UNIVERSIDADE DE ÉVORA

DEPARTAMENTO DE ECONOMIA

DOCUMENTO DE TRABALHO Nº **2004/03**

March

Binary models with misclassification in the variable of interest and nonignorable nonresponse

Esmeralda A. Ramalho
Universidade de Évora, Departamento de Economia
e CEMAPRE

UNIVERSIDADE DE ÉVORA
DEPARTAMENTO DE ECONOMIA
Largo dos Colegiais, 2 – 7000-803 Évora – Portugal
Tel.: +351 266 740 894 Fax: +351 266 742 494
www.decon.uevora.pt wp.economia@uevora.pt

Resumo/ Abstract:

In this paper we propose a general framework to deal with datasets where a binary outcome is subject to misclassification and, for some sampling units, neither the error-prone variable of interest nor the covariates are recorded. A model to describe the observed data is for-malized and eficient likelihood-based generalized method of moments (GMM) estimators are suggested. These estimators merely require the formulation of the conditional distribution of the latent outcome given the covariates. The conditional probabilities which describe the error and the nonresponse mechanisms are estimated simultaneously with the parameters of inter-est. In a small Monte Carlo simulation study our GMM estimators revealed a very promising performance.

Palavras-chave/Keyword: nonignorable nonresponse, misclassification, generalized method of moments estimation.

Classificação JEL/JEL Classification: C25, C51

1 Introduction

In this paper we propose a general framework to deal with datasets where a binary outcome is subject to misclassification and, for some sampling units, neither the error-prone variable of interest nor the covariates are recorded. Specifically, we address a situation where misclassification is due to the nature of the variable of interest and, thus, may be described by the conditional probability of the observable outcome given its true value. On the other hand, we consider that nonresponse depends on the error-prone alternative revealed and define a missing data mechanism in terms of the conditional probability of a response indicator given the error-prone outcome. One variable which is usually affected by measurement error and nonresponse is income [see, for example, Peracchi (2002)]. Assume that the aim is modelling the probability of being poor conditional on a set of covariates. Due to the problems in the measurement of income, the poverty status will suffer from misclassification and nonresponse. It is reasonable to assume that the sampling units first decide whether to report or not the correct income and, then, taking into account the response they provide, decide if they give back the questionnaire or not.

We consider a framework where, besides the contaminated incomplete sample, also an independent supplementary random sample (SRS) from the same population, consisting of individual observations of all covariates, is available. The motivation for addressing this setting is twofold. First, all the procedures may be straightforwardly simplified for cases where the latter dataset is not available. Second, the presence of the SRS allow us to cope with the case where a given error-prone outcome is never observed. In absence of misclassification, this problem was addressed by analogy with choice-based (CB) sampling by Lancaster and Imbens (1996), who consider a situation where in the main sample all units choose alternative 1. In this paper we extend their analysis for the case where some of the observed subjects have actually chosen alternative 0 instead of the reported outcome 1.

To the best of our knowledge, the problem of misclassification has not been analysed in presence of nonignorable nonresponse yet, since all papers dealing with misclassification issues assume that there are no missing data; see, for example, Hausman, Abrevaya and Scott-Morton (1998) for random samples (RS) and Ramalho (2003) for CB samples. On the other hand, the literature on nonresponse assumes that, in the available dataset, all the variables are correctly measured; see, inter alia, the textbooks by Little and Rubin (1987) and Schafer (1997) and the recent proposal by Ramalho and Smith (2003) who deal with random samples subject to several patterns of nonignorable nonresponse when the variable of interest is discrete, which include as particular case the situation considered by Lancaster and Imbens (1996).

The main contribution of this paper is the extension of Ramalho and Smith's (2003) methodology for nonignorable nonresponse to handle also misclassification, which is done in a similar way to that employed by Ramalho (2002) to adapt Imbens' (1992) estimators for CB samples to deal with this class of measurement error. We start by formalizing a model adequate to describe the observed data in terms of the latent structural model and the conditional probabilities which define the mechanisms of misclassification and nonresponse. This setting is then utilized to analyze the distortions imposed by both sampling issues in the structural model and in other probabilities of interest, and to develop efficient likelihood-based generalized method of moments (GMM) estimators for both the parameters of interest and the conditional probabilities which govern the willingness to misreport and to participate in the survey.

The layout of this paper is as follows. Section 2 formalizes a regression model appropriate to deal with misclassification and nonresponse. Efficient GMM estimators for this model are developed in section 3. Section 4 presents some particular cases of interest. Section 5 reports some Monte Carlo evidence on the performance in practice of some of the proposed estimators. Finally, section 6 concludes. Some technical proofs are relegated to the appendix.

2 A regression model accounting for misclassification and nonresponse

Let Y^* be a binary response variable and X a vector of k exogenous variables defined on $\mathcal{Y}^* \times \mathcal{X}$, $\mathcal{Y}^*= \{0,1\}$. Employing also the supercript "*" to denote the latent version of all probabilities and densities, the population joint density function of Y^* and X may be written as

$$f^*\left(y^*,x\right) = \Pr^*\left(y^*|x,\theta\right) f\left(x\right), \tag{1}$$

where the marginal density function $f\left(x\right)$ for X is unknown and $\Pr^*\left(y^*|x,\theta\right)$ is known up to the parameter vector θ. For example, $\Pr^*\left(y^*=1|x,\theta\right) = \Phi\left(x\theta\right)$ in probit models and $\Pr^*\left(y^*=1|x,\theta\right) = \left(1+e^{-x'\theta}\right)^{-1}$ in logit models. Our interest is consistent estimation of and inference on the parameter vector θ. The marginal probability of observing an individual for which $Y^*=y^*$ in the population is $Q^*_{y^*} = \int_{\mathcal{X}} \Pr^*\left(y^*|x,\theta\right) f\left(x\right) dx$, with $\sum_{y^*=0}^{1} Q^*_{y^*} = 1$. For simplicity, we denote $Q^*_1 = Q^*$ and $Q^*_0 = 1-Q^*$.

2.1 Incorporating misclassification

In presence of misclassification, let Y represent the binary observable outcome, $\mathcal{Y} = \{0,1\}$. We assume that, conditional on the latent response, the reported outcome is independent of the individual characteristics X, that is after controlling for Y^*, X does not affect Y. Hence, the error model is described by the conditional probability

$$\Pr\left(Y = y | Y^* = y^*, x\right) = \Pr\left(Y = y | Y^* = y^*\right) = \alpha_{yy^*}, \tag{2}$$

$0 \leq \alpha_{yy^*} \leq 1$ and $\sum_{y=0}^{1} \alpha_{yy^*} = 1$. For each outcome $Y = y$ there is one conditional misclassification probability and one conditional probability of correct classification, obtained when $Y \neq Y^*$ and $Y = Y^*$, respectively. Given that $\sum_{y=0}^{1} \alpha_{yy^*} = 1$, only the two misclassification probabilities, contained in the vector $\alpha = (\alpha_{10}, \alpha_{01})$, are required to define the error mechanism. Similarly to Hausman, Abrevaya and Scott-Morton (1998), we adopt the identification condition $\alpha_{10} + \alpha_{01} < 1$, which implies that the probability that a given value $Y = y$ is misreported is smaller than that of being correctly classified, that is, $\alpha_{yy^*} < \alpha_{yy}$ for $Y \neq Y^*$.

In this setup, the conditional probability of the observable variable Y given X and the marginal probability of Y may be written as, respectively,

$$\Pr\left(y | x, \theta, \alpha\right) = \sum_{y^*=0}^{1} \alpha_{yy^*} \Pr{}^*\left(y^* | x, \theta\right), \tag{3}$$

and

$$Q_y = \sum_{y^*=0}^{1} \alpha_{yy^*} Q^*_{y^*}. \tag{4}$$

Unless misclassification is absent, such that $\alpha_{10} = \alpha_{01} = 0$ and $Y = Y^*$, (3) and (4) differ from their error-free counterparts $\Pr^*\left(y^* | x, \theta\right)$ and $Q^*_{y^*}$. Thus, even under the assumption of RS, this sort of measurement error is nonignorable for likelihood-based inference, in the sense that the use of the likelihood function $\Pr^*\left(y^* | x, \theta\right)$ yields inconsistent estimators for θ. As the error mechanism is defined by the two misclassification probabilities contained in α, the traditional approach is to estimate α jointly with θ by maximum likelihood based on the contaminated likelihood function in (3); see Hausman, Abrevaya and Scott-Morton (1998).[1] Similarly, in this paper, we also propose the estimation of α together with the parameters estimated in the nonresponse problem.

2.2 Incorporating nonignorable nonresponse

Assume that a random sample of size N on Y and X is to be collected, but only n individuals accept to participate in the survey. The n sampling units for which (Y, X) is recorded form the so-called complete sample, in which $n = \sum_{y=0}^{1} n_y = \sum_{y^*=0}^{1} n^*_{y^*}$, where n_y is the number of fully observed subjects reporting $Y = y$ and $n^*_{y^*}$ is the (unknown) number of individuals in the sample for which $Y^* = y^*$. Assume also that an independent SRS of all covariates of size m is drawn from

[1] These authors also suggest alternative estimation methods where neither the error model nor the conditional distribution Y^* given X need to be specified; see also Abrevaya and Hausman (1999).

the population of interest and define $N_m = N + m$ and $n_m = n + m$. The SRS is not affected by misclassification, as only the covariates, which are assumed to be error-free, are measured.

While n_y, n and m are observable in all cases, $n^*_{y^*}$ is always unknown, and the total number of individuals involved in the main survey, N, may or may not be known. Throughout this paper we formalize all the models assuming knowledge on N because, when this information is available, its inclusion is the estimation procedures improves inference on the parameters of interest; see Li and Qin (1998) for a discussion of several examples of biased data where the incorporation of this information improves likelihood-based inference. Moreover, all results are straightforwardly adapted for the case where that information is not available (see section 4).

Define the binary indicators R, which takes the value 1 if (Y, X) is observed or 0 otherwise, and S, which takes the value 1 or 0 when the sampling unit belongs to, respectively, the main or the supplementary dataset. We assume that the mechanism which describes the missingness pattern is defined by

$$\Pr(R = 1|Y = y, Y^* = y^*, x) = \Pr(R = 1|Y = y) = \delta_y, \tag{5}$$

where $0 \leq \delta_y \leq 1$. This formulation implies that, conditional on Y, the willingness to respond is independent of both the individual characteristics contained in X and the true outcome Y^*. Thus, (5) describes cases in which Y^* and X have a similar influence over the reported outcome Y and the willingness to participate in the survey. Note also that due to the independence of the main and the supplementary samples, $\Pr(R = 1|Y = y, Y^* = y^*, x, S = 1) = \delta_y$.

In this framework, the contaminated data are said to be missing completely at random (MCAR), according to the definition of Little and Rubin (1987), when $\delta_1 = \delta_0 = \Pr(R = 1)$. This yields a complete error-prone sample in which nonresponse is ignorable for likelihood-based inference, since, as shown in the next subsection, the formulation of the likelihood function merely has to take misclassification into account.

An interesting situation where our mechanisms of misclassification and nonresponse are likely to hold, at least approximately, is that considered by Nicoletti (2003). This author analyses some methods to deal with the presence of several patterns of missing data in income in a wave of the European Comunity Household Panel (ECHP), when modelling the poverty status. In a situation where each houlsehold is taken as a sampling unit, he distinguishes three types of nonresponse. Household unit nonresponse, which arises when the questionnaire of the houlsehold is not given back, personal unit nonresponse, when some of the members of the household do not give back their questionnaire, and personal iten nonresponse, when some members of the houlsehold do not reveal their income. Our framework is appropriate to deal with this case. In fact, the first individuals obviously generate nonresponse of the kind we consider in this paper. The problem of the two last groups of sampling units generates misclassification in the poverty status of the household. Concerning the mechanism of misclassification, it is reasonable to assume that, conditional on the true poverty status, the observed outcome is independent of other charactristics which may be used as regressors, as household complosition, for example. Moreover, it is likely that the willingness to give back the questionnaire is explained by the observed income, since the reported value for income may lead the household head not to return the questionnaire, namely if there is some personal nonresponse.

2.3 Formulating the model by analogy with CB sampling

Similarly to Ramalho and Smith (2003) in absence of misclassification, we handle the missing data problems of interest by analogy with the CB sampling framework. For each of the two observable outcomes Y, we reinterpret as strata the set of respondents and the set of nonrespondents. The proportion of each stratum of respondents and nonrespondents in the population is the same, Q_y, and in the sample is, respectively, $H_y = \Pr(Y = y, R = 1, S = 1)$ and $H_y^{nr} = \Pr(Y = y, R = 0, S = 1)$. Additionally, the SRS form another stratum with proportion 1 in the population and $H_S = \Pr(S = 0)$ in the sample. Thus, the conditional probability in (5) may be expressed as

$$\delta_y = \frac{H_y}{Q_y(1 - H_S)}, \tag{6}$$

from which it is clear that the data are MCAR only if the ratio $\frac{H_y}{Q_y}$ is constant across all Y; see Ramalho and Smith (2003) for details. The proportions H_y and H_S may be easily estimated as $\hat{H}_y = \frac{n_y}{N_m}$ and $\hat{H}_S = \frac{m}{N_m}$.[2] Q_y is, in principle, unknown, although in some cases there may exist an available estimate for this probability. When this aggregate information is available, similarly to what is usually done in the literature for CB sampling, we assume that is was obtained from a large random sample, for example a census, and deal with it as if it was exact. Alternatively we may have an estimate for the error-free version of this probability, Q^*, and this information is also assumed to be exact. The estimators suggested in the next section do not require prior information on Q or Q^*. However, when one of these probabilities is known, this information may be incorporated in the estimation procedure to improve inference.

To conclude this section, it is interesting to note that, due to misclassification, the conditional

[2] Note that, although H_S remains unaffected by the mismeasurement, the proportion of a given latent outcome in the sample, $H^*_{y^*} = \Pr(Y^* = y^*, R = 1, S = 1)$, differs from H_y. In fact, $H^*_{y^*}$ is written as $H^*_{y^*} = Q^*_{y^*} \sum_{y=0}^{1} \frac{H_y}{Q_y} \alpha_{yy^*}$ and could be estimated as $\hat{H}^*_{y^*} = \frac{n^*_{y^*}}{N_m}$ if $n^*_{y^*}$, the number of individuals in the sample for which $Y^* = y^*$, was known.

probability of response given the true outcome Y^*,

$$\Pr(R = 1|Y^* = y^*) = \sum_{y=0}^{1} \delta_y \alpha_{yy^*} = \eta_{y^*}, \tag{7}$$

differs from δ_y in (5). The two probabilities only become identical when either misclassification is absent, such that $\alpha_{10} = \alpha_{01} = 0$, or the data are MCAR, in which case $\delta_y = \eta_{y^*} = \Pr(R = 1)$. On the other hand, due to nonignorable nonresponse, the conditional probability of observing Y given Y^* among the respondents, $\Pr_S(Y = y|Y^* = y^*, R = 1)$, is not simply α_{yy^*}. In fact, from the sampling joint probability of observing $(Y, Y^*, R = 1, S = 1)$,

$$\Pr_S(Y = y, Y^* = y, R = 1, S = 1) = Q^*_{y^*} \alpha_{yy^*} \delta_y (1 - H_S), \tag{8}$$

we may straightforwardly obtain

$$\begin{aligned} \Pr_S(Y = y|Y^* = y, R = 1, S = 1) &= \frac{\Pr_S(Y = y, Y^* = y, R = 1, S = 1)}{\Pr_S(Y^* = y, R = 1, S = 1)} \\ &= \frac{Q^*_{y^*} \alpha_{yy^*} \delta_y (1 - H_S)}{Q^*_{y^*} (1 - H_S) \sum_{y=0}^{1} \alpha_{yy^*} \delta_y} \\ &= \frac{\alpha_{yy^*} \delta_y}{\sum_{y=0}^{1} \alpha_{yy^*} \delta_y} \\ &= \Pr_S(Y = y|Y^* = y, R = 1) = \varpi_{yy^*}, \end{aligned} \tag{9}$$

which only when the data are MCAR, such that δ_y is constant for all Y, is reduced to α_{yy^*}.

2.4 Observed data likelihood functions

The observed data likelihood functions of interest to analyze the problem of nonignorable nonresponse with misclassification are derived by analogy with Ramalho and Smith (2003), with the difference that here we deal with error-prone data. Consider the observable $V = (Y, X, R, S)$. Taking into account that one observes $(Y, X, R = 1, S = 1)$, $(R = 0, S = 1)$, and $(X, S = 0)$ for, respectively, respondents, nonrespondents and units of the SRS, the likelihood function for an individual in the available dataset is

$$\begin{aligned} l(v) &= \left[h(y, x, r = 1, s = 1)^r \Pr(r = 0, s = 1)^{1-r}\right]^s h(x, s = 0)^{1-s} \\ &= \left\{\left[\frac{H_y}{Q_y} \Pr(y|x, \theta, \alpha) f(x)\right]^r (1 - H_S - H_1 - H_0)^{1-r}\right\}^s [H_S f(x)]^{1-s}. \end{aligned} \tag{10}$$

The information concerning the main sample is associated to indicator S. The contribution of respondents, indexed by R, is the likelihood function of the error-prone complete data. This is the only component of (10) affected by misclassification, since the error-prone variable of interest is only measured for respondent individuals. Note that when data are MCAR, such that $\frac{H_y}{Q_y}$

is constant for all Y, this likelihood function reduces to the joint density function of (Y, X), $f(y, x) = \Pr(y|x, \theta, \alpha) f(x)$. In this case, the failure of participation of some sampling units may be ignored and inference may be based on $\Pr(y|x, \theta, \alpha)$ of (3), since $f(x)$ factors out from the resulting log-likelihood function. From (10), it is also clear that the component associated with $(1 - R)$ merely includes the information on the size of the initial main sample N, since nothing else is known about the nonparticipant units. Finally, the term indexed by $(1 - S)$ is the contribution of the individuals of the SRS for which only X is measured.

The GMM estimators we suggest in the next section will be based on likelihood (10). However, two other likelihood functions are relevant to characterize the problem in analysis: the joint probability of (R, S) and the marginal density of the covariates in the sample, given by, respectively,

$$\begin{aligned}\Pr(R = r, S = s) &= \left\langle \left[\int_{\mathcal{X}} \sum_{y=0}^{1} \frac{H_y}{Q_y} \Pr(y|x, \theta, \alpha) f(x)\, dx \right]^r \right. \\ &\quad \left. (1 - H_S - H_1 - H_0)^{1-r} \right\rangle^s \left[H_S \int_{\mathcal{X}} f(x)\, dx \right]^{1-s} \\ &= \left[(H_1 + H_0)^r (1 - H_S - H_1 - H_0)^{1-r} \right]^s H_S^{1-s} \qquad (11)\end{aligned}$$

and

$$\begin{aligned}h(x) &= \sum_{s=0}^{1} \sum_{r=0}^{1} \left\{ \left[\sum_{y=0}^{1} \frac{H_y}{Q_y} \Pr(y|x, \theta, \alpha) f(x) \right]^r (1 - H_S - H_1 - H_0)^{1-r} \right\}^s \\ &\quad [H_S f(x)]^{1-s} \\ &= f(x) \left[H_S + \sum_{y=0}^{1} \frac{H_y}{Q_y} \Pr(y|x, \theta, \alpha) \right] + 1 - H_S - H_1 - H_0. \qquad (12)\end{aligned}$$

Because (11) [(12)] does not depend (depends) on θ and α, the indicators R and S (the covariates) are (not) ancillary for these parameters vectors. Thus, the analysis must be conditional on R and S and not on X. The next section describes an estimation procedure which yields estimators conditional on R and S, even though the likelihood (10) is not conditional on these indicators. Moreover, although the analysis is not conditional on X, the specification of $f(x)$ is circumvented.

3 Generalized method of moments estimation

In this section we derive likelihood-based GMM estimators which merely require the specification of the structural model $\Pr^*(y^*|x, \theta)$. All the procedures are similar to those suggested by Ramalho and Smith (2003) for nonresponse, but a further vector, α, is incorporated in the parameters of interest. The main consequence is that, analogously to what happens when no missing data is present, large samples sizes are required to estimate properly the misclassification probabilities; see, for example, the discussion in Copas (1988).

The vector of parameters of interest is, thus, $\varphi = (H, \theta, \alpha, Q^*)$, with $H = (H_0,\ H_1, H_S)$, for cases where the marginal choice probabilities Q and Q^* are unknown, or simply $\varphi = (H, \theta, \alpha)$, when one of those probabilities is known. Note that the parameter vector H, which could be estimated separately from $\hat{H}_y = \frac{n_y}{N_m}$ and $\hat{H}_S = \frac{m}{N_m}$, is estimated together with the remaining parameters of interest in order to condition the analysis on the ancillary statistics $\hat{H}_y$ and $\hat{H}_S$; for a discussion on this procedure of conditioning the analysis on ancillary statistics, see Imbens and Lancaster (1996).

In order to avoid the specification of $f(x)$, assume that the covariates follow a discrete distribution with L points of support x^l, $l = 1, 2..., L$, and associated probability mass parameters $\Pr\left(X = x^l\right) = \pi_l$, $\pi_l > 0$, $l = 1, 2..., L$. The resultant log-likelihood function based on (10),

$$\begin{aligned} L(H, \theta, \pi) &= \sum_{i=1}^{N_m} s_i r_i \left[\ln H_{y_i} + \ln \Pr\left(y_i | x^{l_i}, \theta, \alpha\right) - \ln \sum_{l=1}^{L} \pi_l \Pr\left(y_i | x^l, \theta, \alpha\right) + \ln \pi_{l_i} \right] + \\ & \sum_{i=1}^{N_m} s_i (1 - r_i) \ln (1 - H_{S_i} - H_{1_i} - H_{0_i}) + \sum_{i=1}^{N_m} (1 - s_i) (\ln H_{S_i} + \ln \pi_{l_i}), \end{aligned} \tag{13}$$

is maximized with respect to the vector of parameters (H, θ, α, π) subject to the restriction $\sum_{l=1}^{L} \pi_l = 1$. From the first order conditions of (13), we may define the following estimating functions,

$$g(v, \varphi)_{H_t} = srI_{(y=t)} - H_t \tag{14}$$

$$g(v, \varphi)_{H_S} = 1 - s - H_S \tag{15}$$

$$g(v, \varphi)_{\theta} = p \left\{ sr \frac{y - P}{P(1 - P)} - [1 - s(1 - r)] \frac{R}{B} \right\} \tag{16}$$

$$g(v, \varphi)_{\alpha_{yy^*}} = [y - \Pr^*(y^* | x, \theta)] \left\{ sr \frac{y - P}{P(1 - P)} - [1 - s(1 - r)] \frac{R}{B} \right\} \tag{17}$$

$$g(v, \varphi)_{Q^*} = \alpha_{10} + (1 - \alpha_{10} - \alpha_{01}) Q^* - [1 - s(1 - r)] \frac{P}{B}, \tag{18}$$

where $t = \{0, 1\}$, $P = \Pr(Y = 1 | x, \theta, \alpha)$, $p = \nabla_\theta P$, $R = \frac{H_1}{Q} - \frac{H_0}{1-Q}$, $B = H_S + \frac{H_0}{1-Q} + RP$. Note that (14)-(18) do not depend on π_l, which no longer needs to be estimated, and Q^* was introduced in the vector of parameters of interest φ; for details on the derivation of (14)-(18), see appendix A.

The unknown parameters present in the set of estimating equations resulting from (14)-(18) may be estimated by GMM. The objective function to be minimized is $\Upsilon_N(\varphi) = g_N(v, \varphi)' W_N g_N(v, \varphi)'$, where $g_{N_m}(v, \varphi) = \frac{1}{N_m} \sum_{i=1}^{N_m} g(v_i, \varphi)$ is the sample counterpart of the moment conditions $E[g(v, \varphi)] = 0$, with $E[.]$ denoting expectation taken over $l(v)$ of (10), and W_N is a positive semi-definite weighting matrix. Assume that the usual regularity conditions required for GMM estimation are meet; see Newey and McFadden (1994, Theorems 2.6, 3.4). The resulting optimal estimator, $\hat{\varphi}$, obtained from choosing $W_N = \Psi_N^{-1}$, where Ψ_N is a consistent estimator of $\Psi = E\left[g(v,\varphi)\, g(v,\varphi)'\right]$, is consistent for the true value φ^0 and satisfies

$$\sqrt{N_m}\left(\hat{\varphi} - \varphi^0\right) \xrightarrow{d} N\left[0, \left(G'\Psi^{-1}G\right)^{-1}\right], \tag{19}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $G = E\left[\nabla_{\varphi} g(v,\varphi)'\right]$. Asymptotic efficiency, in the semiparametric sense, can also be proved by an analogous demonstration to that of Imbens (1992, Theorem 3.3); see appendix B. When both Q^* and Q are unknown, $\hat{\varphi} = \left(\hat{H}, \hat{\theta}, \hat{\alpha}, \hat{Q}^*\right)$ is a just-identified estimator. Otherwise, when there is prior knowledge on either Q^* or Q, that information is replaced in (14)-(18) [note that $Q = \alpha_{10} + (1 - \alpha_{10} - \alpha_{01})\, Q^*$] and merely $\hat{\varphi} = \left(\hat{H}, \hat{\theta}, \hat{\alpha}\right)$ needs to be estimated, which yields an overidentified GMM estimator.

4 Particular cases

The framework developed previously may be simplified in a number of ways. Furthermore, it nests many of the estimators previously proposed in the literature on misclassification, nonresponse, and CB sampling.

4.1 Unknown N

Consider that the initial sample size N is unknown or this information is not used in the estimation. As now only the respondents and units of the SRS are accounted for, we need to set $R = 1$, replace N_m by n_m in all calculations and results, and, since $H_S + H_0 + H_1 = 1$, suppress one of the moment indicators for H_0, H_1 or H_S.[3] If we consider that none of the sampling units for which $Y = 0$ respond and all subjects for which $Y = 1$ reveal (Y, X), we obtain a generalization to handle misclassification of Lancaster and Imbens' (1996) estimators for case-control studies with missing controls. Indeed, now Y is set to 1, $n_m = n_1 + m$, and, as $H_0 = 0$ and $H_S + H_1 = 1$, both $g(v,\varphi)_{H_0}$ and either $g(v,\varphi)_{H_1}$ or $g(v,\varphi)_{H_S}$ are suppressed. Interestingly, in this case $\delta_1 = 1$ and $\delta_0 = 0$ but, due to misclassification, the probabilities of observing an individual for which the true outcome is $Y^* = 1$ and $Y^* = 0$ is, respectively, $\eta_1 = \alpha_{11}$ and $\eta_0 = \alpha_{10}$ [see equation (7)]. Moreover, the probability of misclassification among the respondents is $\varpi_{10} = 1$ [$\varpi_{01} = 0$], reflecting the fact that all the individuals for which $Y^* = 0$ and $Y = 1$ are included in the sample [none of the individuals for which $Y^* = 1$ and $Y = 0$ is observed].

[3] In terms of model specification, besides these simplifications, one has to take into account that now, as $H_y = \Pr(Y = y, S = 1 | R = 1)$ and $H_y^{nr} = 0$, the relation (6) is no longer valid.

4.2 Absence of the supplementary random sample

When a SRS is not available, we set $S = 1$, $H_S = 0$, replace N_m by N, and eliminate $g(v, \varphi)_{H_S}$. In this framework, there are two particular cases of interest. First, Ramalho's (2002) estimators for CB samples subject to misclassification are obtained from this setting by considering N unknown and implementing similar modifications to those suggested in the previous paragraph: set $R = 1$, replace N by n, and eliminate either $g(v, \varphi)_{H_0}$ or $g(v, \varphi)_{H_1}$. Second, consider the case where the conditional probability of the latent variable of interest given the covariates is a logit model including an intercept term, such that $\Pr^*(y^* = 1|x, \theta) = \left(1 + e^{-x'\theta}\right)^{-1}$, where $\theta = (\theta_0, \theta_1)$, with θ_0 defined as an intercept term. As the nonresponse mechanism depends only on Y, in absence of misclassification the ML RS estimator could be used with the complete dataset for consistent estimation of the slope parameter vector θ_1; see, for example, Ramalho and Smith (2003). Similarly, by an analogous demonstration to that of Caudill and Cosslett (2004) for CBS, it can be shown that nonresponse is also ignorable when the variable of interest is error-prone. In simple terms, this is due to the fact that, in the complete error-prone data, the probability of Y given X,[4]

$$\begin{aligned}
\Pr_S(y|x, R = 1, \theta, \alpha, \delta) &= \frac{\Pr_S(y, x, R = 1, \theta, \alpha, \delta)}{\Pr_S(x, R = 1, \theta, \alpha, \delta)} \\
&= \frac{\sum_{y^*=0}^{1} \delta_j \alpha_{jy^*} \Pr^*(y^*|x, \theta) f(x)}{\sum_{y=0}^{1} \sum_{y^*=0}^{1} \delta_y \alpha_{yy^*} \Pr^*(y^*|x, \theta) f(x)} \\
&= \frac{\delta_j \left(\alpha_{j0} e^{-x'\theta} + \alpha_{j1}\right)}{\delta_0 (\alpha_{00} e^{-x'\theta} + \alpha_{01}) + \delta_1 (\alpha_{10} e^{-x'\theta} + \alpha_{11})} \\
&= \frac{\delta_j \left(\alpha_{j0} e^{-x'\theta} + \alpha_{j1}\right)}{\sum_{y=0}^{1} \delta_y \alpha_{y0} e^{-x'\theta} + \sum_{y=0}^{1} \delta_y \alpha_{y1}} \\
&= \frac{\frac{\delta_j \alpha_{j0}}{\varpi_0} \frac{\varpi_0}{\varpi_1} e^{-x'\theta} + \frac{\delta_j \alpha_{j1}}{\varpi_1}}{1 + \frac{\varpi_0}{\varpi_1} e^{-x'\theta}} \qquad (20) \\
&= \frac{\omega_{j0} \frac{\varpi_0}{\varpi_1} e^{-x'\theta} + \omega_{j1}}{1 + \frac{\varpi_0}{\varpi_1} e^{-x'\theta}} \qquad (21)
\end{aligned}$$

is identical to the error-prone conditional probability of observing the outcome Y given X

$$\Pr(y = j|x, \theta, \alpha) = \frac{\alpha_{j0} e^{-x'\theta} + \alpha_{j1}}{1 + e^{-x'\theta}} \qquad (22)$$

with θ_0 replaced by $\gamma = \theta_0 - \ln \frac{\varpi_0}{\varpi_1}$ and α_{yy^*} replaced by $\varpi_{yy^*} \frac{\varpi_0}{\varpi_1}$. Thus, for consistent estimation of θ_1, one may utilize the simple likelihood (22), where only the problem of misclassification is accounted for, with the complete dataset.

[4] Note that, to obtain (20), we make $\omega_0 = \sum_{y=0}^{1} \delta_y \alpha_{y0}$, $\omega_1 = \sum_{y=0}^{1} \delta_y \alpha_{y1}$, divide and multiply the first and the second term of the numerator by, respectively, ω_0 and ω_1, and, then, divide all the resulting terms by ω_1.

4.3 Absence of misclassification

Ramalho and Smith's (2003) estimators for total nonresponse when a SRS of covariates is available, the ones we extended in this paper to handle misclassification, arise when $\alpha_{10} = \alpha_{01} = 0$ and $g\left(v, \varphi\right)_{\alpha_{yy^*}}$ are eliminated. The same simplification applied to our generalization of Lancaster and Imbens' (1996) estimators yield their original proposal.[5]

5 A Monte Carlo simulation study

This section analyzes the performance of the estimation method proposed in this paper in cases where Y^* given X is described by a logit model, the main sample only contains individuals who reported 1, and a SRS is available. The number of individuals choosing 0, n_0, and, consequently, the size of the main sample, N, are assumed to be unknown. As in absence of misclassification Lancaster and Imbens' (1996) estimator would be appropriate to deal with this dataset, we replicated two of their Monte Carlo experimental designs with two adaptations. First, we admitted the possibility that some of the observed subjects have chosen alternative 0 instead of the reported outcome 1. Second, in order to be able to handle misclassification, we considered a much larger sample size ($n_m = 5000$).

The covariates X were generated from a bivariate normal distribution with zero means, unit variances and zero correlation. In the two experimental designs, designated as A and B, the vector of parameters of interest θ contained in $\Pr^*\left(y^* = 1|x, \theta\right) = \left(1 + e^{-x'\theta}\right)^{-1}$, where $\theta = (\theta_0, \theta_1, \theta_2)$ with θ_0 defined as an intercept term, was set equal to, respectively, $(0.0, 2.0, 0.5)$ and $(-1.89, 1.0, 1.0)$, producing a proportion of individuals choosing alternative $Y^* = 1$ of $Q^* = 0.50$ and $Q^* = 0.20$. In both designs the weight of the main and the supplementary sample is the same, such that $H_1 = H_S = 0.5$ (and $H_0 = 0$), $n = n_1 = 2500$ and $m = 2500$, and the misclassification probabilities are $\alpha_{10} = \alpha_{01} = \bar{\alpha} = \{0.02, 0.05, 0.20\}$. In all experiments we assumed that the marginal probabilities Q^* and Q are unknown. So, two GMM estimators were compared, namelly Lancaster and Imbens' (1996) estimator and its modified version for misclassification developed in this paper, which are denoted by LIE and MLIE, respectively. The vector of parameters estimated in each case is, respectively, $\varphi = (H_1, \theta, Q^*)$ and $\varphi = (H_1, \theta, \bar{\alpha}, Q^*)$.

Table 1 contains the mean and the median bias in percentage terms and the standard deviation

[5] Note that the simplified version of moment indicators (14) and (18) does not coincide with the moment indicators proposed by Lancaster and Imbens (1996). In effect, their proposals are $g\left(v, \varphi\right)_{H_1}^{LI} = H_1 - \frac{H_1}{QB}P$ and $g\left(v, \varphi\right)_{Q}^{LI} = -\frac{1}{Q}\left(y - \frac{H_1}{QB}P\right)$. However, the former uses an alternative consistent estimator for $\hat{H}_1$, $\frac{1}{n_1+m}\sum_{i=1}^{n_1+m}\frac{H_{1_i}}{Q_i B_i}P_i$, instead of $\frac{n_1}{n_1+m}$. The latter is proportional to $g\left(v, \varphi\right)_Q$, because it may be written as $g\left(v, \varphi\right)_Q^{LI} = -\frac{1}{Q}\left[\frac{H}{Q}g\left(v, \varphi\right)_Q + g\left(v, \varphi\right)_{H_1}\right]$.

across 1000 replications of the slope estimates. The number of replications that failed to converge (F.C.) is also reported, since it was very large for $\bar{\alpha} = 0.2$, mainly in LIE, which ignore the presence of misclassification. This problem is more serious in design B, similarly to the results reported by Lancaster and Imbens (1996). The justification presented by these authors for this fact is also valid in presence of misclassification: as the main sample merely contains units for which $Y = 1$ and the SRS when Q is small contains mainly units that would report $Y = 0$, the level of overlapping of the two samples is very small. The available dataset is, thus, near a CB sampling design in which one stratum includes individuals that reveal $Y = 1$ and other includes individuals reporting $Y = 0$, a situation where the problems of identification of the intercept term and the marginal choice probabilities are well known; see Manski and Lerman (1977), who first discussed this issue.

Table 1 about here

The behaviour of MLIE in terms of mean and median bias is very promising, namely for the two smallest misclassification probabilities, where the worst distortion for the slope parameters is 1.3% (for the mean of θ_2, in design B, when $\bar{\alpha} = 0.02$). Naturally, the performance decays with the highest level of misclassification, but even in these cases the median bias is smaller than 4.1%.[6] On the other hand, as expected, in all cases the biases in LIE are far larger than those presented by our modified estimators, the divergence being increased when the probability of misclassification gets large. With regard to the standard deviation across the replications, it is always substantially larger in MLIE, specially for $\bar{\alpha} = 0.20$, as these estimators capture the additional variability induced by the measurement error.

Thus, the overall performance of MLIE is very satisfactory, namely when the probability of misclassification is moderate. Obviously, when the amount of measurement error grows, the quality of our modified estimators worsens, mainly due to the increment in both the dispersion of the estimates and the failures of convergence. However, even in these cases, the behaviour of MLIE is far better than that of the uncorrected estimators.

6 Conclusion

In this paper we proposed a general framework to deal with the presence of misclassification and missing data in datasets where the variable of interest is binary. By modifying the setup usually employed with CB sampling, we specified the regression model of interest in terms of the struc-

[6] Note that in design A, the results for $\alpha = 0.2$ were negatively affected by the presence of 4 replications where the estimate for θ_1 was larger than 30. Eliminating these replications, the mean bias for θ_1 and θ_2 is reduced to, respectively, 8.0% and 9.3% and their standard deviations across the replications are 0.713 and 0.240.

tural model and the conditional probabilities which define the mechanisms of misclassification and nonresponse. This formulation emphasized the distortions imposed by both sampling issues over the parametric model maintained in the population of interest, which, obviously, cause the inconsistency of all the likelihood-based estimators which ignore both or one of the sampling problems. The model for the observed data was also utilized to derive efficient GMM estimators. The performance in practice of some of these estimators was assessed through a small Monte Carlo simulation study. The results were very promising, particularly when the amount of misclassification was moderate.

Appendix A: derivation of the moment indicators

The maximization of the log-likelihood (13) with respect to (H, θ, α, π) yields the following first order conditions:

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{H_t} = \sum_{i=1}^{N_m} s_i I_{(y_i=t)} \left(\frac{r_i}{\hat{H}_t} - \frac{1-r_i}{1-\hat{H}_S-\hat{H}_0-\hat{H}_1}\right) = 0 \tag{23}$$

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{H_S} = \sum_{i=1}^{N_m} I_{(s_i=0)} \left[\frac{1-s_i}{\hat{H}_S} - \frac{s_i\left(1-r_i\right)}{1-\hat{H}_S-\hat{H}_0-\hat{H}_1}\right] = 0 \tag{24}$$

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{\theta} = \sum_{i=1}^{N_m} s_i r_i \left[\nabla_\theta \ln \Pr\left(y_i|x^{l_i}, \hat{\theta}, \hat{\alpha}\right) - \frac{\sum_{l=1}^{L} \hat{\pi}_l \nabla_\theta \Pr\left(y_i|x^{l}, \hat{\theta}, \hat{\alpha}\right)}{\sum_{l=1}^{L} \hat{\pi}_l \Pr\left(y_i|x^{l}, \hat{\theta}, \hat{\alpha}\right)}\right] = 0 \tag{25}$$

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{\alpha_{yy*}} = \sum_{i=1}^{N_m} s_i r_i \left[\nabla_{\alpha_{yy*}} \ln \Pr\left(y_i|x^{l_i}, \hat{\theta}, \hat{\alpha}\right) - \frac{\sum_{l=1}^{L} \hat{\pi}_l \nabla_{\alpha_{yy*}} \Pr\left(y_i|x^{l}, \hat{\theta}, \hat{\alpha}\right)}{\sum_{l=1}^{L} \hat{\pi}_l \Pr\left(y_i|x^{l}, \hat{\theta}, \hat{\alpha}\right)}\right] = 0 \tag{26}$$

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{\pi_z} = \sum_{i=1}^{N_m} \left\{ s_i r_i \left[\frac{I_{(l_i=z)}}{\hat{\pi}_z} - \frac{\Pr\left(y_i|x^{z}, \hat{\theta}, \hat{\alpha}\right)}{\sum_{l=1}^{L} \hat{\pi}_l \Pr\left(y_i|x^{z}, \hat{\theta}, \hat{\alpha}\right)}\right] + \left(1-s_i\right) \frac{I_{(l_i=z)}}{\hat{\pi}_z} \right\} - \hat{\mu} = 0 \tag{27}$$

$$s\left(\hat{H}, \hat{\theta}, \hat{\pi}\right)_{\mu} = \sum_{l=1}^{L} \hat{\pi}_l - 1 = 0, \tag{28}$$

where $t = \{0, 1\}$, $z = \{1, ..., L\}$, $I_{(s=t)}$ takes the value 1 for $s = t$ and 0 for $s \neq t$, ∇_β denotes derivative with respect to β, and μ is the Lagrange multiplier associated with the restriction $\sum_{l=1}^{L} \pi_l = 1$.

In order to transform the system (23)-(28) into (14)-(18), we first multiply all the terms in

(27) by $\hat{\pi}_z$ and sum over z, to obtain

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

Taking into account that the maximum likelihood estimator for Q_y may be written from (4) as $\hat{Q}_y = \sum_{l=1}^{L}\hat{\pi}_l \Pr\left(y|x^l,\hat{\theta},\hat{\alpha}\right)$ and replacing $\hat{\mu}$ by $\hat{H}_S$ in (27), yields

$$\begin{aligned}
\hat{\pi}_z &= \frac{1}{N_m}\sum_{i=1}^{N_m}\left[1-s_i\left(1-r_i\right)\right] I_{(l_i=z)}\left[\hat{H}_S + \frac{1}{N_m}\sum_{i=1}^{N_m}\frac{s_i r_i}{\hat{Q}_{y_i}}\Pr\left(y_i|x^z,\hat{\theta},\hat{\alpha}\right)\right]^{-1} \\
&= \frac{1}{N_m}\sum_{i=1}^{N_m}\left[1-s_i\left(1-r_i\right)\right] I_{(l_i=z)}\left[\hat{H}_S + \sum_{y=0}^{1}\frac{\hat{H}_y}{\hat{Q}_y}\Pr\left(y|x^z,\hat{\theta},\hat{\alpha}\right)\right]^{-1} \\
&= \frac{1}{N_m}\sum_{i=1}^{N_m}\left[1-s_i\left(1-r_i\right)\right] I_{(l_i=z)}\left[\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)\right]^{-1} \qquad (29)
\end{aligned}$$

Then, $\hat{\pi}_l$ is substituted in both the last terms of (25) and (26). As the calculations are similar, only those for (25) are presented:

$$\begin{aligned}
&\sum_{i=1}^{N_m}\frac{s_i r_i}{\hat{Q}_{y_i}}\sum_{l=1}^{L}\hat{\pi}_l\nabla_\theta \Pr\left(y_i|x^l,\hat{\theta},\hat{\alpha}\right) = \\
&= \sum_{i=1}^{N_m}\frac{s_i r_i}{\hat{Q}_{y_i}}\sum_{l=1}^{L}\frac{1}{N_m}\sum_{j=1}^{N_m}\frac{\left[1-s_j\left(1-r_j\right)\right] I_{(l_j=l)}}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}\nabla_\theta \Pr\left(y_i|x^l,\hat{\theta},\hat{\alpha}\right) \\
&= \sum_{i=1}^{N_m}\frac{s_i r_i}{\hat{Q}_{y_i}}\frac{1}{N_m}\sum_{j=1}^{N_m}\frac{1-s_j\left(1-r_j\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}\nabla_\theta \Pr\left(y_i|x^l,\hat{\theta},\hat{\alpha}\right) \\
&= \sum_{j=1}^{N_m}\frac{1-s_j\left(1-r_j\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}\frac{1}{N_m}\sum_{i=1}^{N_m}\frac{s_i r_i}{\hat{Q}_{y_i}}\nabla_\theta \Pr\left(y_i|x^l,\hat{\theta},\hat{\alpha}\right) \\
&= \sum_{j=1}^{N_m}\frac{1-s_j\left(1-r_j\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}\sum_{y=0}^{1}\frac{\hat{H}_y}{\hat{Q}_y}\nabla_\theta \Pr\left(y|x^l,\hat{\theta},\hat{\alpha}\right) \\
&= \sum_{j=1}^{N_m}\frac{\left[1-s_j\left(1-r_j\right)\right]\left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\nabla_\theta \Pr\left(y=1|x^l,\hat{\theta},\hat{\alpha}\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}.
\end{aligned}$$

As the transformed versions of (25) and (26) depend on Q which may be unknown for the researcher, an estimating function for this probability is required. Similarly to Imbens (1992) procedure for CB sampling, we use the definition of the maximum likelihood estimator for Q with $\hat{\pi}_l$

replaced by the estimator given in (29)

$$\begin{aligned}\hat{Q} &= \sum_{l=1}^{L} \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{\left[1-s_j\left(1-r_j\right)\right] I_{(l_i=l)} \Pr\left(y_i=1|x^l,\hat{\theta},\hat{\alpha}\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)} \\ &= \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{\left[1-s_j\left(1-r_j\right)\right] \Pr\left(y_i=1|x^l,\hat{\theta},\hat{\alpha}\right)}{\hat{H}_S + \frac{\hat{H}_0}{1-\hat{Q}} + \left(\frac{\hat{H}_1}{\hat{Q}} - \frac{\hat{H}_0}{1-\hat{Q}}\right)\Pr\left(y=1|x^z,\hat{\theta},\hat{\alpha}\right)}.\end{aligned}$$

To obtain moment indicator (18) $\hat{Q}$ is written from (4) as $\hat{Q} = \hat{\alpha}_{10} + \left(1-\hat{\alpha}_{10}-\hat{\alpha}_{01}\right)\hat{Q}^*$. This last substitution allow us to deal with both cases where we have information on $\hat{Q}$ and $\hat{Q}^*$.

Appendix B: efficiency of the generalized method of moments estimators

Similarly to Imbens (1992), the efficiency of the GMM estimators proposed previously is proved by showing that the Cramér-Rao lower bounds associated with a sequence of parametric models which satisfy the same regularity conditions as our model, converges to the asymptotic covariance matrix of our semiparametric estimators.

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

The parametric model indexed by ε, which results from substituting $f(x,\varpi)$ in (10), is

$$\begin{aligned}l_\varepsilon(v) &= \left\{\left[H_y \frac{\Pr\left(y|x,\theta,\alpha\right) f_\varepsilon(x) \sum_{l=1}^{L_\varepsilon}\phi_{lx}\varpi_l}{\sum_{l=1}^{L_\varepsilon}\varpi_l\int_{\mathcal{X}_l}\Pr\left(y|x,\theta,\alpha\right) f_\varepsilon(x)\phi_{lx}dx}\right]^r \left(1-H_S-H_1-H_0\right)^{1-r}\right\}^s \\ &\quad \left(H_S f_\varepsilon(x)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\varpi_l\right)^{1-s},\end{aligned}$$

which, as $f_\varepsilon(x)$ is a known function, depend on the unknown vector of parameters $(H,\theta,\alpha,\phi_{lx})$. Constructing the log-likelihood function, taking the first order derivatives and noting that the maximum likelihood estimator for Q is written from (4) as $\hat{Q}_y = \sum_{l=1}^{L_\varepsilon}\varpi_l\int_{\mathcal{X}_l}\Pr\left(y|x,\hat{\theta},\hat{\alpha}\right) f_\varepsilon(x)\phi_{lx}dx$, the dependence on ϖ_l can be removed following the same procedure described in Appendix A to remove dependence on $\hat{\pi}_l$ in the system (23)-(28). The resultant moment indicators are

$$g_\varepsilon(v,\varphi)_{H_y} = srI_{(y=t)} - H_t$$

$$g_\varepsilon(v,\varphi)_{H_S} = 1 - s - H_S$$

$$g_\varepsilon(v,\varphi)_{\theta_\varepsilon} = srp\frac{y-P}{P(1-P)} - \frac{\left[1-s(1-r)\right]\left(\frac{H_1}{Q}-\frac{H_0}{1-Q}\right)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} pf_\varepsilon(x)\,dx}{H_S + \frac{H_0}{1-Q} + \left(\frac{H_1}{Q}-\frac{H_0}{1-Q}\right)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} Pf_\varepsilon(x)\,dx}$$

$$g_\varepsilon\left(v,\varphi\right)_{\alpha_{yy^*_\varepsilon}} = \left[y - \Pr^*\left(y^*|x,\theta\right)\right]\left\{sr\frac{y-P}{P\left(1-P\right)} - \frac{\left[1-s\left(1-r\right)\right]\left(\frac{H_1}{Q}-\frac{H_0}{1-Q}\right)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} pf_\varepsilon\left(x\right)dx}{H_S+\frac{H_0}{1-Q}+\left(\frac{H_1}{Q}-\frac{H_0}{1-Q}\right)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} Pf_\varepsilon\left(x\right)dx}\right\}$$

$$g_\varepsilon\left(v,\varphi\right)_{Q^*_\varepsilon} = \alpha_{10} + \left(1-\alpha_{10}-\alpha_{01}\right)Q^* - \frac{\left[\left(1-s\right)r+s\right]\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} Pf_\varepsilon\left(x\right)dx}{H_S+\frac{H_0}{1-Q}+\left(\frac{H_1}{Q}-\frac{H_0}{1-Q}\right)\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} Pf_\varepsilon\left(x\right)dx}.$$

To compare the asymptotic covariance matrix of this parametric estimator with that of our semiparametric estimator, define $E_\varepsilon\left(P\right)=\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l}Pf_\varepsilon\left(x\right)dx$ and $E_\varepsilon\left(p\right)$, $E_\varepsilon\left(\nabla_{\theta\theta'}P\right)$, $E_\varepsilon\left(\nabla_{\alpha\alpha'}P\right)$ and $E_\varepsilon\left(\nabla_{\theta\alpha'}P\right)$ similarly. Hence, it is clear that these systems correspond to, respectively, (14)-(18) with P and p replaced by their expectations.

Assuming that P, $\nabla_\theta P$, $\nabla_{\theta\theta'}P$, $\nabla_{\alpha\alpha'}P$ and $\nabla_{\theta\alpha'}P$ are continuously differentiable with respect to x, there is uniform convergence of $E_\varepsilon\left(P\right)$, $E_\varepsilon\left(p\right)$, $E_\varepsilon\left(\nabla_{\theta\theta'}P\right)$, $E_\varepsilon\left(\nabla_{\alpha\alpha'}P\right)$ and $E_\varepsilon\left(\nabla_{\theta\alpha'}P\right)$ to P, p, $\nabla_{\theta\theta'}P$, $\nabla_{\alpha\alpha'}P$ and $\nabla_{\theta\alpha'}P$, respectively. Thus, the limits of $\Psi_\varepsilon = E_\varepsilon\left[g_\varepsilon\left(v,\varphi\right)\ g_\varepsilon\left(v,\varphi\right)'\right]$ and $G_\varepsilon = E_\varepsilon\left[\nabla_\varphi g_\varepsilon\left(v,\varphi\right)'\right]$ equal those of Ψ and G and the covariance matrix, $\left(G_\varepsilon\Psi_\varepsilon^{-1}G_\varepsilon'\right)^{-1}$, the Cramér-Rao bound, converges to $\left(G\Psi^{-1}G'\right)^{-1}$, which implies that our semiparametric estimators are efficient.

References

Abrevaya, J. and Hausman, J. A., 1999. Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells. Working Paper, MIT.

Copas, J. B., 1988. Binary regression models for contaminated data. Journal of the Royal Statistical Society B 50, 225-265.

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.

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

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

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

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, pp. 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, pp. 1977-1988.

Nicoletti, C. (2003), "Poverty analysis with unit and iten nonresponses: alternative estimators compared", Working Papers of the Institute for Social and Economic Research, paper 2003-20, University of Essex.

Newey, W. K. and McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R. F. and McFadden, D. L. (Eds), Handbook of Econometrics, vol. IV. Elsevier Science.

Peracchi, F. (2002), "The European Community Household Panel: a review", Empirical Economics, 1, pp.63-90.

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

Ramalho, E.A. and Smith, R.J. (2003), “Discrete choice nonresponse”, CeMMAP working paper CWP07/03.

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

Table 1: Summary statistics for GMM estimators from 1000 replications

$\theta^{\text{design A}}$=(0.0,2.0,0.5), $\theta^{\text{design B}}$=(-1.89,1.0,1.0)									
				$\hat{\theta}_1$			$\hat{\theta}_2$		
Design	$\bar{\alpha}$	Estimator	F. C.	Bias		St. D.	Bias		St. D.
				Mean	Med.		Mean	Med.	
A	.02	LIE	0	-.129	-.132	.150	-.128	-.132	.071
		MLIE	4	-.006	-.001	.321	.002	-.008	.119
	.05	LIE	1	-.275	-.277	.133	-.272	-.240	.068
		MLIE	7	.010	.005	.338	.012	.004	.116
	.20	LIE	63	-.647	-.651	.105	-.646	-.650	.055
		MLIE	13	.166	.041	2.850	.188	.020	.799
B	.02	LIE	5	-.069	-.073	.069	-.067	-.066	.065
		MLIE	9	.009	-.005	.423	.013	-.003	.449
	.05	LIE	5	-.166	-.169	.065	-.163	-.163	.062
		MLIE	9	.005	.006	.255	.011	.009	.272
	.20	LIE	363	-.467	-.472	.043	-.465	-.468	.041
		MLIE	66	.026	.023	.586	.018	.026	.414