the estimated sampling distributions of these estimators for Experiments b, c and d. Figure 1 considers the four estimators in which information on Q_1 is not used (RSMLE, UNRE, UNRSE and INRE) together with QUNRE which, among the estimators which use the known value of Q_1, gave the worst performance. Figure 2 illustrates the behaviour of RSMLE compared with that of the four estimators where Q_1 is known, QRSMLE, QUNRE, QUNRSE and QINRE.

Table 3 about here

Figures 1 and 2 about here

As expected, RSMLE performs well in Experiment a, where there are no missing values, and Experiment e, where data are MCAR. In these experiments the incorporation

of aggregate information on Q_1 reduces the standard deviations across replications by more than 50%. For Experiments b, c and d, however, where the probability of response differs for $Y = 1$ and $Y = 0$, RSMLE suffers from substantial mean and median biases. These biases are less in Experiments b and d, which are characterized by a ratio P^* close to 1, and are, thus, closer to a MCAR pattern. But even in these cases, the biases are still unacceptably high, being of the order of at least 10%. In these three experiments, the incorporation of information on Q_1 produces substantial improvements. Except for Experiment c, where P^* is very small, QRSMLE displays relatively small mean and median distortions, which are smaller than some of those for the GMM estimators UNRE and UNRSE, and a variability similar to that of GMM estimators when Q_1 is known; see also Figure 2. These results are especially interesting. Imbens and Lancaster (1994) show that combining macro and micro information results in more efficient estimators. Clearly, as these experiments reveal, an improvement in efficiency is not the only advantage of their proposal, since it also produces more robust estimators in the presence of nonignorable nonresponse.[17]

All of the GMM estimators proposed to deal with nonignorable missing data perform relatively well. None of their results appears to be strongly affected by the experimental design, apart from some adverse effects on estimator variability when the complete sample size n is reduced, which become more serious when P^* is close to 1 in Experiment d. In effect, the mean and median biases of the GMM estimators are small; see also Figures 1 and 2, in which the estimated densities for these estimators are always centrally located around the true value of $\theta_0 = 0.251$. Table 3 and these Figures show that the inclusion of information on X from nonrespondents and units of the SRS is only relevant when Q_1 is unknown. In fact, while UNRSE and INRE exhibit, in general, better results than UNRE, especially when P^* is reduced, the mean and the median biases are very similar for QUNRE, QUNRSE, and QINRE. On the other hand, the inclusion of information on Q_1 only appears to substantially ameliorate bias under UNR, the case where X is only

[17]Similar conclusions were also reached in simulation studies conducted in Ramalho (2001, 2002). These experiments concerned problems of misclassification in the reponse variable under CB sampling and measurement error in the covariates.

measured for respondents.

Standard deviations across replications reported in Table 3 exemplify the improvements due to the knowledge of Q_1. The estimators QUNRE, QUNRSE and QINRE, relative to their respective versions with Q_1 estimated, have dispersion reduced by at least 31%. Figure 1 reinforces this observation; compare QUNRE in Figure 1 with UNRE, UNRSE and INRE. Moreover, for Experiments b, c and d, standard deviations across replications are reduced for both UNRSE and INRE relative to UNRE, and QUNRSE and QINRE relative to QUNRE, which results from including information on X from the incomplete questionnaires. These improvements are more considerable in Experiment c, where the differential between P_1 and P_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_1 is unknown the reductions in variability are more significant for INRE than UNRSE. Thus, as the sample size of the SRS in UNRS equals the number of nonrespondents in INR in our experiments, we may conclude that the observations on X contributed by nonrespondents are more informative than those from the SRS. Finally, it is also worth noting that RSMLE underestimates the variability of the data, which is a common feature if a sampling problem, not only nonresponse but also several forms of measurement error, is ignored; see, for example, Hausman et al. (1998) and Chesher (1998), who examine two different forms of measurement error.

Additionally, the ratio P^* for cases where Q_1 is unknown was estimated using (2.9) for experiments with missing data (Experiments b, c, d and e). Mean and the median biases in percentage terms and standard deviations across the replications are presented in Table 4. The conclusions for the P^* estimates are similar to those for the estimators of θ. The mean and median biases are small and worst 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^*: standard deviations are smaller in Experiment c, with the smallest P^*, and then increase dramatically in the other cases, especially in Experiment d, where a relatively large value of P^* is associated

with a small complete sample size n.

Table 4 about here

These experiments show the importance of using all the available information in the estimation procedure. Undoubtedly, aggregate information on Q_1 is the major source of improvement followed by data on X from incomplete responses, and, finally, data on X from a SRS. The use of one or other of these two last forms of information when Q_1 is available does not appear to offer any advantage. However, the incorporation of known Q_1 appears to be beneficial in all cases.

7 Conclusion

This paper considers several nonignorable missing data problems when the dependent variable is discrete. A unified GMM estimation and inference methodology is proposed for such circumstances which adapts and extends that usually employed with choice based sampling. The advantages of an integrated approach are obvious. The same methodology is employed for both model specification and estimator derivation in all cases. Additionally, it also allows the investigation of and comparison between the different nonresponse problems. The nonresponse pattern, INRS, encompasses all the rest. Discarding the information on covariates provided by nonrespondents and individuals in the SRS, respectively, UNRS and INR, are straightforwardly obtained. The additional suppression of similar information from these two cases yields UNR. Moreover, for a MIM structural model, RS estimation methods using the complete data set can be employed in all cases for consistent estimation of the slope parameters, but at the expense of the loss of information on the covariates from nonrespondents and/or SRS units.

The critical assumption in our framework, besides the correct specification of the structural model, is that the probability of response conditional on the dependent variable and covariates is independent of the covariates. This assumption might be expected to relevant in many practical situations. In cases of INR (and INRS), it is not necessarily too unreasonable to assume that covariates influence the choice variable and the

willingness to report that choice in a similar fashion. Under UNR (and UNRS), this assumption is likely to be appropriate in cases where the refusal to participate in the survey is especially motivated by an unwillingness to reveal the value of the choice variable. We suggest how this assumption may be weakened to allow the response mechanism to depend additionally on a finite partition of the covariate sample space. Specification tests are presented both for MCAR and the missingness assumption.

A small simulation study revealed very promising results. The GMM estimators suggested here display negligible bias, which is especially apparent in cases where data on the covariates, from either nonrespondents or units of a SRS, are incorporated in the estimation procedure. In contradistinction, RSML estimators are considerably biased in all cases where response rates across the alternatives are different, even in experiments where this differential was not very substantial. The incorporation of aggregate information on the marginal population choice probabilities only greatly improved the properties of both the proposed GMM estimators and the RS estimators based on the complete data set when response was nonignorable.

Appendix A: Derivation of Moment Indicators

A.1 INRS

Let $\mathcal{L}$ denote the Lagrangean arising from (4.1) with μ as the Lagrange multiplier associated with the constraint $\sum_{l=1}^{L}\pi^l = 1$. The resultant first order derivatives are

$$\frac{\partial \mathcal{L}}{\partial H_y} = \sum_{i=1}^{N_m}(1-s_i)\left[\frac{r_i I(y_i=y)}{H_y} - \frac{(1-r_i)}{1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x^{l_i},\theta\}}\frac{\mathcal{P}\{y|x^{l_i},\theta\}}{Q_y}\right], \quad \text{(A.1)}$$

$$\frac{\partial \mathcal{L}}{\partial H_S} = \sum_{i=1}^{N_m}\left[\frac{s_i}{H_S} - \frac{(1-s_i)(1-r_i)}{1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x^{l_i},\theta\}}\right], \quad \text{(A.2)}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{N_m}(1-s_i)\Bigg\{ & r_i\left[\frac{\partial \log \mathcal{P}\{y_i|x^{l_i},\theta\}}{\partial\theta} - \frac{1}{Q_{y_i}}\sum_{l=1}^{L}\pi_l\frac{\partial \mathcal{P}\{y_i|x^l,\theta\}}{\partial\theta}\right] \\ & -\frac{(1-r_i)}{1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x^{l_i},\theta\}} \\ & \times\left[\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\frac{\partial \mathcal{P}\{y|x^{l_i},\theta\}}{\partial\theta} - \sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y^2}\mathcal{P}\{y|x^{l_i},\theta\}\sum_{l=1}^{L}\pi_l\frac{\partial \mathcal{P}\{y|x^l,\theta\}}{\partial\theta}\right]\Bigg\}, \end{aligned} \quad \text{(A.3)}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_l} = \sum_{i=1}^{N_m}(1-s_i)\Bigg\{ & r_i\left[\frac{I(l_i=l)}{\pi_l} - \frac{1}{Q_{y_i}}\mathcal{P}\{y_i|x^l,\theta\}\right] \\ & (1-r_i)\left[\frac{\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y^2}\mathcal{P}\{y|x^{l_i},\theta\}\mathcal{P}\{y|x^l,\theta\}}{1-H_S-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x^{l_i},\theta\}} + \frac{I(l_i=l)}{\pi_l}\right]\Bigg\} + \sum_{i=1}^{N_m}s_i\frac{I(l_i=l)}{\pi_l} - \mu, \end{aligned} \quad \text{(A.4)}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{l=1}^{L}\pi_l - 1, \quad \text{(A.5)}$$

where $y \in \mathcal{Y}$ and $l = 1, ..., L$.

Equating (A.1) to zero and solving yields the ML estimator for H_y

$$\hat{H}_y = n_y\hat{Q}_y\left[\sum_{i=1}^{N_m}\frac{(1-s_i)(1-r_i)\mathcal{P}\{y|x^{l_i},\hat{\theta}\}}{1-\hat{H}_S-\sum_{y\in\mathcal{Y}}\frac{\hat{H}_y}{\hat{Q}_y}\mathcal{P}\{y|x^{l_i},\hat{\theta}\}}\right]^{-1}, \quad \text{(A.6)}$$

where $\hat{Q}_y = \sum_{l=1}^{L}\hat{\pi}_l\mathcal{P}\{y|x^l,\hat{\theta}\}$. The ancillary statistic $\hat{H}_S = m/N_m$ is the ML estimator for H_S and is obtained by multiplying (A.1) by H_y, summing over y and then equating the resultant expression and (A.2) to zero.

The mass point probabilities π_l, $l = 1, ..., L$, can be concentrated out, thus removing the dependence on the discrete distribution of X; cf. Imbens (1992). Firstly, note that, from (A.6), the second and third terms in (A.4) sum to zero. Secondly, multiplying (A.4) by $\hat{\pi}_l$ and summing over $l = 1, ..., L$ yields

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{N_m} \{(1 - s_i)\,[(r_i + (1 - r_i))] + s_i\} \sum_{l=1}^{L} I(l_i = l) \\ &= N_m. \end{aligned}$$

Substituting for $\hat{\mu}$ in (A.4),

$$\hat{\pi}_l = \frac{1}{N_m} \sum_{i=1}^{N_m} I(l_i = l), \tag{A.7}$$

which is the the usual nonparametric ML estimator for a probability mass point at each of L points of support; see, for example, Cosslett (1997). Note that X is observed for all units and, thus, $h_{INRS}(x)$ of (3.3) coincides with $f_X(x)$. Hence, the ML estimator for Q_y is given from (2.2) by

$$\begin{aligned} \hat{Q}_y &= \sum_{l=1}^{L} \hat{\pi}_l \mathcal{P}\{y|x^l, \hat{\theta}\} \qquad (A.8) \\ &= \frac{1}{N_m} \sum_{i=1}^{N_m} \mathcal{P}\{y|x^{l_i}, \hat{\theta}\},\ y \in \mathcal{Y}. \end{aligned}$$

Substitution for $\hat{\pi}_l$ in (A.3) from (A.7) results in the second and fourth terms summing to zero after using (A.6). Therefore, (A.3) becomes

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{i=1}^{N_m} (1 - s_i) \left\{ r_i \frac{\partial \log \mathcal{P}\{y_i|x^{l_i}, \theta\}}{\partial \theta} - \right. \qquad (A.9) \\ &\quad \left. (1 - r_i) \left[1 - H_S - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \mathcal{P}\{y|x^{l_i}, \theta\} \right]^{-1} \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \frac{\partial \mathcal{P}\{y|x^{l_i}, \theta\}}{\partial \theta} \right\}. \end{aligned}$$

A.2 UNRS

The Lagrangean $\mathcal{L}$ now arises from (4.6). The resultant first order derivatives are

$$\frac{\partial \mathcal{L}}{\partial H_y} = \sum_{i=1}^{N_m} (1 - s_i) \left[\frac{r_i I(y_i = y)}{H_y} - \frac{(1 - r_i)}{1 - H_S - \sum_{y \in \mathcal{Y}} H_y} \right], \tag{A.10}$$

$$\frac{\partial \mathcal{L}}{\partial H_S} = \sum_{i=1}^{N_m} \left[\frac{s_i}{H_S} - \frac{(1 - s_i)\,(1 - r_i)}{1 - H_S - \sum_{y \in \mathcal{Y}} H_y} \right], \tag{A.11}$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{N_m} (1-s_i)\, r_i \left[\frac{\partial \log \mathcal{P}\{y_i|x^{l_i},\theta\}}{\partial \theta} - \frac{1}{Q_{y_i}} \sum_{l=1}^{L} \pi_l \frac{\partial \mathcal{P}\{y_i|x^l,\theta\}}{\partial \theta} \right], \tag{A.12}$$

$$\frac{\partial \mathcal{L}}{\partial \pi_l} = \sum_{i=1}^{N_m} \left\{ (1-s_i)\, r_i \left[\frac{I(l_i=l)}{\pi_l} - \frac{1}{Q_{y_i}} \mathcal{P}\{y_i|x^l,\theta\} \right] + s_i \frac{I(l_i=l)}{\pi_l} \right\} - \mu, \tag{A.13}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{l=1}^{L} \pi_l - 1, \tag{A.14}$$

where $y \in \mathcal{Y}$ and $l = 1, ..., L$.

Equating (A.10) and (A.11) to zero, we obtain the ancillary statistics $\hat{H}_y = n_y/N_m$ and $\hat{H}_S = m/N_m$ as ML estimators for H_y and H_S respectively.

Similarly to INRS in Appendix A.1, the dependence on the discrete distribution for X may be removed. Recall that the ML estimator for Q_y from (2.2) is $\hat{Q}_y = \sum_{l=1}^{L} \hat{\pi}_l \mathcal{P}\{y|x^l,\hat{\theta}\}$. Hence, multiplying (A.13) by $\hat{\pi}_l$ and summing over $l = 1, ..., L$ yields

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{N_m} (1-s_i)\, r_i \left[\sum_{l=1}^{L} I(l_i=l) - \frac{\sum_{l=1}^{L} \hat{\pi}_l \mathcal{P}\{y_i|x^l,\hat{\theta}\}}{\sum_{l=1}^{L} \hat{\pi}_l \mathcal{P}\{y_i|x^l,\hat{\theta}\}} \right] \\ &\quad + \sum_{i=1}^{N_m} s_i \sum_{l=1}^{L} I(l_i=l) = m = N_m \hat{H}_S. \end{aligned}$$

Replacing $\hat{\mu}$ in (A.13) and solving,

$$\begin{aligned} \hat{\pi}_l &= \frac{1}{N_m} \sum_{i=1}^{N_m} \left[(1-s_i)\, r_i + s_i \right] I(l_i=l) \left[\hat{H}_S + \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{(1-s_i)\, r_i}{\hat{Q}_{y_i}} \mathcal{P}\{y_i|x^l,\hat{\theta}\} \right]^{-1} \\ &= \frac{1}{N_m} \sum_{i=1}^{N_m} \left[(1-s_i)\, r_i + s_i \right] I(l_i=l) \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^l,\hat{\theta}\} \right]^{-1}. \end{aligned} \tag{A.15}$$

The estimator (A.15) for π_l reflects the distortion of $h_{UNRS}(x)$ in (3.5) relative to $f_X(x)$ induced by the pattern of nonresponse. If there were no missing values in the main sample, then $\hat{\pi}_l$ would equal the usual nonparametric ML estimator $\sum_{i=1}^{N_m} I(l_i=l)/N_m$; cf. (A.7) above. This result obtains since $R = 1$ for all units of the main sample. So $\sum_{y \in \mathcal{Y}} \hat{H}_y \mathcal{P}\{y|x^l,\hat{\theta}\}/\hat{Q}_y = (\hat{H}_y/\hat{Q}_y) \sum_{y \in \mathcal{Y}} \mathcal{P}\{y|x^l,\hat{\theta}\} = (1-\hat{H}_S) \sum_{y \in \mathcal{Y}} \mathcal{P}\{y|x^l,\hat{\theta}\} = 1-\hat{H}_S$, the first and second equalities resulting from $P_y = 1$ in eq. (2.9). Substituting $\hat{\pi}_l$ of (A.15) in the last term of (A.12),

$$\sum_{i=1}^{N_m} (1-s_i)\, r_i \frac{1}{\hat{Q}_{y_i}} \sum_{l=1}^{L} \hat{\pi}_l \frac{\partial \mathcal{P}\{y_i|x^l,\hat{\theta}\}}{\partial \theta} =$$

$$
\begin{aligned}
&= \sum_{i=1}^{N_m} \frac{(1-s_i)\, r_i}{\hat{Q}_{y_i}} \sum_{l=1}^{L} \left[\frac{1}{N_m} \sum_{j=1}^{N_m} [(1-s_j)\, r_j + s_j]\, I(l_j = l) \right. \\
&\quad \left. \times \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^l, \hat{\theta}\} \right]^{-1} \right] \frac{\partial \mathcal{P}\{y_i|x^l, \hat{\theta}\}}{\partial \theta} \\
&= \sum_{i=1}^{N_m} \frac{(1-s_i)\, r_i}{\hat{Q}_{y_i}} \frac{1}{N_m} \sum_{j=1}^{N_m} [(1-s_j)\, r_j + s_j] \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^{l_j}, \hat{\theta}\} \right]^{-1} \\
&\quad \times \frac{\partial \mathcal{P}\{y_i|x^{l_j}, \hat{\theta}\}}{\partial \theta} \\
&= \sum_{j=1}^{N_m} [(1-s_j)\, r_j + s_j] \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^{l_j}, \hat{\theta}\} \right]^{-1} \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{(1-s_i)\, r_i}{\hat{Q}_{y_i}} \\
&\quad \times \frac{\partial \mathcal{P}\{y_i|x^{l_j}, \hat{\theta}\}}{\partial \theta} \\
&= \sum_{i=1}^{N_m} [(1-s_i)\, r_i + s_i] \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^{l_i}, \hat{\theta}\} \right]^{-1} \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \frac{\partial \mathcal{P}\{y|x^{l_i}, \hat{\theta}\}}{\partial \theta}.
\end{aligned}
$$

The ML estimator $\hat{Q}_y$ becomes

$$
\begin{aligned}
\hat{Q}_y &= \sum_{l=1}^{L} \frac{1}{N_m} \sum_{i=1}^{N_m} [(1-s_i)\, r_i + s_i]\, I(l_i = l) \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^l, \hat{\theta}\} \right]^{-1} \mathcal{P}\{y|x^l, \hat{\theta}\} \\
&= \frac{1}{N_m} \sum_{i=1}^{N_m} [(1-s_i)\, r_i + s_i] \left[\hat{H}_S + \sum_{y \in \mathcal{Y}} \frac{\hat{H}_y}{\hat{Q}_y} \mathcal{P}\{y|x^{l_i}, \hat{\theta}\} \right]^{-1} \mathcal{P}\{y|x^{l_i}, \hat{\theta}\}. \qquad \text{(A.16)}
\end{aligned}
$$

Appendix B: Semiparametric Efficiency

Following Imbens (1992), estimator efficiency, when the exact value Q_y is known or unknown, 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, converge to the asymptotic covariance matrix of our semiparametric estimators. For reasons of expositional simplicity only we confine attention here to UNRS and UNR.

To construct the sequence of parametric models recall that X has density $f_X(.)$ defined on $\mathcal{X}$. For any $\varepsilon > 0$, partition $\mathcal{X}$ into L_ε subsets $\mathcal{X}_l$, $l = 1, ..., L_\varepsilon$, where $\mathcal{X}_l \cap \mathcal{X}_m = \emptyset$ if $l \neq m$ and $\|x - z\| < \varepsilon$ if $x, z \in \mathcal{X}_l$. Define $\phi_l(x) = 1$ if $x \in \mathcal{X}_l$ and 0 otherwise

and $f_X^{\varepsilon}(x) = f_X(x)\left[\sum_{l=1}^{L_\varepsilon}\phi_l(x)\int_{\mathcal{X}_l} f_X(x)dx\right]^{-1}$. Define the parameters $\delta_l = \mathcal{P}\{x \in \mathcal{X}_l\} = \int_{\mathcal{X}_l} f_X(x)dx$, $l = 1, ..., L_\varepsilon$.

Under UNRS, the sequence of parametric models indexed by ε, and which result from substituting for $f_X(x)$ in (3.2), is

$$\begin{aligned} h_{UNRS}^{\varepsilon}(y,x,r,s) &= \left\{\left[H_y \frac{\mathcal{P}\{y|x,\theta\}f_X^{\varepsilon}(x)\sum_{l=1}^{L_\varepsilon}\phi_l(x)\delta_l}{\sum_{l=1}^{L_\varepsilon}\delta_l\int_{\mathcal{X}_l}\mathcal{P}\{y|x,\theta\}f_X^{\varepsilon}(x)\phi_l(x)dx}\right]^r \left(1 - H_S - \sum_{y\in\mathcal{Y}} H_y\right)^{1-r}\right\}^{1-s} \\ &\quad \times \left(H_S f_X^{\varepsilon}(x)\sum_{l=1}^{L_\varepsilon}\phi_l(x)\delta_l\right)^s, \end{aligned}$$

where $f_X^{\varepsilon}(x)$ is a known function and H_y, $y \in \mathcal{Y}$, H_S, θ and δ_l, $l = 1, ..., L_\varepsilon$ are the unknown parameters.

The ML estimator for Q_y from (2.2) is

$$\hat{Q}_y = \sum_{l=1}^{L_\varepsilon}\hat{\delta}_l\phi_l(x)\int_{\mathcal{X}_l}\mathcal{P}\{y|x,\hat{\theta}\}f_X^{\varepsilon}(x)dx.$$

Hence, the dependence of the likelihood equations obtained from $h_{UNRS}^{\varepsilon}(y,x,r,s)$ on δ_l may be removed by the same procedure employed to remove dependence on $\hat{\pi}_l$ in the system (A.10)-(A.14). The resultant score vector is described by the moment indicators

$$H_t \;:\; (1-s)\,rI(y=t) - H_t, \tag{B.1}$$

$$H_S \;:\; s - H_S, \tag{B.2}$$

$$\begin{aligned} \theta \;:\; & (1-s)\,r\frac{\partial \log \mathcal{P}\{y|x,\theta\}}{\partial\theta} - \\ & [(1-s)\,r+s]\left[H_S + \sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\sum_{l=1}^{L_\varepsilon}\phi_l(x)\int_{\mathcal{X}_l}\mathcal{P}\{y|x,\theta\}f_X^{\varepsilon}(x)dx\right]^{-1} \\ & \times\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\sum_{l=1}^{L_\varepsilon}\phi_l(x)\int_{\mathcal{X}_l}\frac{\partial\mathcal{P}\{y|x,\theta\}}{\partial\theta}f_X^{\varepsilon}(x)dx, \end{aligned} \tag{B.3}$$

$$\begin{aligned} Q_y \;:\; & Q_y - \\ & [(1-s)\,r+s]\left[H_S + \sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\sum_{l=1}^{L_\varepsilon}\phi_l(x)\int_{\mathcal{X}_l}\mathcal{P}\{y|x,\theta\}f_X^{\varepsilon}(x)dx\right]^{-1} \\ & \times\sum_{l=1}^{L_\varepsilon}\phi_l(x)\int_{\mathcal{X}_l}\mathcal{P}\{y|x,\theta\}f_X^{\varepsilon}(x)dx. \end{aligned} \tag{B.4}$$

Define the expectations $\mathcal{E}_\varepsilon[\mathcal{P}\{y|x,\theta\}] = \sum_{l=1}^{L_\varepsilon} \phi_l(x) \int_{\mathcal{X}_l} \mathcal{P}\{y|x,\theta\} f_X^\varepsilon(x)dx$ and $\mathcal{E}_\varepsilon[\partial \mathcal{P}\{y|x,\theta\}/\partial\theta]$ and $\mathcal{E}_\varepsilon[\partial^2\mathcal{P}\{y|x,\theta\}/\partial\theta\partial\theta']$ similarly. Therefore, the score vector based on the system of moment indicators (B.1)-(B.4) corresponds to (4.7)-(4.10) with $\mathcal{P}\{y|x,\theta\}$ and $\partial\mathcal{P}\{y|x,\theta\}/\partial\theta$ replaced by their respective expectations.

Continuous differentiability of $\mathcal{P}\{y|x,\theta\}$, $\partial\mathcal{P}\{y|x,\theta\}/\partial\theta$ and $\partial^2\mathcal{P}\{y|x,\theta\}/\partial\theta\partial\theta'$ in x implies uniform convergence of $\mathcal{E}_\varepsilon[\mathcal{P}\{y|x,\theta\}]$, $\mathcal{E}_\varepsilon[\partial\mathcal{P}\{y|x,\theta\}/\partial\theta]$ and $\mathcal{E}_\varepsilon[\partial^2\mathcal{P}\{y|x,\theta\}/\partial\theta\partial\theta']$ to $\mathcal{P}\{y|x,\theta\}$, $\partial\mathcal{P}\{y|x,\theta\}/\partial\theta$ and $\partial^2\mathcal{P}\{y|x,\theta\}/\partial\theta\partial\theta'$ respectively. Let $\Omega_\varepsilon = \mathcal{E}_\varepsilon[g^\varepsilon(\varphi)g^\varepsilon(\varphi)']$ and $G_\varepsilon = \mathcal{E}_\varepsilon[\partial g^\varepsilon(\varphi)/\partial\varphi']$ where $g^\varepsilon(\varphi)$ stacks the moment indicators (B.1)-(B.4). When Q_y, $y \in \mathcal{Y}$, are unknown, $\lim_{\varepsilon\to 0}\Omega_\varepsilon = \Omega$ and $\lim_{\varepsilon\to 0} G_\varepsilon = G$. Thus, the asymptotic variance matrix $G_\varepsilon^{-1}\Omega_\varepsilon G_\varepsilon'^{-1}$, which is the Cramér-Rao lower bound for the parametric estimator defined by (B.1)-(B.4), also converges to $G^{-1}\Omega G'^{-1}$, the asymptotic variance matrix of the GMM estimator. Therefore, the GMM estimator is semiparametrically efficient. Analogously, in the presence of exact information on Q_y, a suitable re-definition of Ω_ε and G_ε allows a similar conclusion to be reached, since the asymptotic variance matrix $(G_\varepsilon'\Omega_\varepsilon^{-1}G_\varepsilon)^{-1}$ of the ML estimator converges to $(G'\Omega^{-1}G)^{-1}$.

References

Allison, P.D. (2001), *Missing Data*, Sage University Papers Series on Quantitative Applications in the Social Sciences.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall.

Chesher, A. (1998), "Measurement error bias reduction", Discussion Paper 98/449, Department of Economics, University of Bristol.

Cosslett, S. (1981a), "Efficient estimation of discrete-choice models", in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 51-111.

Cosslett, S. (1981b), "Maximum likelihood estimator for choice-based samples", *Econometrica*, 49, 1289-1316.

Cosslett, S.R. (1997), "Nonparametric maximum likelihood methods", in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics, volume 15*, Elsevier Science Publishers, 385-404.

Cox, D.R., and D.V. Hinkley (1974), *Theoretical Statistics*, Chapman Hall: London.

Fitzgerald, J., Gottschslk, P. and Moffit, R. (1997), "An analysis of sample attrition in panel data: the Michigan Panel Study of Income Dynamics", *Journal of Human Resourses*, 33, 251-299.

Hausman, J.A., Abrevaya, F. and Scott-Morton, F.M. (1998), "Misclassification of the dependent variable in a discrete-response setting", *Journal of Econometrics,* 87, 239-269.

Hausman, J. and Wise, D. (1981), "Stratification on endogenous variables and estimation: the Gary Income Maintenance Experiment", in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 365-391.

Heckman, J.J. (1976), "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models", *Annals of Economic and Social Measurement*, 5, 475-492.

Hirano, K., Imbens, G.W., Ridder, G. and Rubin, D.B. (1998), "Combining panel data sets with attrition and refreshment samples", *Econometrica*, 69, 1645-1659.

Horowitz, J.L. and Manski, C.F. (1995), "Identification and robustness with contaminated and corrupted data", *Econometrica*, 63, 281-302.

Horowitz, J.L. and Manski, C.F. (1998), "Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations", *Journal of Econometrics*, 84, 37-58.

Horowitz, J.L. and Manski, C.F. (2001), "Imprecise identification from incomplete data", Working Paper.

Hsieh, D.A., Manski, C.F. and McFadden, D. (1985), "Estimation of response probabilities from augmented retrospective observations", *Journal of the American Statistical Association,* 80, 651-662.

Imbens, G. (1992), "An efficient method of moments estimator for discrete choice models with choice-based sampling", *Econometrica*, 60, 1187-1214.

Imbens, G.W. and Lancaster, T. (1994), "Combining micro and macro data in microeconometric models", *Review of Economic Studies*, 61, 655-680.

Imbens, G.W. and Lancaster, T. (1996), "Efficient estimation and stratified sampling", *Journal of Econometrics*, 74, 289-318.

Lancaster, T. and Imbens, G. (1996), "Case-control studies with contaminated controls", *Journal of Econometrics*, 71, 145-160.

Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999), "Semiparametric methods for response-selective and missing data problems in regression", *Journal of the Royal Statistical Society*, Series B, 61, 413-438.

Li, G. and Qin, J. (1998), "Semiparametric likelihood-based inference for biased and truncated data when the total sample size is known", *Journal of the Royal Statistical Society*, Series B, 60, 243-254.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical analysis with missing data*, John Wiley & Sons.

Manski, C. and Lerman, S. (1977), "The estimation of choice probabilities from choice based samples", *Econometrica*, 45, 1977-1988.

Newey, W. and McFadden, D. (1994), "Large sample estimation and hypothesis testing", in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, volume IV, Elsevier Science Publishers, 2111-2245.

Newey, W.K. and K.D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

Ramalho, E.A. (2001), "Covariate measurement error in endogenous stratified samples", mimeo.

Ramalho, E.A. (2002), "Regression models for choice-based samples with misclassification in the response variable", *Journal of Econometrics*, 106, 171-201.

Ridder, G. (1990), "Attrition in multi-wave panel data", in J. Hartog, G. Ridder and J. Theeuwes (eds.), *Panel Data and Labour Market Studies*, Elsevier Science Publishers, North-Holland, 45-68.

Rubin, D.B. (1976), "Inference and missing data", *Biometrika*, 63(3), 581-592.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Weinberg, C.R. and Wacholder, S. (1993), "Prospective analysis to case-control data under general multiplicative-intercept risk models", *Biometrika*, 80, 461-465.

Wooldridge, J.M. (1999), "Asymptotic properties of weighted m-estimators for variable probability samples", *Econometrica*, 67, 1385-1406.

Wooldridge, J.M. (2001), "Asymptotic properties of weighted m-estimators for standard stratified samples", *Econometric Theory*, 17, 451-470.

Table 1: Experimental Designs: Missing Data Patterns

Experiment	P_1	P_0	P^*	H_S	H_1	H_0	n_1	n_0	n	$N-n$
a	1.000	1.000	1.000	0	.750	.250	225	75	300	0
b	.760	.920	.826	.167	.475	.192	171	69	240	60
c	.227	.920	.247	.375	.106	.144	51	69	120	180
d	.373	.480	.778	.375	.175	.075	84	36	120	180
e	.920	.920	1.000	.074	.639	.213	207	69	276	24

Table 2: Binary Models: Individual Moment Indicators

	Estimators	
	INRSE	UNRSE
H_1	$(1-s)\,ry - H_1$	$(1-s)\,ry - H_1$
H_0	$(1-s)\,r\,(1-y) - H_0$	$(1-s)\,r\,(1-y) - H_0$
H_S	$s - H_S$	$s - H_S$
θ	$(1-s)\,xp\left[\frac{r(y-P)}{P(1-P)} - \frac{(1-r)[(H_1/Q_1)-(H_0/Q_0)]}{1-H_S-(H_0/Q_0)(1-P)-(H_1/Q_1)P}\right]$	$xp\left[\frac{(1-s)r(y-P)}{P(1-P)} - \frac{[(1-s)r+s][(H_1/Q_1)-(H_0/Q_0)]}{H_S+(H_0/Q_0)(1-P)+(H_1/Q_1)P}\right]$
Q_1	$Q_1 - P$	$Q_1 - \frac{[(1-s)r+s]}{H_S+(H_0/Q_0)(1-P)+(H_1/Q_1)P}P$

Note: $P = P\{1|x,\theta\}$, $p = \partial P\{1|x,\theta\}/\partial(x\theta)$.

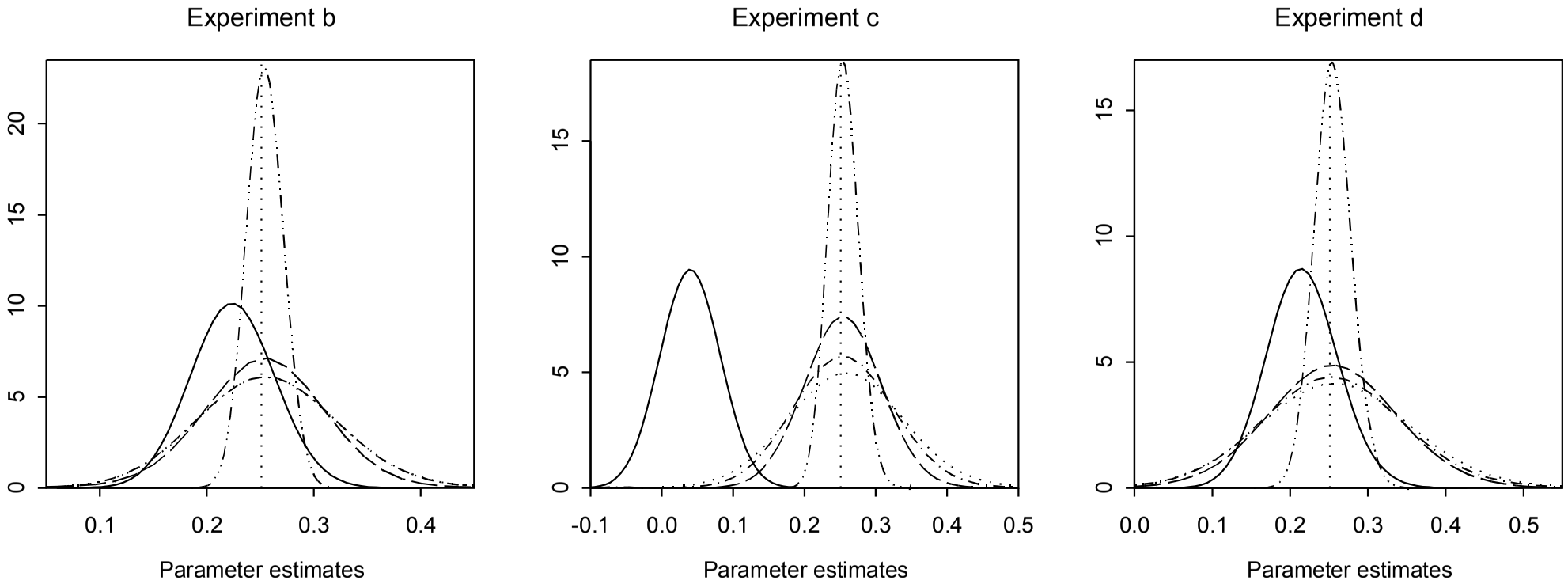
Table 3: Probit Model: Summary Statistics for GMM Estimators

θ = 0.251, Q_1 = 0.75				
Experiment	Estimator	Bias		St. Dev.
		Mean	Median	
a	RSMLE	.008	.005	.028
	QRSMLE	.015	.012	.011
b	RSMLE	$-$.105	$-$.110	.028
	UNRE	.024	.013	.048
	UNRSE	.022	.017	.046
	INRE	.010	.011	.047
	QRSMLE	.015	.012	.012
	QUNRE	.013	.011	.012
	QUNRSE	.013	.011	.011
	QINRE	.015	.013	.012
c	RSMLE	$-$.841	$-$.842	.029
	UNRE	.032	.036	.059
	UNRSE	.016	.008	.052
	INRE	.014	.012	.039
	QRSMLE	$-$.130	$-$.127	.023
	QUNRE	.014	.011	.016
	QUNRSE	.016	.015	.011
	QINRE	.016	.014	.012
d	RSMLE	$-$.138	$-$.145	.033
	UNRE	.034	.027	.068
	UNRSE	.010	.011	.060
	INRE	.015	.021	.061
	QRSMLE	.018	.014	.017
	QUNRE	.016	.013	.017
	QUNRSE	.015	.013	.012
	QINRE	.015	.014	.012
e	RSMLE	.009	.009	.028
	UNRE	.021	.015	.046
	QRSMLE	.016	.014	.012
	QUNRE	.016	.013	.012

Table 4: Probit Model: Summary statistics for P^* Estimates

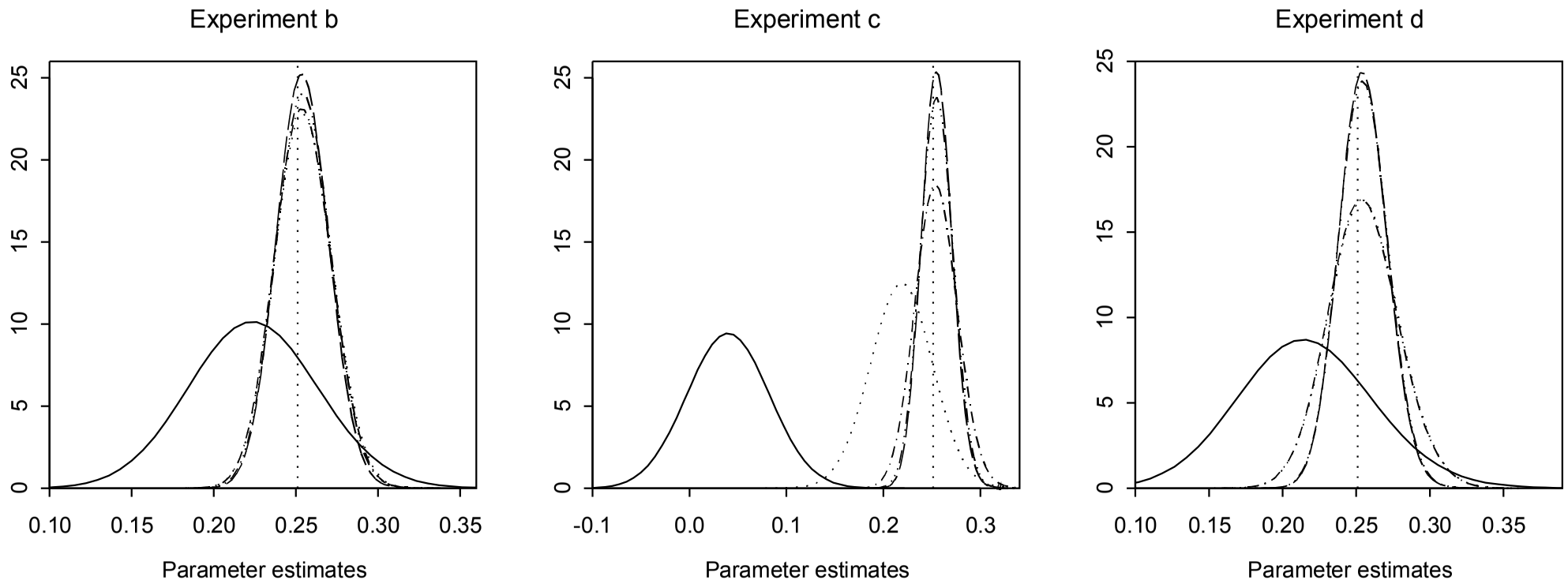
θ = .251, Q_1 = 0.75				
Experiment	Estimator	Bias		St. Dev.
		Mean	Median	
b	UNRE	.006	$-$.020	.204
	UNRSE	$-$.001	$-$.030	.192
	INRE	.020	$-$.021	.203
c	UNRE	.006	$-$.047	.078
	UNRSE	.036	$-$.019	.107
	INRE	.005	$-$.023	.044
d	UNRE	.017	$-$.034	.253
	UNRSE	.060	$-$.021	.284
	INRE	.031	$-$.018	.216
e	UNRE	.007	$-$.023	.239

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.